LEHRBUCH

Klaus Backhaus
Bernd Erichson
Wulff Plinke
Rolf Weiber

Multivariate Analysemethoden

Eine anwendungsorientierte Einführung

15. Auflage

Preis der Deutschen Marktforschung 2015

EXTRAS ONLINE

Springer Gabler

Multivariate Analysemethoden

Klaus Backhaus · Bernd Erichson ·
Wulff Plinke · Rolf Weiber

Multivariate Analysemethoden

Eine anwendungsorientierte Einführung

15., vollständig überarbeitete Auflage

Springer Gabler

Klaus Backhaus
Marketing Center Münster
WWU Münster
Münster, Deutschland

Bernd Erichson
Otto-von-Guericke-Universität Magdeburg
Magdeburg, Deutschland

Wulff Plinke
European School of Management and Technology (ESMT)
Berlin, Deutschland

Rolf Weiber
Marketing, Innovation & E-Business
Universität Trier
Trier, Deutschland

ISBN 978-3-662-56654-1 ISBN 978-3-662-56655-8 (eBook)
https://doi.org/10.1007/978-3-662-56655-8

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

Springer Gabler

Lektorat: Barbara Roscher

Gedruckt auf säurefreiem und chlorfrei gebleichtem Papier

Springer Gabler ist ein Imprint der eingetragenen Gesellschaft Springer-Verlag GmbH, DE und ist ein Teil von Springer Nature.
Die Anschrift der Gesellschaft ist: Heidelberger Platz 3, 14197 Berlin, Germany

Vorwort zur 15. Auflage

Die erfreuliche Akzeptanz unseres Buches macht es erforderlich, nach drei Jahren eine weitere Neuauflage herauszubringen. Dies kommt unserem Bemühen entgegen, das Buch immer wieder zu verbessern und zu aktualisieren. Auch die 15. Auflage der Multivariaten Analysemethoden behält das mit der 12. Auflage realisierte Konzept bei und präsentiert neun „grundlegende“ Verfahren der multivariaten Datenanalyse, die besonders in der Bachelorausbildung häufig eingesetzt werden. Demgegenüber werden Verfahren, die aus unserer Sicht eher im Master- oder Doktorandenstudium vermittelt werden, in unserem Buch

> **Backhaus, Klaus/Erichson, Bernd/Weiber, Rolf:**
> Fortgeschrittene Multivariate Analysemethoden.
> Eine anwendungsorientierte Einführung, 3. Aufl., Berlin 2015

behandelt. In Teil III des vorliegenden Buches geben wir aber eine kurze Einführung in diese sog. „fortgeschrittenen“ Verfahren, um die Anwendungsbereiche für die „Fortgeschrittenen Verfahren“ auch für den Leser des Grundlagenwerkes transparent zu machen.

Für die 15. Auflage haben wir alle Verfahren mit SPSS 25 neu gerechnet und die entsprechenden SPSS-Outputs sowie die verwendeten SPSS-Screenshots ausgetauscht. Weiterhin haben wir inhaltliche Änderungen vor allem in den Kapiteln Regressionsanalyse, Diskriminanzanalyse und Faktorenanalyse vorgenommen. Darüber hinaus wurden Fehler korrigiert, wobei wir uns hier insbesondere bei den Nutzern unserer Internetseite **www.multivariate.de** bedanken, über die uns immer wieder Hinweise zu Fehlern und Vorschläge für ein besseres Verständnis der Texte erreichen. Ein besonderer Dank gilt an dieser Stelle Herrn Prof. Dr. habil. Gerhard Kockläuner, Institut für Statistik und Operations Research, Fachhochschule Kiel, der unser Buch sorgfältigst durchgesehen und uns auf eine Reihe von Fehlern sowie Inkonsistenzen hingewiesen hat.

Da wir das Buch selbst formatieren und die von uns verwendete Software zur Erstellung des Buchlayouts manchmal nicht nachvollziehbare Veränderungen vornimmt, müssen wir leider konstatieren, dass mit jeder Auflage Fehler eliminiert werden, gleichzeitig aber auch wieder neue Fehler im Text auftreten. Wir versuchen dies zwar mit großer Sorgfalt zu verhindern, leider lässt sich aber das Entstehen neuer Fehler nicht vollends vermeiden. Wir bitten unsere Leserinnen und Leser daher, die Internet-Seite **www.multivariate.de** intensiv zu nutzen, um uns Mängel mitzuteilen. Wir werden schnellstmöglich darauf reagieren.

Auch bei der Erstellung der 15. Auflage haben uns wieder unsere Söhne Dr. Max Backhaus, Dr. N. Benjamin Erichson und Dr. Thomas Weiber unterstützt und insb. mit Diskussionsbeiträgen zur Verbesserung der vorliegenden Auflage beigetragen. Darüber hinaus waren auch unsere wissenschaftlichen Mitarbeiterinnen und Mitarbeiter sowie verschiedenen Hilfskräfte in die Erstellung der Neuauflage eingeschaltet. Ein besonderer Dank gilt dabei an der Universität Trier Herrn M. Sc. David Lichter, der die Gesamtkoordination der neuen Auflage übernommen und uns bei der Korrespondenz mit dem Verlag unterstützt hat. Weiterhin haben in Trier die Herren M. Sc. Lorenz Gabriel, M. Sc. Lukas Mohr und M. Eng. Julian Morgen vor allem bei der Rechnung der Fallbeispiele mit SPSS 25 sowie der Fehlerkorrektur mitgewirkt. In Münster geht unser Dank an Herrn M. Sc. Ulf König. In Magdeburg waren Frau Dr. Franziska Rumpel und Herr Christian Stahr behilflich. Ein herzlicher Dank geht auch an Frau Barbara Roscher sowie Frau Birgit Borstelmann vom SpringerGabler-Verlag für Ihre Hilfestellungen und die gute Zusammenarbeit bei der Erstellung der vorliegenden Auflage.

Insgesamt können wir wieder eine Auflage präsentieren, die auf der aktuellen Version SPSS 25 basiert und an den Stand der aktuellen Entwicklungen angepasst ist. Evtl. noch vorhandene Fehler gehen selbstverständlich zu unseren Lasten.

Im Juni 2018

Klaus Backhaus, Münster Bernd Erichson, Magdeburg Rolf Weiber, Trier

Vorwort zur 14. Auflage

Anlässlich des 50-jährigen Bestehens des Berufsverbandes Deutscher Markt- und Sozialforscher e. V. (BVM) wurden im Juni 2015 die beiden von uns verfassten Bücher zu multivariaten Analysemethoden mit dem BVM-Preis in der Kategorie „Persönlichkeit des Jahres“ ausgezeichnet. Der Preis wird nur verliehen, wenn nach Einschätzung der BVM-Jury eine Leistung erkennbar ist, die als herausragend und zukunftsweisend zu bezeichnen ist. Es ehrt uns in besonderer Weise, dass wir diesen Preis erhalten haben und damit vor allem der Einfluss unseres Buches auf die Marktforschungspraxis in den letzten 35 Jahren gewürdigt wird. Der Preis ist uns Verpflichtung und Ansporn für die Zukunft und gegenüber unseren Lesern. Zeitgleich mit der Preisverleihung durch den BVM können wir bereits die 14. Auflage der „Multivariaten Analysemethoden“ präsentieren, die vor allem durch Verbesserungen in Einzelaspekten sowie Aktualisierungen gekennzeichnet ist. Auch die 14. Auflage „MVA“ behält das mit der 12. Auflage realisierte Konzept bei und präsentiert neun „grundlegende“ Verfahren der multivariaten Datenanalyse, die besonders häufig in der Bachelorausbildung gelehrt werden. Demgegenüber werden Verfahren, die aus unserer Sicht eher im Master- oder Doktorandenstudium vermittelt werden, in unserem ebenfalls neu aufgelegten Buch

> **Backhaus, Klaus/Erichson, Bernd/Weiber, Rolf:**
> Fortgeschrittene Multivariate Analysemethoden.
> Eine anwendungsorientierte Einführung, 3. Aufl., Berlin 2015

behandelt. Um auch den Lesern des vorliegenden Werkes eine kurze Einführung in die sog. „Fortgeschrittenen Verfahren“ zu geben, werden diese in Teil III des vorliegenden Buches auf jeweils ca. sieben Seiten in ihren wesentlichen Charakteristika dargestellt. Für die vorliegende 14. Auflage der „Multivariaten“ haben wir uns erstmals entschlossen, in die Überarbeitung unsere Söhne einzubeziehen.

Max Backhaus	Dipl.-Kfm., CEM MiM, Wiss. Mitarbeiter Universität zu Köln
Ben Erichson	M.Sc., Doktorand Universität St Andrews
Dr. Thomas Weiber	Dipl.-Math., Dipl.-Kfm., Senior Consultant bei EbelHofer Strategy & Management Consultants, München

Der Entschluss beruht auf den vielfältigen Diskussionen, die wir mit unseren Söhnen in den vergangenen Jahren zu Fragen der multivariaten Analysemethoden geführt haben sowie den einschlägigen Methodenkenntnissen unserer Söhne und ihren aktuellen Tätigkeitsfeldern, die ebenfalls stark methodisch ausgerichtet sind. Über ihr Mitwirken bei der vorliegenden Auflage sind wir besonders erfreut, und es macht das Lehrbuch zu einem generationenübergreifenden Werk. Wir sind gespannt, welche Auswirkungen die gemeinsame Arbeit von Vätern und Söhnen für die Zukunft haben wird.
Die 14. Auflage weist insbesondere folgende Änderungen bzw. Neuerungen auf:

1. Beispiele, soweit sie mit IBM SPSS gerechnet wurden, sind auf die neueste SPSS-Version umgestellt worden (SPSS 23). Dabei zeigte sich, dass in SPSS 22 viele Outputs in meist kleinen Details geändert wurden, was dann aber in SPSS 23 wieder rückgängig gemacht wurde. Im Ergebnis entsprechen damit die SPSS-Outputs der Version 23 weitgehend wieder denen der Version 19, die der 13. Auflage dieses Werkes zu Grunde lag. Durch den Abgleich der Outputs mit der aktuell verfügbaren SPSS-Version möchten wir sicherstellen, dass der Leser die eigenen Ergebnisse - auch wenn er frühere Versionen von SPSS verwendet - mit den Outputs der neuesten Programmversion vergleichen kann.

2. Zum Teil haben wir die Erklärungen in einzelnen Kapiteln erweitert und/oder verbessert. Dabei haben uns auch Reaktionen von Leserinnen und Lesern wertvolle Anregungen gegeben, wofür wir uns sehr herzlich bedanken. Ein besonderer Dank gilt den Kollegen Prof. Dr. Björn Christensen, Prof. Dr. Dominik Papies, Dr. Denis Proppe und Prof. Dr. Michel Clement sowie auch Herrn Dr. Dirk Windelberg für wertvolle Hinweise. Auch in Zukunft sind uns Anregungen oder kritische Anmerkungen jederzeit höchst willkommen.

3. In einzelnen Kapiteln haben wir Erweiterungen und Ergänzungen vorgenommen, wobei wir uns bei der Entscheidung über die Aufnahme neuer Auswertungsoptionen daran orientiert haben, ob diese für unsere Zielgruppe (Nutzer mit starker Anwendungsorientierung) sinnvoll sind: So wurde z. B. bei der *Regressionsanalyse* der Datensatz aus didaktischen Gründen verändert und ein Kapitel zur Residuen-Analyse aufgenommen, die *Varianzanalyse* wurde um die Kontrastanalyse sowie multiple Vergleichstests (Post hoc-Tests) erweitert, das Kapitel zur *Logistischen Regression* wurde grundlegend überarbeitet und die *Faktorenanalyse* um verschiedene Extraktionsverfahren, z.B. das Maximum Likelihood-Verfahren, ergänzt. Für alle Kapitel wurden die Literaturempfehlungen und die zitierten Literaturquellen überarbeitet und auf den neuesten Stand gebracht.

Auch bei der 14. Auflage haben uns die wissenschaftlichen Mitarbeiterinnen und Mitarbeiter sowie viele Hilfskräfte tatkräftig unterstützt. Ein besonderer Dank geht dabei an Herrn Stefan Benthaus, M. Sc., und Herrn Matthias Reese, alle Universität Münster. Sie haben die zentrale Koordination des Werkes übernommen und in unermüdlichen kleinen Schritten die Tücken der verwendeten Schreib-Software LaTeX überwunden, was sich bei dieser Auflage wirklich als Sisyphusarbeit herausstellte. In Trier haben Frau Katharina Ferreira, Dipl.-Kffr. und die Herren Michael Bathen, Dipl.-Kfm., David Lichter, M. Sc. und Dominic Link, B. Sc., vor allem Literatur aktualisiert und akribisch Änderungen in den SPSS-Outputs gegenüber vorherigen Versionen abgeglichen.

Insgesamt können wir wieder eine Auflage präsentieren, die auf der aktuellen Version von IBM SPSS 23 basiert und an den Stand der aktuellen Entwicklungen angepasst ist. Evtl. noch vorhandene Fehler gehen selbstverständlich zu unseren Lasten.

Im Juni 2015
Klaus Backhaus, Münster Bernd Erichson, Magdeburg Rolf Weiber, Trier

Vorwort zur 12. Auflage

Mit der 12. Auflage der „Multivariaten Analysemethoden“ liegt eine grundlegende Überarbeitung und wesentliche Erweiterung der methodischen Inhalte der 11. Auflage vor, ohne dass die bisherigen Verfahren eingeschränkt oder verdrängt wurden. Als neue Verfahren wurden die Zeitreihenanalyse, die Nichtlineare Regressionsanalyse, die Konfirmatorische Faktorenanalyse und ein Kapitel zu auswahlbasierten Verfahren der Conjoint-Analyse aufgenommen. Alle übrigen Verfahren wurden überarbeitet und insbesondere um die für Einsteiger zentralen Optionen von SPSS 16.0 ergänzt und bezüglich der SPSS-Screenshots aktualisiert.

Diese umfangreichen inhaltlichen Erweiterungen und Überarbeitungen hatten zur Folge, dass mit der 12. Auflage das Volumen des Buches nochmals erheblich gestiegen ist. Die 11. Auflage hatte bereits mit ihrem äußeren Umfang die Grenze einer nutzergerechten Handhabung des Buches erreicht. Es war deshalb unausweichlich, eine Grundsatzentscheidung hinsichtlich des Umfangs zu treffen. Wir haben uns deshalb entschieden, das Buch *nicht* inhaltlich zu kürzen, sondern für die Leserinnen und Leser einen neuen Weg des Zugangs zu allen Methoden zu schaffen. Das Gesamtwerk wurde deshalb in zwei Teile gegliedert:

1. Das **vorliegende Buch** umfasst in ausführlicher Darstellung in den Kapiteln 1 bis 9 *„Grundlegende Verfahren der multivariaten Analyse“*, die in der bisher bewährten Form im Detail dargestellt werden. In den Kapiteln 10 bis 16 werden *„Komplexe Verfahren der multivariaten Analyse“* jeweils auf ca. 6 Seiten in ihren elementaren Grundzügen erläutert.

2. Über die **Internetplattform** (www.multivariate.de) zu diesem Buch stellen wir unseren Leserinnen und Lesern jeweils auch eine Darstellung der *„Komplexen Verfahren der multivariaten Analyse“* (Kapitel 10 bis 16) im Detail zur Verfügung.

Nachfolgende Tabelle auf Seite VI gibt einen Überblick über die Zuordnung der Verfahren jeweils zum Buch oder zur Internetplattform.

Die neue Aufteilung des Inhalts erlaubt es, den Verkaufspreis des Buches trotz erheblich anspruchsvollerer Designqualität zu halten. Mit der gefundenen neuen äußeren Form haben wir uns auch bemüht, die Lesefreundlichkeit durch ein vergrößertes Seitenformat, durch Farbgebung, durch professionelle Satztechnik sowie die Hinzufügung von Marginalien zu erhöhen.

Wir danken einer Vielzahl von Leserinnen und Lesern, die uns durch ihre kritischen Hinweise auf Fehler aufmerksam gemacht haben. Wir bedauern sehr, dass sich trotz größter Sorgfalt Fehler eingeschlichen haben und befürchten aus der Erfahrung früherer Auflagen, dass dieses auch bei der 12. Auflage nicht völlig ausgeschlossen werden kann. Umso mehr schätzen wir den offenen Dialog mit unseren Leserinnen und Lesern.

MVA-Buch "Grundlegende Verfahren der multivariaten Analyse"	MVA-Internetplattform "Komplexe Verfahren der multivariaten Analyse"
1. Regressionsanalyse	10. Nichtlineare Regression
2. Zeitreihenanalyse	11. Strukturgleichungsmodelle
3. Varianzanalyse	12. Konfirmatorische Faktorenanalyse
4. Diskriminanzanalyse	13. Auswahlbasierte Conjoint-Analyse
5. Logistische Regression	14. Neuronale Netze
6. Kreuztabellierung und Kontingenzanalyse	15. Multidimensionale Skalierung
7. Faktorenanalyse	16. Korrespondenzanalyse
8. Clusteranalyse	
9. Conjoint-Analyse	

Wiederum sind wir mit der neuen Auflage unserer bewährten Leitlinie gefolgt, die seit Anbeginn von unseren Lesern geschätzt wurde: „Geringstmögliche Anforderungen an mathematische Vorkenntnisse und Gewährleistung einer allgemein verständlichen Darstellung anhand eines für mehrere Methoden entwickelten Beispiels." Das konsequente Verfolgen dieser Konzeption führt natürlich dazu, dass wir auf eine Fülle von Detailfragen nicht eingehen können, weil das Grundverständnis vor dem Detail rangiert. Auf unserer Plattform **www.multivariate.de** haben wir aber für jedes Verfahren Angaben zu weiterer Spezialliteratur bereitgestellt, die wir kontinuierlich aktualisieren. Hier können auch Anwendungsfragen diskutiert werden. Aber dennoch möchten wir an unserem Grundsatz festhalten: Das Buch ist kein Lehrbuch von Spezialisten für Spezialisten, sondern von Anwendern für Anwender!

Für die neue umfänglich bearbeitete, erweiterte und äußerlich neu gestaltete 12. Auflage schulden wir unseren Mitarbeiterinnen und Mitarbeitern Dank für vielfältige und umfassende Hilfe, nicht nur bei der Lektüre der einzelnen Kapital, sondern auch in Form der kritischen Begleitung der neuen Textfassung ebenso wie die großen Mühen der Dokumentation:
In Münster haben sich Dipl.-Ing. Harald Neun und Dipl.-Kfm. Alfred Zerres in unermüdlicher Sisyphusarbeit um die Koordination der Erstellung der 12. Auflage gekümmert. Sie haben einen Stab von studentischen Hilfskräften geführt, die mit der Transformation des Manuskriptes in das neue Design befasst waren. Besonderen Dank schulden wir cand. rer. pol. Oliver Behrla, Hossein Ghodrati, Alexander Heck, Silja Motullo, Marie Louise Orth, Daniel Piegsa und Christopher Vierhaus.
In Magdeburg haben Dipl.-Kffr. Franziska Rumpel und Frau cand. rer. pol. Betül Kural sowie in Trier Dipl.-Kfm. Steffen Freichel, Dipl.-Kfm. Robert Hörstrup, Dipl.-Volksw. Dipl.-Kfm. Daniel Mühlhaus und Dipl.-Kffr. Nina Pecornik immer wieder neue Textfassungen gelesen, konstruktive Verbesserungsvorschläge unterbreitet und bei der abschließenden Kontrolle der Verlagsversion mitgewirkt.

Selbstverständlich gehen alle eventuellen Mängel zu unseren Lasten.

Im Juli 2008

Klaus Backhaus, Münster — Bernd Erichson, Magdeburg
Wulff Plinke, Berlin — Rolf Weiber, Trier

Inhaltsverzeichnis

Teil I

Benutzungshinweise

www.multivariate.de

Zu den Büchern „Backhaus/Erichson/Plinke/Weiber: Multivariate Analysemethoden, 15. Aufl., Berlin 2018“ und „Backhaus/Erichson/Weiber: Fortgeschrittene Multivariate Analysemethoden, 3. Aufl., Berlin 2015“ finden die Leserinnen und Leser im Internet unter der Adresse

www.multivariate.de

unterschiedliche Unterstützungsleistungen zu den in beiden Büchern behandelten Verfahren der multivariaten Datenanalyse. Ziel dieser Internetpräsenz ist es, ergänzend zu den beiden Lehrbüchern auch zwischen den verschiedenen Auflagen auf aktuelle Entwicklungen hinzuweisen und eine Plattform für den Erfahrungsaustausch auch unter den Nutzern der Bücher zu schaffen. Den Kern der Internetpräsenz bilden die folgenden Serviceleistungen:

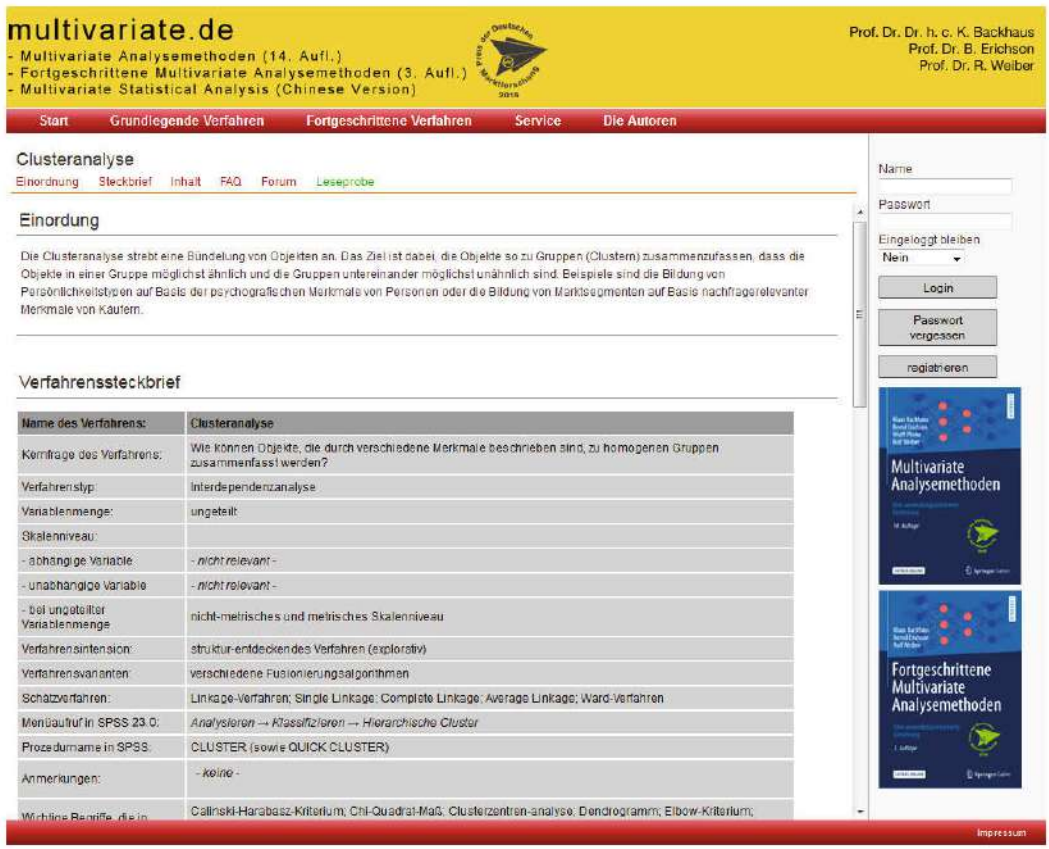
- **MVA-Grundlegende Verfahren**
  Zu den im Buch „*Multivariate Analysemethoden, 15. Aufl.*“ behandelten grundlegenden Verfahren der multivariaten Analyse finden die interessierten Leserinnen und Leser jeweils eine Einordnung dieser Verfahren, einen kurzen Verfahrenssteckbrief sowie eine Übersicht der jeweiligen Kapitelinhalte.

- **MVA-Fortgeschrittene Verfahren**
  Zu den im Buch „*Fortgeschrittene Multivariate Analysemethoden, 3. Aufl.*“ behandelten Verfahren der multivariaten Analyse finden die interessierten Leserinnen und Leser jeweils eine Einordnung dieser Verfahren, einen kurzen Verfahrenssteckbrief sowie eine Übersicht der jeweiligen Kapitelinhalte.

- **MVA-FAQ**
  Häufig gestellte Fragen und Hinweise zu den Verfahren werden unter der Rubrik „*Frequently Asked Questions*“ übersichtlich archiviert, so dass eine schnelle Problemlösung bei häufigen Anwenderfragen gewährleistet ist. Die FAQs sind geordnet und bei jedem Verfahren gesondert aufgeführt.

- **MVA-Forum zu den einzelnen Analysemethoden**
  Die Internetseite bietet spezielle verfahrensspezifische Foren, in denen sich die Anwenderinnen und Anwender über Verfahrensprobleme austauschen und z. B. gemeinsam Lösungen für spezifische Anwendungssituationen finden können. Dabei sind sowohl Fragen von Experten, wie auch von Nicht-Experten, Lösungshinweise oder Verbesserungsvorschläge gerne erwünscht. Überdies werden die Diskussionen regelmäßig von den Autoren verfolgt und durch konstruktive Beiträge das Umfeld „des gemeinsamen Lernens“ unterstützt.

K. Backhaus, B. Erichson, W. Plinke, R. Weiber, *Multivariate Analysemethoden*,
https://doi.org/10.1007/978-3-662-56655-8_1



www.multivariate.de

- **MVA-Forum zum Buchkonzept**
  Unter dem *Register „Service“* wird ein allgemeines Forum zu beiden Buchkonzepten angeboten. Hier freuen sich die Autoren auch über neue Konzeptvorschläge und beantworten spezielle Fragen zu beiden Büchern.

- **MVA-Anwender- und Dozentensupport**
  Über den MVA-Support können unter dem *Register „Service“* sowohl alle Abbildungen als PowerPoint-Datei als auch die Datensätze und SPSS-Jobs zu allen Verfahren schnell und bequem bestellt werden.

- **MVA-Korrekturliste**
  Das *Register „Service“* enthält für beide Bücher jeweils eine Korrekturliste, in der die Autoren über nach Drucklegung ggf. bemerkte Fehler in der jeweils aktuellen Auflage informieren.

- **MVA-Feedback an Autoren**
  Die Autoren freuen sich, wenn die Leserinnen und Leser in den beiden Büchern entdeckte Fehler über das Feedbackformular direkt an die Autoren melden.

- - - - - - - - - - - - - - - **Faxantwort an +49/651/2013910** - - - - - - - - - - - - - - -

Absender:

_____

_____

_____

Tel.: _____

Mail: _____

Professur für Marketing,
Innovation & E-Business
Univ.-Prof. Dr. Rolf Weiber
Universitätsring 15
D-54296 **Trier**

Betr.: Multivariate Analysemethoden 15. Auflage

Hiermit bestelle ich (zuzüglich Versandkosten):

- ☐ die Datensätze und Syntaxdateien zu allen *„Grundlegenden Verfahren“* (Teil II des Buches) zum Gesamtpreis von 3 Euro;
- ☐ das komplette Set der Abbildungen zu den *„Grundlegenden Verfahren“* (Teil II des Buches) sowie den Kurzfassungen der *„Fortgeschrittenen Verfahren“*(Teil III des Buches) als geschützte Powerpoint-Dateien (15 Euro);

Das Set der *Abbildungen* als geschützte Powerpoint-Dateien für die einzelnen Kapitel der Grundlegenden Verfahren kann zum Preis von je 2,50 Euro erworben werden. Die Abbildungen zu den einzelnen Kapiteln der *Kurzfassungen* der Fortgeschrittenen Verfahren werden unter der Adresse **www.multivariate.de** im Internet kostenlos zum Download bereitgestellt.
Hiermit bestelle ich (zuzüglich Versandkosten) die Abbildungen zu folgenden Kapiteln:

- ☐ Zur Verwendung dieses Buches
- ☐ Clusteranalyse
- ☐ Conjoint-Analyse
- ☐ Diskriminanzanalyse
- ☐ Faktorenanalyse
- ☐ Kreuztabellierung und Kontingenzanalyse
- ☐ Logistische Regression
- ☐ Regressionsanalyse
- ☐ Varianzanalyse
- ☐ Zeitreihenanalyse

Die Bestellung soll

- ☐ postalisch als CD oder Stick versendet werden (plus Versandkosten)
- ☐ elektronisch zugesendet werden (hier entstehen *keine* Versandkosten)

_____ _____

Datum Unterschrift

Die Bestellung ist auch über **www.multivariate.de** möglich!

Zur Verwendung dieses Buches

1 Zielsetzung des Buches

Mittels multivariater Analysemethoden werden mehrere Variablen simultan betrachtet und deren Zusammenhang quantitativ analysiert, sei es, um ihn zu beschreiben, zu erklären oder um zukünftige Entwicklungen zu prognostizieren. Einen Spezialfall bilden bivariate Analysemethoden, bei denen nur jeweils zwei Variablen betrachtet werden. Die Realität aber ist komplex und deren adäquate Abbildung macht daher in der Regel die Anwendung multivariater Verfahren erforderlich. Die Bewältigung des damit verbundenen oft recht hohen Rechenaufwandes bildet seit der Verfügbarkeit von leistungsfähigen Computern und geeigneter Software keinen Engpass mehr.

Multivariate Analysemethoden sind daher heute eines der Fundamente der empirischen Forschung in den Realwissenschaften. Die Methoden sind immer noch in stürmischer Entwicklung. Es werden ständig neue methodische Varianten entwickelt, neue Anwendungsbereiche erschlossen und neue oder verbesserte Computer-Programme entwickelt. Mancher Interessierte aber empfindet Zugangsbarrieren zur Anwendung der Methoden, die aus

Überwindung von Zugangsbarrieren

- Vorbehalten gegenüber mathematischen Darstellungen,
- einer gewissen Scheu vor dem Einsatz des Computers und
- mangelnder Kenntnis der Methoden und ihrer Anwendungsmöglichkeiten

resultieren. Es ist eine Kluft zwischen interessierten Fachleuten und Methodenexperten festzustellen, die bisher nicht genügend durch das Angebot der Fachliteratur überbrückt wird.

Die Autoren dieses Buches haben sich deshalb das Ziel gesetzt, zur Überwindung dieser Kluft beizutragen. Daraus ist ein Text entstanden, der folgende Charakteristika besonders herausstellt:

Allgemeinverständlichkeit

1. Es ist größte Sorgfalt darauf verwendet worden, die Methoden *allgemeinverständlich* darzustellen. Der Zugang zum Verständnis durch den mathematisch ungeschulten Leser hat in allen Kapiteln Vorrang gegenüber dem methodischen Detail. Dennoch wird der rechnerische Gehalt der Methoden in den wesentlichen Grundzügen erklärt, damit sich der Leser, der sich in die Methoden einarbeitet, eine Vorstellung von der Funktionsweise, den Möglichkeiten und Grenzen der Methoden verschaffen kann.

Beispiele

2. Das Verständnis wird erleichtert durch die ausführliche Darstellung von *Beispielen*, die es erlauben, die Vorgehensweise der Methoden leicht nachzuvollziehen und zu verstehen.

3. Darüber hinaus wurde – soweit die Methoden das zulassen – ein Beispiel durchgehend für mehrere Methoden benutzt, um das Einarbeiten zu erleichtern und um die Ergebnisse der Methoden vergleichen zu können. Die Rohdaten der Beispiele können über den Bestellschein oder über die Internetadresse www.multivariate.de angefordert werden. Rohdaten

   Die Beispiele sind dem Marketing-Bereich entnommen. Die Darstellung ist jedoch so gehalten, dass jeder Leser die Fragestellung versteht und auf seine spezifischen Anwendungsprobleme in anderen Bereichen übertragen kann.

4. Der Umfang des zu verarbeitenden Datenmaterials ist in aller Regel so groß, dass die Rechenprozeduren der einzelnen Verfahren mit vertretbarem Aufwand nur computergestützt durchgeführt werden können. Deshalb erstreckt sich die Darstellung der Methoden sowohl auf die Grundkonzepte der Methoden als auch auf die *Nutzung geeigneter Computer-Programme* als Arbeitshilfe. Es existiert heute eine Reihe von Programmpaketen, die die Anwendung multivariater Analysemethoden nicht nur dem Computer-Spezialisten erlauben. Insbesondere bedingt durch die zunehmende Verbreitung und Leistungsfähigkeit des PCs sowie die komfortablere Gestaltung von Benutzeroberflächen wird auch die Nutzung der Programme zunehmend erleichtert. Damit wird der Fachmann für das Sachproblem unabhängig vom Computer-Spezialisten. Software-Unterstützung

   Das Programmpaket bzw. Programmsystem, mit dem die meisten Beispiele durchgerechnet werden, ist *IBM SPSS Statistics* oder kurz *SPSS*. Als Programmsystem wird dabei eine Sammlung von Programmen mit einer gemeinsamen Benutzeroberfläche bezeichnet. IBM SPSS hat sehr weite Verbreitung gefunden, besonders im Hochschulbereich, aber auch in der Praxis. Es ist unter vielen Betriebssystemen auf Großrechnern, Workstations und PC verfügbar. IBM SPSS

5. Das vorliegende Buch hat den Charakter eines *Arbeitsbuches*. Die Darstellungen sind so gewählt, dass der Leser in jedem Fall alle Schritte der Lösungsfindung nachvollziehen kann. Alle Ausgangsdaten, die den Beispielen zugrunde liegen, können für die umfangreicheren Fallbeispiele über www.multivariate.de bestellt werden. Die Syntaxkommandos für die Computer-Programme werden im Einzelnen aufgeführt, so dass der Leser durch eigenes Probieren sehr schnell erkennen kann, wie leicht letztlich der Zugang zur Anwendung der Methoden unter Einsatz des Computers ist, wobei er seine eigenen Ergebnisse gegen die im vorliegenden Buch ausgewiesenen kontrollieren kann. Arbeitsbuch

6. Die Ergebnisse der computergestützten Rechnungen in den einzelnen Methoden werden jeweils anhand der betreffenden *Programmausdrucke* erläutert und kommentiert. Dadurch kann der Leser, der sich in die Handhabung der Methoden einarbeitet, schnell in den eigenen Ergebnissen eine Orientierung finden. Programmausdrucke

Interpretation

7. Besonderes Gewicht wurde auf die *inhaltliche Interpretation* der Ergebnisse der einzelnen Verfahren gelegt. Wir haben es uns dabei zur Aufgabe gemacht, die *Ansatzpunkte für Ergebnismanipulationen* in den Verfahren offenzulegen und die Gestaltungsspielräume aufzuzeigen, damit der Anwender der Methoden objektive und subjektive Bestimmungsfaktoren der Ergebnisse unterscheiden kann. Dies macht es u. a. erforderlich, dass methodische Details offengelegt werden. Dies macht deutlich, dass dem Anwender der Methoden eine Verantwortung für seine Interpretation der Ergebnisse zukommt.

Fasst man die genannten Merkmale des Buches zusammen, dann ergibt sich ein Konzept, das geeignet ist, sowohl dem Anfänger, der sich in die Handhabung der Methoden einarbeitet, als auch demjenigen, der mit den Ergebnissen dieser Methoden arbeiten muss, die erforderliche Hilfe zu geben. Die Konzeption lässt es dabei zu, dass *jede dargestellte Methode für sich verständlich* ist. Der Leser ist also an keine Reihenfolge der Kapitel gebunden.

Im Folgenden wird ein knapper Überblick über die von uns behandelten Verfahren der multivariaten Analysetechnik gegeben. Da sich die einzelnen Verfahren vor allem danach unterscheiden lassen, welche Anforderungen sie an das Datenmaterial stellen, seien hierzu einige Bemerkungen vorausgeschickt, die für Anfänger gedacht und deshalb betont knapp gehalten sind.[1]

2 Daten, Skalen und Variablen

Das „Rohmaterial“ für multivariate Analysen sind die (vorhandenen oder noch zu erhebenden) *Daten*. Die Qualität von Daten wird u. a. bestimmt durch die Art und Weise der *Messung*. Daten sind nämlich das Ergebnis von Messvorgängen. Messen bedeutet, dass Eigenschaften von Objekten nach bestimmten Regeln in Zahlen ausgedrückt werden.

Messung

Im wesentlichen bestimmt die jeweils betrachtete Art einer Eigenschaft, wie gut man ihre Ausprägung messen, d. h. wie gut man sie in Zahlen ausdrücken kann. Im Wesentlichen wird z. B. die Körpergröße eines Menschen sehr leicht in Zahlen auszudrücken sein, seine Intelligenz, seine Motivation oder sein Gesundheitszustand dagegen sehr schwierig.

Das Skalenniveau von Messungen

Die „Messlatte“, auf der die Ausprägungen einer Eigenschaft abgetragen werden, heisst *Skala*. Je nachdem, in welcher Art und Weise eine Eigenschaft eines Objektes in Zahlen ausgedrückt (gemessen) werden kann, unterscheidet man Skalen mit unterschiedlichem *Skalenniveau*:

Skalen(niveau)

1. Nominalskala
2. Ordinalskala
3. Intervallskala
4. Ratioskala.

[1] Zu den statistischen Grundlagen vgl. z. B. Bleymüller/Weißbach (2015) oder Fahrmeir et al. (2016).

Das Skalenniveau bedingt sowohl den *Informationsgehalt der Daten* wie auch die *Anwendbarkeit von Rechenoperationen*. Abbildung 1 gibt hierzu einen zusammenfassenden Überblick. Im Folgenden werden die Skalentypen und ihre Eigenschaften kurz erläutert:

Skala		Merkmale	Mögliche rechnerische Handhabung
nicht- metrische Skalen (kategorial)	NOMINALSKALA	Klassifizierung qualitativer Eigenschaftsausprägungen	Bildung von Häufigkeiten
	ORDINALSKALA	Rangwert mit Ordinalzahlen	Median, Quantile
metrische Skalen (kardinal)	INTERVALLSKALA	Skala mit gleichgroßen Abschnitten ohne natürlichen Nullpunkt	Subtraktion, Mittelwert
	RATIOSKALA	Skala mit gleichgroßen Abschnitten und natürlichem Nullpunkt	Summe, Division, Multiplikation

Abbildung 1: Skalenniveaus

Die *Nominalskala* stellt die „primitivste“ Grundlage des Messens dar. Beispiele für Nominalskalen sind

Nominalskala

- Geschlecht (männlich – weiblich)
- Religion (katholisch – evangelisch – andere)
- Farbe (rot – gelb – grün – blau ...)
- Werbemedium (Fernsehen – Zeitungen – Plakattafeln).

Nominalskalen stellen also Klassifizierungen qualitativer Eigenschaftsausprägungen dar. Zwecks leichterer Verarbeitung mit Computern werden die Ausprägungen von Eigenschaften häufig durch Zahlen ausgedrückt. So lassen sich z. B. die Farben einer Verpackung wie folgt kodieren:

$$\begin{aligned} \text{rot} &= 1 \\ \text{gelb} &= 2 \\ \text{grün} &= 3 \end{aligned}$$

Die Zahlen hätten auch in anderer Weise zugeordnet werden können, solange diese Zuordnung eineindeutig ist, d. h. solange durch eine Zahl genau eine Farbe definiert ist. Mit derartigen Zahlen sind daher keine arithmetischen Operationen (wie Addition, Subtraktion, Multiplikation oder Division) erlaubt. Vielmehr lassen sich lediglich durch Zählen der Merkmalsausprägungen (bzw. der sie repräsentierenden Zahlen) Häufigkeiten ermitteln.

Rechenoperationen

Eine *Ordinalskala* stellt das nächsthöhere Messniveau dar. Die Ordinalskala erlaubt die Aufstellung einer Rangordnung mit Hilfe von Rangwerten (d. h. ordinalen Zahlen). Beispiele: Produkt A wird Produkt B vorgezogen, Herr M. ist tüchtiger als Herr N. Die Untersuchungsobjekte können immer nur in eine Rangordnung gebracht werden. Die Rangwerte 1., 2., 3. etc. sagen nichts über die Abstände zwischen den Objekten

Ordinalskala

aus. Aus der Ordinalskala kann also nicht abgelesen werden, um wieviel das Produkt A besser eingeschätzt wird als das Produkt B. Daher dürfen auch ordinale Daten, ebenso wie nominale Daten, nicht arithmetischen Operationen unterzogen werden. Zulässige statistische Maße sind neben Häufigkeiten z. B. der Median oder Quantile.

Intervallskala

Das wiederum nächsthöhere Messniveau stellt die *Intervallskala* dar. Diese weist gleichgroße Skalenabschnitte aus. Ein typisches Beispiel ist die Celsius-Skala zur Temperaturmessung, bei der der Abstand zwischen Gefrierpunkt und Siedepunkt des Wassers in hundert gleichgroße Abschnitte eingeteilt wird. Bei intervallskalierten Daten besitzen auch die Differenzen zwischen den Daten Informationsgehalt (z. B. großer oder kleiner Temperaturunterschied), was bei nominalen oder ordinalen Daten nicht der Fall ist.

Oftmals werden – auch in dem vorliegenden Buch – Skalen benutzt, von denen man lediglich annimmt, sie seien intervallskaliert. Dies ist z. B. der Fall bei Ratingskalen: Eine Auskunftsperson ordnet einer Eigenschaft eines Objektes einen Zahlenwert auf einer Skala von 1 bis 7 (oder einer kürzeren oder längeren Skala) zu. Solange die Annahme gleicher Skalenabstände unbestätigt ist, handelt es sich allerdings strenggenommen um eine Ordinalskala.

Intervallskalierte Daten erlauben die arithmetischen Operationen der Addition und Subtraktion. Zulässige statistische Maße sind, zusätzlich zu den oben genannten, z. B. der Mittelwert (arithmetisches Mittel) und die Standardabweichung, nicht aber die Summe.

Ratio(Verhältnis)Skala

Die *Ratio- (oder Verhältnis)skala* stellt das höchste Messniveau dar. Sie unterscheidet sich von der Intervallskala dadurch, dass zusätzlich ein natürlicher Nullpunkt existiert, der sich für das betreffende Merkmal im Sinne von „nicht vorhanden“ interpretieren lässt. Das ist z. B. bei der Celsius-Skala oder der Kalenderzeit nicht der Fall, dagegen aber bei den meisten physikalischen Merkmalen (z. B. Länge, Gewicht, Geschwindigkeit) wie auch bei den meisten ökonomischen Merkmalen (z. B. Einkommen, Kosten, Preis). Bei verhältnisskalierten Daten besitzen nicht nur die Differenz, sondern, infolge der Fixierung des Nullpunktes, auch der Quotient bzw. das Verhältnis (Ratio) der Daten Informationsgehalt (daher der Name).

Ratioskalierte Daten erlauben die Anwendung aller arithmetischen Operationen wie auch die Anwendung aller obigen statistischen Maße. Zusätzlich sind z. B. die Anwendung des geometrischen Mittels oder des Variationskoeffizienten erlaubt.

Nominalskala und Ordinalskala bezeichnet man als nichtmetrische oder auch kategoriale Skalen, Intervallskala und Ratioskala als metrische Skalen.

Informationsgehalt

Zusammenfassend kann folgendes festgehalten werden: Je höher das Skalenniveau ist, desto größer ist auch der Informationsgehalt der betreffenden Daten und desto mehr Rechenoperationen und statistische Maße lassen sich auf die Daten anwenden. Es ist generell möglich, Daten von einem höheren Skalenniveau auf ein niedrigeres Skalenniveau zu transformieren, nicht aber umgekehrt. Dies kann sinnvoll sein, um die Übersichtlichkeit der Daten zu erhöhen oder um ihre Analyse zu vereinfachen. So werden z. B. häufig Einkommensklassen oder Preisklassen gebildet. Dabei kann es sich um eine Transformation der ursprünglich ratio-skalierten Daten auf eine Intervall-, Ordinal- oder Nominal-Skala handeln. Mit der Transformation auf ein niedrigeres Skalenniveau ist natürlich immer auch ein Informationsverlust verbunden.

Variable, Variablenwerte und Skalenniveaus

Die numerisch kodierten Eigenschaften (Merkmale) von Objekten werden als *Variablen* bezeichnet und zwecks kompakter Darstellung meist durch Buchstaben symbolisiert. Ihre Werte drücken die Eigenschaftsausprägungen von Objekten aus. Eine Variable variiert also über die betrachteten Objekte und möglicherweise auch über die Zeit.

Die Werte einer Variablen können das Ergebnis einer Messung oder auch Berechnung sein. Variablen lassen sich daher u. a. auch nach dem Skalenniveau ihrer Messung einteilen (daneben existieren vielfältige weitere Einteilungen). Besonders bedeutsam hinsichtlich Informationsgehalt und rechnerischer Handhabung von Variablen ist die Einteilung in

- metrische vs. nicht-metrische
- quantitative vs. qualitative
- kardinale vs. kategoriale

Variablen. Diese Begriffspaare werden weitgehend synonym verwendet.

Dabei bereitet die Einordnung von ordinalen Variablen allerdings manchmal Schwierigkeiten, da sie nicht immer eindeutig erfolgen kann. Sie werden oft wie nominale (und damit qualitative) Variablen behandelt, z. B. soziale Schicht (Unter-, Mittel-, Oberschicht) oder politische Einstellung (progressiv, gemäßigt, konservativ), da eine Transformation auf ein niedrigeres Skalenniveau immer möglich ist. Manchmal aber werden ordinale Variablen auch wie quantitative Variablen behandelt, was streng genommen nicht korrekt ist, z. B. wenn die Mittelwerte von Schulnoten oder Ratings berechnet werden.[2] Im engeren Sinn bezeichnet man daher als qualitative Variablen nur solche mit nominal-skalierten Werten.

Ähnliches gilt für kategoriale Variablen. Sie umfassen zwar sowohl nominale wie auch ordinale Variablen, letzteres allerdings nur, wenn die Anzahl der Ausprägungen nicht zu groß ist, so dass sie zur Kategorisierung dienen können.[3] So wären z. B. die ordinalen Variablen Schulnote (sehr gut bis ungenügend), Dienstgrad beim Militär (Soldat bis General), Medaillen bei Olympia (Gold, Silber, Bronze) auch kategoriale Variablen, nicht aber die ordinale Variable Rangplatz bei Wettbewerben mit meist zahlreichen Teilnehmern.

3 Einteilung multivariater Analysemethoden

Multivariate Analysemethoden unterliegen einer permanenten Weiter- und auch Neuentwicklung, und auch die Anwendung der Verfahren ist durch die Verfügbarkeit leistungsfähiger sowie benutzerfreundlicher Analyseprogramme weiterhin stark zunehmend. Vor diesem Hintergrund haben die Autoren dieses Buches versucht, eine Auswahl an multivariaten Analysemethoden zu treffen, denen sowohl in der Hochschul-Ausbildung als auch bei praktischen Anwendungen eine besondere Bedeutung beizumessen ist. Da in den letzten Jahren aber auch das Bildungssystem der Hochschulen mit der Einführung von Bachelor- und Master-Studiengängen eine grundlegende Veränderung erfahren hat, haben die Autoren weiterhin versucht,

[2] Siehe dazu z. B. Fahrmeir et al. (2016), S. 14 f.
[3] Zur Analyse kategorialer Daten siehe insbesondere Agresti (1990); Tutz (2000).

eine allgemeine Einschätzung vorzunehmen, welche multivariaten Verfahren eher in der Bachelor-Ausbildung und welche eher in der Master- und/oder Doktoranden-Ausbildung eingesetzt werden. Obwohl eine Einteilung der multivariaten Verfahren in „Grundlegende Verfahren“ und „Fortgeschrittene Verfahren“ weder leicht noch eindeutig ist, haben die Autoren dieses Buches eine solche Unterscheidung wie folgt vorgenommen:[4]

(1) Multivariate Analysemethoden: Grundlegende Verfahren

- Regressionsanalyse (Lineare Einfachregression und multiple Regression)
- Zeitreihenanalyse
- Varianzanalyse
- Diskriminanzanalyse
- Logistische Regression
- Kreuztabellierung und Kontingenzanalyse
- Faktorenanalyse
- Clusteranalyse
- (Traditionelle) Conjoint-Analyse

Die hier als „Grundlegende Verfahren“ bezeichneten multivariaten Analysemethoden werden in dem vorliegenden Buch ausführlich behandelt und jeweils an einem ausführlichen Fallbeispiel unter Verwendung von IBM SPSS erklärt.

(2) Multivariate Analysemethoden: Fortgeschrittene Verfahren

- Nichtlineare Regressionsanalyse
- Strukturgleichungsanalyse
- Konfirmatorische Faktorenanalyse
- Auswahlbasierte Conjoint-Analyse
- Neuronale Netze
- Multidimensionale Skalierung
- Korrespondenzanalyse

Die „Fortgeschrittenen Verfahren“ werden in diesem Buch nur in einer Kurzfassung behandelt, um so dem geneigten Leser einen Ausblick auf weitere gängige Verfahren der multivariaten Datenanalyse zu geben. Eine ausführliche Behandlung dieser Verfahren liefert das Buch:

> **Backhaus, Klaus/Erichson, Bernd/Weiber, Rolf:**
> Fortgeschrittene Multivariate Analysemethoden.
> Eine anwendungsorientierte Einführung, 3. Aufl., Berlin 2015.

[4] Einen Überblick über multivariate Analysemethoden geben auch die folgenden Bücher: Hair et al. (2010); Härdle/Simar (2015); Herrmann/Homburg/Klarmann (2008); Norusis/SPSS Inc. (2008); Schlittgen (2009).

Unabhängig von der obigen Zweiteilung nehmen wir im Folgenden eine Einordnung der multivariaten Analysemethoden nach ihrem Anwendungsbezug vor. Dabei sei jedoch betont, dass eine *überschneidungsfreie Zuordnung* der Verfahren zu praktischen Fragestellungen nicht immer möglich ist, da sich die Zielsetzungen der Verfahren z. T. überlagern. Versucht man jedoch eine Einordnung der Verfahren nach anwendungsbezogenen Fragestellungen, so bietet sich eine Einteilung in primär *strukturen-entdeckende Verfahren* und primär *strukturen-prüfende Verfahren* an. Diese beiden Kriterien werden in diesem Zusammenhang wie folgt verstanden:

1. *Strukturen-prüfende Verfahren* sind solche multivariaten Verfahren, deren primäres Ziel in der *Überprüfung von Zusammenhängen* zwischen Variablen liegt. Dabei wird überwiegend die kausale Abhängigkeit einer interessierenden Variablen von einer oder mehreren sog. unabhängigen Variablen (Einflussfaktoren) betrachtet. Der Anwender besitzt eine auf sachlogischen oder theoretischen Überlegungen basierende Vorstellung über die Zusammenhänge zwischen Variablen und möchte diese mit Hilfe multivariater Verfahren überprüfen.

   Strukturen-prüfende Verfahren

   Verfahren, die diesem Bereich der multivariaten Analyse zugeordnet werden können, sind die lineare und nichtlineare Regressionsanalyse, die Zeitreihenanalyse, die Varianzanalyse, die Diskriminanzanalyse, die Kontingenzanalyse sowie die Logistische Regression, die Strukturgleichungsanalyse und die Conjoint-Analyse zur Analyse von Präferenzstrukturen.

2. *Strukturen-entdeckende Verfahren* sind solche multivariaten Verfahren, deren Ziel in der *Entdeckung von Zusammenhängen* zwischen Variablen oder zwischen Objekten liegt. Der Anwender besitzt zu Beginn der Analyse noch keine Vorstellungen darüber, welche Beziehungszusammenhänge in einem Datensatz existieren.

   Strukturen-entdeckende Verfahren

   Verfahren, die primär eingesetzt werden, um mögliche Beziehungszusammenhänge aufzudecken, sind die Faktorenanalyse, die Clusteranalyse, die Multidimensionale Skalierung, die Korrespondenzanalyse und die Neuronalen Netze.

3.1 Strukturen-prüfende Verfahren

Die strukturen-prüfenden Verfahren werden primär zur Durchführung von *Kausalanalysen* eingesetzt, um herauszufinden, ob und wie stark sich z.B. das Wetter, die Bodenbeschaffenheit sowie unterschiedliche Düngemittel und -mengen auf den Ernteertrag auswirken oder wie stark die Nachfrage eines Produktes von dessen Qualität, dem Preis, der Werbung und dem Einkommen der Konsumenten abhängt.

Kausalanalysen

Voraussetzung für die Anwendung der entsprechenden Verfahren ist, dass der Anwender *a priori (vorab)* eine sachlogisch möglichst gut fundierte Vorstellung über den Kausalzusammenhang zwischen den Variablen entwickelt hat, d. h. er weiß bereits oder vermutet, welche der Variablen auf andere Variablen einwirken. Zur Überprüfung seiner (theoretischen) Vorstellungen werden die von ihm betrachteten Variablen i. d. R. in *abhängige* und *unabhängige* Variablen eingeteilt und dann mit Hilfe von multivariaten Analysemethoden an den empirisch erhobenen Daten überprüft. Nach dem Skalenniveau der Variablen lassen sich die grundlegenden strukturen-prüfenden Verfahren gemäß Abbildung 2 charakterisieren.

Hypothesen

		UNABHÄNGIGE VARIABLE	
		metrisches Skalenniveau	nominales Skalenniveau
ABHÄNGIGE VARIABLE	metrisches Skalennivau	Regressionsanalyse, Zeitreihenanalyse	Varianzanalyse, Regression mit Dummies
	nominales Skalenniveau	Diskriminanzanalyse, Logistische Regression	Kontingenzanalyse Auswahlbasierte Conjoint-Analyse

Abbildung 2: Grundlegende strukturen-prüfende Verfahren

Regressionsanalyse

Erklärung und Prognose

Die Regressionsanalyse ist ein außerordentlich vielseitiges und flexibles Analyseverfahren, das sowohl für die *Beschreibung* und *Erklärung von Zusammenhängen* als auch für die *Durchführung von Prognosen* große Bedeutung besitzt. Sie ist damit sicherlich das wichtigste und am häufigsten angewendete multivariate Analyseverfahren. Insbesondere kommt sie in Fällen zur Anwendung, wenn Wirkungsbeziehungen zwischen einer abhängigen und einer oder mehreren unabhängigen Variablen untersucht werden sollen. Mit Hilfe der Regressionsanalyse können derartige Beziehungen quantifiziert und damit weitgehend exakt beschrieben werden. Außerdem lassen sich mit ihrer Hilfe Hypothesen über Wirkungsbeziehungen prüfen und auch Prognosen erstellen.

Beispiel

Ein Beispiel bildet die Frage, ob und wie die Absatzmenge eines Produktes vom Preis, den Werbeausgaben, der Zahl der Verkaufsstätten und dem Volkseinkommen abhängt. Sind diese Zusammenhänge mit Hilfe der Regressionsanalyse quantifiziert und empirisch bestätigt worden, so lassen sich Prognosen (What-if-Analysen) erstellen, die beantworten, wie sich die Absatzmenge verändern wird, wenn z. B. der Preis oder die Werbeausgaben oder auch beide Variablen zusammen verändert werden.

Dummy Variablen

Die Regressionsanalyse ist prinzipiell anwendbar, wenn sowohl die abhängige als auch die unabhängigen Variablen metrisches Skalenniveau besitzen. Dies ist der klassische Fall, wobei die Beziehungen zwischen unabhängigen und abhängigen Variablen auch nicht linear sein können. Durch Anwendung der sog. *Dummy-Variablen-Technik* lassen sich aber auch qualitative (nominal skalierte) Variable in die Regressionsanalyse einbeziehen und deren Anwendungsbereich somit ausweiten. Dummy-Variable sind binäre Variable, die nur die Werte 0 oder 1 annehmen. Stellen wir uns vor, es sollen die Einflüsse verschiedener Produkteigenschaften auf das Kaufverhalten von Konsumenten untersucht werden. Die Dummy-Variable q_1 würde dann in allen Fällen, bei denen das Produkt eine rote Verpackung hat, den Wert 1 annehmen, und wenn dies nicht der Fall ist, den Wert 0.

$$q_1 = \begin{cases} 1 & falls \quad Farbe = rot \\ 0 & sonst \end{cases}$$

In analoger Weise lassen sich auch eine Dummy-Variable q_2 für die Farbe Gelb und eine Dummy-Variable q_3 für die Farbe Grün definieren. Wenn allerdings nur Verpackungen in den drei Farben Rot, Gelb und Grün vorkommen, so wäre eine der drei Dummies überflüssig. Denn wenn $q_1 = 0$ und $q_2 = 0$ gilt, so muss zwangsläufig $q_3 = 1$ gelten. Die drei Farben lassen sich also eindeutig mittels der zwei Dummies

(q_1, q_2) beschreiben: rot = (1, 0), gelb = (0, 1), grün = (0, 0). Generell gilt, dass sich eine nominale Variable mit n Ausprägungen durch $n - 1$ Dummy-Variablen ersetzen lässt.

Die Bedeutung von Dummy-Variablen liegt darin, dass sie sich wie metrische Variable behandeln lassen. Somit lassen sich mit ihrer Hilfe auch nominal skalierte Variable in eine Regressionsanalyse einbeziehen. Dies gilt aber generell nur für die unabhängigen Variablen und nicht für die abhängige Variable. Nachteilig ist, dass sich dadurch u. U. die Zahl der Variablen und der damit verbundene Kodierungs- und Rechenaufwand stark erhöht. Deshalb kann in solchen Fällen die Anwendung einer Varianzanalyse einfacher und übersichtlicher sein.

Zeitreihenanalyse

Die Zeitreihenanalyse dient neben der Beschreibung und Erklärung der zeitlichen Entwicklung einer Variablen insbesondere auch deren Prognose, d. h. der Schätzung von Werten dieser Variablen für zukünftige Zeitpunkte oder Perioden. Jede weitreichende Entscheidung basiert auf Prognosen. Die Zeitreihenanalyse ist daher für die Stützung von Entscheidungsproblemen jeglicher Art von großer Wichtigkeit. Für die Produktions- und Absatzplanung eines Herstellers ist z. B. von Wichtigkeit, wie sich seine Absatzmenge oder das Volumen seines Marktes langfristig entwickeln werden oder welchen periodischen Schwankungen diese Größen unterworfen sind. Neben anderen Verfahren bildet insbesondere die zuvor behandelte Regressionsanalyse ein wichtiges Instrument zur Durchführung von Zeitreihenanalysen. Sie ermöglicht die Erstellung von Punktprognosen sowie auch die Berechnung von Prognosefehlern und von Prognoseintervallen, innerhalb derer das vorhergesagte Ereignis mit einer festgelegten Wahrscheinlichkeit liegen wird.

Prognose

Varianzanalyse

Werden die unabhängigen Variablen auf nominalem Skalenniveau gemessen und die abhängigen Variablen auf metrischem Skalenniveau, so findet die Varianzanalyse Anwendung. Dieses Verfahren besitzt besondere Bedeutung für die *Analyse von Experimenten*, wobei die nominalen unabhängigen Variablen die experimentellen Einwirkungen repräsentieren. So kann z. B. in einem Experiment untersucht werden, welche Wirkung alternative Verpackungen eines Produktes oder dessen Plazierung im Geschäft auf die Absatzmenge haben.

Experimente

Diskriminanzanalyse

Ist die abhängige Variable nominal skaliert, und besitzen die unabhängigen Variablen metrisches Skalenniveau, so findet die Diskriminanzanalyse Anwendung. Die Diskriminanzanalyse ist ein Verfahren zur *Analyse von Gruppenunterschieden*. Ein Beispiel bildet die Frage, ob und wie sich die Wähler der verschiedenen Parteien hinsichtlich soziodemografischer und psychografischer Merkmale unterscheiden. Die abhängige nominale Variable identifiziert die Gruppenzugehörigkeit, hier die gewählte Partei, und die unabhängigen Variablen beschreiben die Gruppenelemente, hier die Wähler.

Gruppenunterschiede

Ein weiteres Anwendungsgebiet der Diskriminanzanalyse bildet die *Klassifizierung von Elementen*. Nachdem für eine gegebene Menge von Elementen die Zusammenhänge zwischen der Gruppenzugehörigkeit der Elemente und ihren Merkmalen analysiert wurden, lässt sich darauf aufbauend eine Prognose der Gruppenzugehörigkeit

Klassifizierung

von neuen Elementen vornehmen. Derartige Anwendungen finden sich z. B. bei der Kreditwürdigkeitsprüfung (Einstufung von Kreditkunden einer Bank in Risikoklassen) oder bei der Personalbeurteilung (Einstufung von Außendienstmitarbeitern nach erwartetem Verkaufserfolg).

Logistische Regression

Gruppenzugehörigkeit

Ganz ähnliche Fragestellungen, wie mit der Diskriminanzanalyse können auch mit dem Verfahren der logistischen Regression untersucht werden. Hier wird die *Wahrscheinlichkeit* der Zugehörigkeit zu einer Gruppe (einer Kategorie der abhängigen Variablen) in Abhängigkeit von einer oder mehreren unabhängigen Variablen bestimmt. Dabei können die unabhängigen Variablen sowohl nominales als auch metrisches Skalenniveau aufweisen. Über die Analyse der Gruppenunterschiede hinaus kann z. B. auch das Herzinfarktrisiko von Patienten in Abhängigkeit von ihrem Alter und ihrem Cholesterin-Spiegel ermittelt werden. Da zur Schätzung der Eintrittswahrscheinlichkeiten der Kategorien der abhängigen Variablen auf die (s-förmige) logistische Funktion zurückgegriffen wird, gehört dieses Verfahren zu den *nicht-linearen Analyseverfahren*.

Kreuztabellierung und Kontingenzanalyse

Kreuztabelle

Eine weitere Methodengruppe, die der Analyse von Beziehungen zwischen ausschließlich nominalen Variablen dient, wird als Kontingenzanalyse bezeichnet. Hier kann es z. B. darum gehen, die Frage nach dem Zusammenhang zwischen Rauchen (Raucher versus Nichtraucher) und Lungenerkrankung (ja, nein) statistisch zu überprüfen. Die Überprüfung erfolgt dabei auf der Basis von in Form einer Kreuztabelle (Kontingenztabelle) angeordneten Daten. Mit Hilfe weiterführender Verfahren, wie der sog. Logit-Analyse, lässt sich weiterhin auch die Abhängigkeit einer nominalen Variablen von mehreren nominalen Einflussgrößen untersuchen (vgl. hierzu auch das Verfahren der logistischen Regression).

Conjoint-Analyse

Analyse von Präferenzen

Bei den bisher aufgezeigten Verfahren wurde nur zwischen metrischem und nominalem Skalenniveau der Variablen unterschieden. Ein Verfahren, bei dem die abhängige Variable häufig auf ordinalem Skalenniveau gemessen wird, ist die Conjoint-Analyse. Insbesondere lassen sich mit Hilfe der Conjoint-Analyse ordinal gemessene Präferenzen und auch Auswahlentscheidungen (auswahlbasierte CA) analysieren. Ziel ist es dabei, den *Beitrag einzelner Merkmale* von Produkten oder sonstigen Objekten *zum Gesamtnutzen* bzw. zur Kaufentscheidung bzgl. dieser Objekte herauszufinden. Einen Anwendungsbereich bildet die Gestaltung neuer Produkte. Dazu ist es von Wichtigkeit, den Einfluss oder Beitrag alternativer Produktmerkmale, z. B. alternativer Materialien, Formen, Farben oder Preisstufen, auf die Nutzenbeurteilung zu kennen.

Bei der Conjoint-Analyse muss der Forscher vorab festlegen, welche Merkmale in welchen Ausprägungen berücksichtigt werden sollen. Hierauf basierend wird sodann ein Erhebungsdesign gebildet, im Rahmen dessen Präferenzen, z. B. bei potenziellen Käufern eines neuen Produktes, gemessen werden. Auf Basis dieser Daten erfolgt dann die Analyse zur Ermittlung der Nutzenbeiträge der berücksichtigten Merkmale und ihrer Ausprägungen. Die Conjoint-Analyse bildet damit also eine *Kombination aus Erhebungs- und Analyseverfahren*.

Nichtlineare Regression

Durch die Nichtlineare Regression wird das Anwendungsspektrum der Regressionsanalyse erheblich erweitert. Es lassen sich nahezu beliebige Modellstrukturen schätzen. Anwendungen finden sich z. B. im Rahmen der Werbewirkungsforschung (Abhängigkeit der Werbeerinnerung von der Zahl der Werbekontakte, Abhängigkeit der Absatzmenge von der Höhe des Werbebudgets) oder in der Marktforschung bei der Untersuchung des Wachstums von neuen Produkten. Die Nichtlineare Regression ist allerdings mit einer Reihe von Schwierigkeiten verbunden. Der Rechenaufwand ist um ein Vielfaches größer als bei der traditionellen Regressionsanalyse, da iterative Algorithmen für die Berechnung der Schätzwerte verwendet werden müssen. Ob diese Algorithmen konvergieren, hängt u. a. davon ab, welche Startwerte der Untersucher vorgibt. Es werden somit auch erhöhte Anforderungen an den Untersucher gestellt. Ein weiterer Nachteil ist, dass die statistischen Tests, die bei der linearen Regressionsanalyse zur Prüfung der Güte des Modells oder der Signifikanz der Parameter verwendet werden, für die nichtlineare Regression nicht anwendbar sind. Der Untersucher sollte daher, wenn möglich, der linearen Regressionsanalyse den Vorzug geben. Wie gezeigt werden wird, lassen sich auch mit Hilfe der linearen Regressionsanalyse vielfältige nichtlineare Problemstellungen behandeln.

Werbewirkungsforschung

Wachstumsmodelle

Strukturgleichungsmodelle und Konfirmatorische Faktorenanalyse

Die bisher betrachteten Analysemethoden gehen davon aus, dass alle Variablen in der Realität beobachtbar und gegebenenfalls auch messbar sind. Bei vielen theoriegestützten Fragestellungen hat man es aber auch mit nicht beobachtbaren Variablen zu tun, sog. *hypothetischen Konstrukten* oder *latenten Variablen*. Beispiele hierfür sind psychologische Konstrukte wie Einstellung und Motivation oder soziologische Konstrukte wie Kultur und soziale Schicht. In solchen Fällen kann die Analyse von Strukturgleichungen zur Anwendung kommen.

Latente Variable

Zur Behandlung von Strukturgleichungsmodellen wird in diesem Buch auf das Programmpaket AMOS (Analysis of Moment Structures) zurückgegriffen, das Datenmatrizen aus IBM SPSS analysieren und Ergebnisse mit IBM SPSS austauschen kann.[5] Mit Hilfe von AMOS lassen sich komplexe Kausalstrukturen überprüfen. Insbesondere können Beziehungen mit mehreren abhängigen Variablen, mehrstufigen Kausalbeziehungen und mit nicht beobachtbaren (latenten) Variablen überprüft werden. Der Benutzer muss, wenn er latente Variable in die Betrachtungen einbeziehen will, zwei Modelle spezifizieren:

AMOS

- Das *Messmodell*, das die Beziehungen zwischen den latenten Variablen und geeigneten Indikatoren vorgibt, mittels derer sich die latenten Variablen indirekt messen lassen. Die empirische Überprüfung erfolgt mit Hilfe der Konfirmatorischen Faktorenanalyse (KFA).
- Das *Strukturmodell*, welches die Kausalbeziehungen zwischen den latenten Variablen vorgibt, die letztlich dann zu überprüfen sind.

Messmodell

KFA

Die Variablen des Strukturmodells können alle latent sein, müssen es aber nicht. Ein Beispiel, bei dem nur die unabhängigen Variablen latent sind, wäre die Abhängig-

Strukturmodell

[5] Bis zur 9. Auflage wurde bei der Behandlung von Strukturgleichungsmodellen auf das Programm LISREL (LInear Structural RELationships) zurückgegriffen.

keit der Absatzmenge von der subjektiven Produktqualität und Servicequalität eines Anbieters.

Auswahlbasierte Conjoint-Analyse

CBC

Während bei der traditionellen Conjoint-Analyse zwecks Analyse von Nutzenstrukturen die Präferenzen von Probanden bezüglich alternativer Objekte (Stimuli) auf ordinalem Skalenniveau gemessen werden (mittels Ranking- oder Ratingskalen), erfolgt bei der Auswahlbasierten Conjoint-Analyse (Choice-Based Conjoint) nur eine Abfrage von Auswahlentscheidungen. Aus einer Menge von Alternativen (Choice Set) muss der Proband nur jeweils die am meisten präferierte Alternative auswählen, wobei meist auch die Option besteht, keine der Alternativen zu wählen. Dies ist für ihn nicht nur einfacher, sondern kommt auch seinem realen Entscheidungsverhalten (z. B. in Kaufsituationen) sehr viel näher, als das Ranking oder Rating aller Alternativen im Choice Set. Die erhöhte Realitätsnähe wird allerdings mit einem Verlust an Information erkauft, da bei dieser Vorgehensweise die Präferenz nur noch auf nominalem Skalenniveau gemessen wird. Zur Schätzung der Nutzenbeiträge einzelner Merkmale (Teilnutzenwerte) muss daher eine anderes Schätzverfahren verwendet werden. Während bei der traditionellen Conjoint-Analyse die Schätzung meist durch Regression mit Dummy-Variablen erfolgt, kommt bei der Auswahlbasierten Conjoint-Analyse die Maximum-Likelihood-Methode zur Anwendung. Dabei wird dem Verhalten der

Kaufsimulation

Probanden ein probabilistisches Entscheidungsmodell zugrunde gelegt. Wegen des geringeren Informationsgehalts ist es meist nur möglich, die Teilnutzenwerte aggregiert zu schätzen, während es bei der traditionellen Conjoint-Analyse üblich ist, sie individuell für jeden Probanden zu schätzen.

3.2 Strukturen-entdeckende Verfahren

Die hier den strukturen-entdeckenden Verfahren zugeordneten Analysemethoden werden primär zur *Entdeckung von Zusammenhängen* zwischen Variablen oder zwischen Objekten eingesetzt. Es erfolgt daher vorab durch den Anwender *keine* Zweiteilung der Variablen in abhängige und unabhängige Variablen, wie es bei den strukturenprüfenden Verfahren der Fall ist.

Faktorenanalyse

Die Faktorenanalyse findet insbesondere dann Anwendung, wenn im Rahmen einer Erhebung eine Vielzahl von Variablen zu einer bestimmten Fragestellung erhoben

Variablenbündelung

wurde, und der Anwender nun an einer Reduktion bzw. *Bündelung der Variablen* interessiert ist. Von Bedeutung ist die Frage, ob sich möglicherweise sehr zahlreiche Merkmale, die zu einem bestimmten Sachverhalt erhoben wurden, auf einige wenige „zentrale Faktoren“ zurückführen lassen. Ein einfaches Beispiel hierzu bildet die Verdichtung der zahlreichen technischen Eigenschaften von Kraftfahrzeugen auf wenige Dimensionen, wie Größe, Leistung und Sicherheit.

Positionierung

Einen wichtigen Anwendungsbereich der Faktorenanalyse bilden *Positionierungsanalysen*. Dabei werden die subjektiven Eigenschaftsbeurteilungen von Objekten (z. B. Produktmarken, Unternehmen oder Politiker) mit Hilfe der Faktorenanalyse auf zugrundeliegende Beurteilungsdimensionen verdichtet. Ist eine Verdichtung auf zwei oder drei Dimensionen möglich, so lassen sich die Objekte im Raum dieser Dimensionen grafisch darstellen. Im Unterschied zu anderen Formen der Positionierungsanalyse spricht man hier von faktorieller Positionierung.

Clusteranalyse

Während die Faktorenanalyse eine Verdichtung oder Bündelung von Variablen vornimmt, wird mit der Clusteranalyse eine *Bündelung von Objekten* angestrebt. Das Ziel ist dabei, die Objekte so zu Gruppen (Clustern) zusammenzufassen, dass die Objekte in einer Gruppe möglichst ähnlich und die Gruppen untereinander möglichst unähnlich sind. Beispiele sind die Bildung von Persönlichkeitstypen auf Basis der psychografischen Merkmale von Personen oder die Bildung von Marktsegmenten auf Basis nachfragerelevanter Merkmale von Käufern.

Objektbündelung

Zur Überprüfung der Ergebnisse einer Clusteranalyse kann die Diskriminanzanalyse herangezogen werden. Dabei wird untersucht, inwieweit bestimmte Variablen zur Unterscheidung zwischen den Gruppen, die mittels Clusteranalyse gefunden wurden, beitragen bzw. diese erklären.

Überprüfung

Neuronale Netze

Neuronale Netze werden heute in der Praxis in zunehmendem Maße sowohl ergänzend zu den klassischen multivariaten Methoden eingesetzt, als auch in den Fällen, in den die klassischen Methoden versagen. Anwendungsgebiete sind Klassifikationen von Objekten, Prognosen von Zuständen oder Probleme der Gruppenbildung. Insofern bestehen hinsichtlich der Aufgabenstellungen Ähnlichkeiten zur Diskriminanzanalyse und zur logistischen Regression wie auch zur Clusteranalyse. Die Methodik neuronaler Netze lehnt sich an biologische Informationsverarbeitungsprozesse im Gehirn an (daher der Name). Es werden künstliche neuronale Netze gebildet, die in der Lage sind, selbständig aus Erfahrung zu lernen. Insbesondere vermögen sie, komplexe Muster in vorhandenen Daten (z. B. Finanzdaten, Verkaufsdaten) zu erkennen und eröffnen so eine sehr einfache Form der Datenanalyse. Besonders vorteilhaft lassen sie sich zur Behandlung von schlecht strukturierten Problemstellungen einsetzen.

Substitut

Gehirn

Innerhalb neuronaler Netze werden künstliche Neuronen (Nervenzellen) als Grundelemente der Informationsverarbeitung in Schichten organisiert, wobei jedes Neuron mit denen der nachgelagerten Schicht verbunden ist. Dadurch lassen sich auch hochgradig nicht-lineare und komplexe Zusammenhänge ohne spezifisches Vorwissen über die etwaige Richtung und das Ausmaß der Wirkungsbeziehungen zwischen einer Vielzahl von Variablen modellieren.

Nicht-Linearität

Zum Erlernen von Strukturen wird das Netz zunächst in einer sog. *Trainingsphase* mit beobachteten Daten „gefüttert“. Dabei wird unterschieden zwischen Lernprozessen, bei denen die richtigen Ergebnisse bekannt sind und diese durch das Netz reproduziert werden sollen (*überwachtes Lernen*), und solchen, bei denen die richtigen Ergebnisse nicht bekannt sind und lediglich ein konsistentes Verarbeitungsmuster erzeugt werden soll (*unüberwachtes Lernen*). Nach der Trainingsphase ist das Netz konfiguriert und kann für die Analyse neuer Daten eingesetzt werden.

Überwachtes vs. unüberwachtes Lernen

Multidimensionale Skalierung

Den Hauptanwendungsbereich der Multidimensionalen Skalierung (MDS) bilden Positionierungsanalysen, d. h. die *Positionierung von Objekten im Wahrnehmungsraum* von Personen. Sie bildet somit eine Alternative zur faktoriellen Positionierung mit Hilfe der Faktorenanalyse.

Positionierung

Im Unterschied zur faktoriellen Positionierung werden bei Anwendung der MDS nicht die subjektiven Beurteilungen von Eigenschaften der untersuchten Objekte er-

Ähnlichkeiten

hoben, sondern es werden nur wahrgenommene globale Ähnlichkeiten zwischen den Objekten erfragt. Mittels der MDS werden die diesen Ähnlichkeiten zugrundeliegenden Wahrnehmungsdimensionen abgeleitet. Wie schon bei der faktoriellen Positionierung lassen sich sodann die Objekte im Raum dieser Dimensionen positionieren und grafisch darstellen. Die MDS findet insbesondere dann Anwendung, wenn der Forscher keine oder nur vage Kenntnisse darüber hat, welche Eigenschaften für die subjektive Beurteilung von Objekten (z. B. Produktmarken, Unternehmen oder Politiker) von Relevanz sind.

Beziehung zu anderen Verfahren

Zwischen der Multidimensionalen Skalierung und der Conjoint-Analyse besteht sowohl inhaltlich wie auch methodisch eine enge Beziehung, obgleich wir sie hier unterschiedlich zum einen den strukturen-entdeckenden und zum anderen den strukturenprüfenden Verfahren zugeordnet haben. Beide Verfahren befassen sich mit der Analyse psychischer Sachverhalte und bei beiden Verfahren können auch ordinale Daten analysiert werden, weshalb sie z. T. auch identische Algorithmen verwenden. Ein gewichtiger Unterschied besteht dagegen darin, dass der Forscher bei Anwendung der Conjoint-Analyse bestimmte Merkmale auszuwählen hat.

Korrespondenzanalyse

Visualisierung

Die Korrespondenzanalyse dient, wie auch die Faktorenanalyse und die Multidimensionale Skalierung (MDS), zur Visualisierung komplexer Daten. Sie wird daher in der Marktforschung ebenfalls zur Durchführung von Positionierungsanalysen verwendet. Insbesondere kann sie als ein Verfahren der multidimensionalen Skalierung von nominal skalierten Variablen charakterisiert werden. Sie ermöglicht es, die Zeilen und Spalten einer zweidimensionalen Kreuztabelle (Kontingenztabelle) grafisch in einem gemeinsamen Raum darzustellen.

Beispiel

Beispiel: Gegeben sei eine Häufigkeitstabelle, deren Zeilen Automarken betreffen und in deren Spalten wünschenswerte Merkmale von Autos (z. B. hohe Sicherheit, schönes Design) stehen. Die Zellen der Matrix sollen beinhalten, mit welcher Häufigkeit ein bestimmtes qualitatives Merkmal den verschiedenen Automarken im Rahmen einer Käuferbefragung zugeordnet wurde. Marken und Merkmale lassen sich sodann mit Hilfe der Korrespondenzanalyse in einem gemeinsamen Raum als Punkte darstellen. Dadurch lässt sich dann erkennen, wie die Automarken relativ zueinander und in Bezug auf die Merkmale von den Käufern beurteilt werden. Für die Korrespondenzanalyse spielt es dabei *keine* Rolle (im Unterschied zur Faktorenanalyse), welche Elemente in den Zeilen und welche in den Spalten angeordnet werden.

Vorteil

Ein besonderer Vorteil der Korrespondenzanalyse liegt darin, dass sie kaum Ansprüche an das Skalenniveau der Daten stellt. Die Daten müssen lediglich nichtnegativ sein. Die Korrespondenzanalyse kann daher auch zur Quantifizierung qualitativer Daten verwendet werden. Da sich qualitative Daten leichter erheben lassen als quantitative Daten, kommt diesem Verfahren eine erhebliche praktische Bedeutung zu.

3.3 Zusammenfassende Betrachtung

Die vorgenommene Zweiteilung der multivariaten Verfahren in strukturen-prüfende und strukturen-entdeckende Verfahren kann keinen Anspruch auf Allgemeingültigkeit erheben, sondern kennzeichnet nur den vorwiegenden Einsatzbereich der Verfahren. So kann und wird auch die Faktorenanalyse zur Überprüfung von hypothetisch ge-

bildeten Strukturen eingesetzt, und viel zu häufig werden in der empirischen Praxis auch Regressions- und Diskriminanzanalyse im heuristischen Sinne zur Auffindung von Kausalstrukturen verwendet. Diese Vorgehensweise wird nicht zuletzt auch durch die Verfügbarkeit leistungsfähiger Rechner und Programme unterstützt. Der gedankenlose Einsatz von multivariaten Verfahren kann leicht zu einer Quelle von Fehlinterpretationen werden, da ein statistisch signifikanter Zusammenhang keine hinreichende Bedingung für das Vorliegen eines kausal bedingten Zusammenhangs bildet. („Erst denken, dann rechnen!“) Es sei daher generell empfohlen, die strukturen-prüfenden Verfahren auch in diesem Sinne, d. h. zur empirischen Überprüfung von theoretisch oder sachlogisch begründeten Hypothesen, einzusetzen. In Abbildung 3 sind die oben skizzierten multivariaten Verfahren noch einmal mit jeweils einem Anwendungsbeispiel zusammengefasst.

Fehlinterpretation

Verfahren	Beispiel
Regressionsanalyse	Abhängigkeit der Absatzmenge eines Produktes von Preis, Werbeausgaben und Einkommen.
Zeitreihenanalyse	Analyse und Prognose der zeitlichen Entwicklung des Absatzvolumens eines Produktes oder eines Marktes.
Varianzanalyse	Wirkung alternativer Verpackungsgestaltungen auf die Absatzmenge eines Produktes.
Diskriminanzanalyse	Unterscheidung der Wähler der verschiedenen Parteien hinsichtlich soziodemografischer und psychografischer Merkmale.
Logistische Regression	Ermittlung des Herzinfarktrisikos von Patienten in Abhängigkeit ihres Alters und ihres Cholesterin-Spiegels.
Kontingenzanalyse	Zusammenhang zwischen Rauchen und Lungenerkrankung.
Faktorenanalyse	Verdichtung einer Vielzahl von Eigenschaftsbeurteilungen auf zugrundeliegende Beurteilungsdimensionen.
Clusteranalyse	Bildung von Persönlichkeitstypen auf Basis der psychografischen Merkmale von Personen.
Conjoint-Analyse	Ableitung der Nutzenbeiträge alternativer Materialien, Formen u. Farben von Produkten zur Gesamtpräferenz.
Nichtlineare Regression	Untersuchung des Wachstums von neuen Produkten, der Diffusion von Innovationen oder der Ausbreitung von Epidemien.
Strukturgleichungsanalyse	Abhängigkeit der Käufertreue von der subjektiven Produktqualität und Servicequalität eines Anbieters.
Konfirmatorische Faktorenanalyse	Überprüfung der Eignung vorgegebener Indikatorvariablen für die Messung von hypothetischen Konstrukten wie z. B. Einstellung, Kaufabsicht, Loyalität, Vertrauen oder Reputation.
Neuronale Netze	Untersuchung von Aktienkursen und möglichen Einflussfaktoren zwecks Prognose von Kursentwicklungen.
Multidimensionale Skalierung	Positionierung von konkurrierenden Produktmarken im Wahrnehmungsraum der Konsumenten.
Korrespondenzanalyse	Darstellung von Produktmarken und Produktmerkmalen in einem gemeinsamen Raum.
Auswahlbasierte Conjoint-Analyse	Schätzung der Nutzenbeiträge einzelner Merkmale von Produkten zur Gesamtpräferenz auf Basis simulierter Kaufentscheidungen.

Abbildung 3: Synopsis der multivariaten Analysemethoden

4 Zur Verwendung von IBM SPSS

IBM SPSS

Zur rechnerischen Durchführung der Analysen, die in diesem Buch behandelt werden, wurde vornehmlich das Programmsystem *IBM SPSS Statistics* oder kurz *SPSS* verwendet, da dieses in Wissenschaft und Praxis eine besonders große Verbreitung gefunden hat.[6] Der Name „SPSS“ stand ursprünglich als Akronym für „*S*tatistical *P*ackage for the *S*ocial *S*ciences“.[7] Der Anwendungsbereich von SPSS wurde im Laufe der Zeit ständig erweitert und erstreckt sich inzwischen auf nahezu alle Bereiche der Datenanalyse. „SPSS“ steht daher heute als Markenname, der nicht weiter interpretiert wird.

Grafische Benutzeroberfläche

Während SPSS ursprünglich nur über eine Kommandodatei gesteuert werden konnte (Batch-Betrieb), besitzt es heute eine grafische Benutzeroberfläche, über die der Benutzer mit dem Programm kommunizieren kann (Dialog-Betrieb). Diese Benutzeroberfläche wird ständig verbessert und erweitert. Über die dort vorhandenen Menüs und Dialogfelder lassen sich auch komplexe Analysen sehr bequem durchführen. Die früher zur Steuerung des Programms benötigte Kommandosprache (Befehlssyntax) findet daher zunehmend weniger Anwendung, ist aber nicht überflüssig geworden. Intern wird sie weiterhin verwendet und auch für den Benutzer besitzt sie gewisse Vorteile. In den einzelnen Kapiteln sind daher auch jeweils die erforderlichen Kommando-Sequenzen zum Nachvollzug der Analysen wiedergegeben. An dieser Stelle sollen in sehr kurzer Form einige allgemeine Hinweise zur Handhabung von IBM SPSS angeführt werden. Bezüglich näherer Ausführungen muss auf die einschlägige Literatur verwiesen werden.[8]

4.1 Die Daten

Daten

Die Datenanalyse mit IBM SPSS setzt voraus, dass die Daten in Form einer *Matrix* angeordnet werden (vgl. Abbildung 4). IBM SPSS erwartet, dass die *Spalten der Matrix* sich auf *Variablen* (variables), z. B. Eigenschaften, Merkmale, Dimensionen, beziehen. Die *Zeilen der Matrix* bilden *Beobachtungen bzw. Fälle* (cases), die sich auf unterschiedliche Personen, Objekte oder Zeitpunkte beziehen können. Ein kleines Beispiel zeigt Abbildung 5.

4.1.1 Der Daten-Editor

Daten-Editor

Der Daten-Editor dient der Eingabe der zu analysierenden Daten in IBM SPSS. Neben der Erstellung neuer Datensätze können hier aber auch bereits bestehende Datensätze modifiziert werden. Abbildung 6 zeigt zunächst den Aufbau des Daten-Editors. Er besteht ähnlich einem Tabellenkalkulationsblatt (Spreadsheet) aus Zeilen und Spalten.

[6] Das Programmsystem IBM SPSS Statistics kann unter den Betriebssystemen Windows, Macintosh und Linux verwendet werden und ist zur Zeit in der Version 25 auf dem Markt. Es umfasst ein Basismodul und diverse Erweiterungsmodule. Neben der Vollversion von IBM SPSS Statistics Base wird zu Lehrzwecken auch eine preiswertere Studentenversion angeboten. Diese weist einige Einschränkungen auf, die aber für die Mehrzahl der Nutzer kaum relevant sein dürften: Datendateien dürfen maximal 50 Variablen und 1500 Fälle enthalten und die SPSS-Befehlssyntax (Kommandosprache) sowie die Erweiterungsmodule sind nicht verfügbar.

[7] Zeitweilig wurde IBM SPSS auch interpretiert als „*S*tatistical *P*roduct and *S*ervice *S*olutions“ oder „Superior Performing Software System“. Nachdem IBM die Firma SPSS Inc. im Jahr 2009 übernommen hatte, wurde das Programmsystem kurzzeitig in der Version 18.0 unter dem Namen PASW (Predictive Analysis SoftWare) angeboten.

[8] Vgl. hierzu die Handbücher zu IBM SPSS 25, die als PDF-Dateien mit dem Programm ausgeliefert werden sowie z. B. Bühl (2014) oder Janssen/Laatz (2017).

Fälle k	Variablen j 1	2	3	.....	J
1	x_{11}	x_{21}	x_{31}	.....	X_{J1}
2	x_{12}	x_{22}	x_{32}	.....	X_{J2}
.	.				.
.	.				.
.	.		Werte x_{jk}		.
.	.				.
.	.				.
K	x_{1K}	x_{2K}	x_{3K}	.....	X_{JK}

Abbildung 4: Datenmatrix

Person	Geschlecht	Größe [cm]	Gewicht [kg]
1	1	178	68
2	0	166	50
3	1	183	75
4	0	168	52
5	1	195	100
6	1	175	73

Abbildung 5: Beispiel einer Datenmatrix

Die einzelnen Zeilen entsprechen dabei den Beobachtungen bzw. Fällen (z. B. Personen, Objekte) und die Spalten den Variablen (Merkmalen). In die einzelnen Felder (Zellen) sind für jeden Fall die jeweiligen Messwerte der entsprechenden Variablen einzugeben. Für das Beispiel aus Abbildung 5 sind dies die vier Variablen Person, Geschlecht, Größe und Gewicht. Zunächst geben wir in die erste Spalte die Nummern der sechs Personen ein, wodurch eine Variable VAR00001 erzeugt wird.

Beispiel

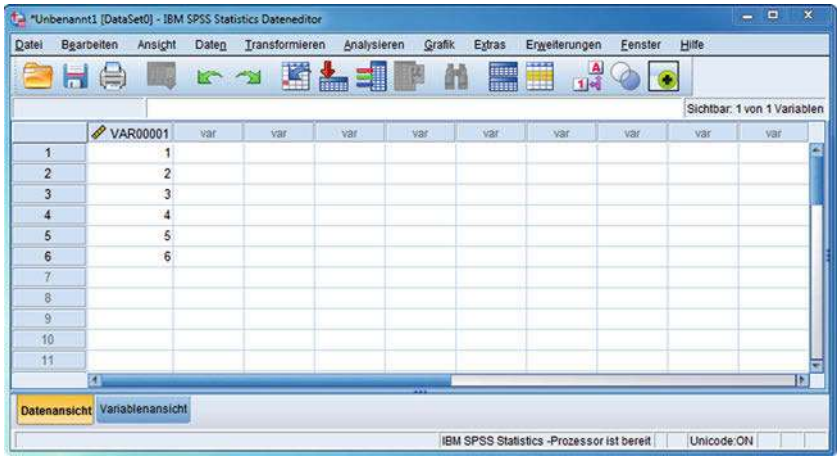


Abbildung 6: Der Daten-Editor

Über dem Eingabefeld enthält der Daten-Editor eine *Menüleiste* mit den Optionen „Datei“, „Bearbeiten“, „Ansicht“ etc. Durch Klick auf diese Optionen öffnen sich jeweils weitere Menüs, auf deren Anwendung bzw. Nutzung noch eingegangen wird.

4.1.2 Erstellung einer neuen Datendatei

4.1.2.1 Variablen definieren

Variablendefinition

Bevor mit der Eingabe der zu analysierenden Daten in den Daten-Editor begonnen wird, ist es zweckmäßig, die relevanten Variablen (hier Person, Geschlecht, Größe, Gewicht) zu definieren. Der Eintrag „VAR“ in den jeweiligen Spaltenköpfen zeigt zunächst an, dass für die entsprechende Spalte noch keine Variable definiert wurde.

Zur Definition der Variablen wechseln wir von der *Datenansicht* (Abbildung 6) in die *Variablenansicht*(Abbildung 7) des Dateneditors. Dies erfolgt z. B. durch Klick auf die Registerkarte „Variablenansicht“ (Abbildung 6 links unten).[9]

Folgende Eigenschaften können in der Variablenansicht zwecks Definition einer Variablen festgelegt werden:

- Variablenname,
- Variablentyp (einschließlich Spaltenformat und Dezimalstellen),
- Variablen- und Wertelabels,
- fehlende Werte,
- Spaltenbreite und Ausrichtung
- Messniveau.

Zunächst werden für diese Spezifikationen in der Variablenansicht die Voreinstellungen angezeigt.

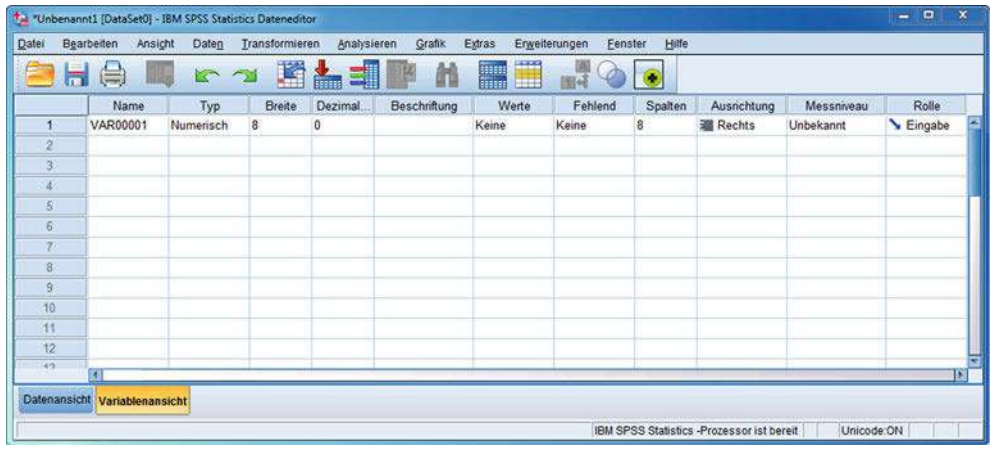


Abbildung 7: Die Variablenansicht

[9] Alternativ kann dies durch
- Aufruf der Option „Ansicht/ Variablen“ aus der Menüleiste (oben)
- Doppelklick auf einen Variablennamen im Spaltenkopf

erfolgen.

Variablenname

Im Eingabefeld „Name“ kann der Variablen ein Name zugewiesen werden. Wir tragen hier die Namen „Person“, „Geschlecht“, „Größe“ und „Gewicht“ ein (vgl. Abbildung 8).

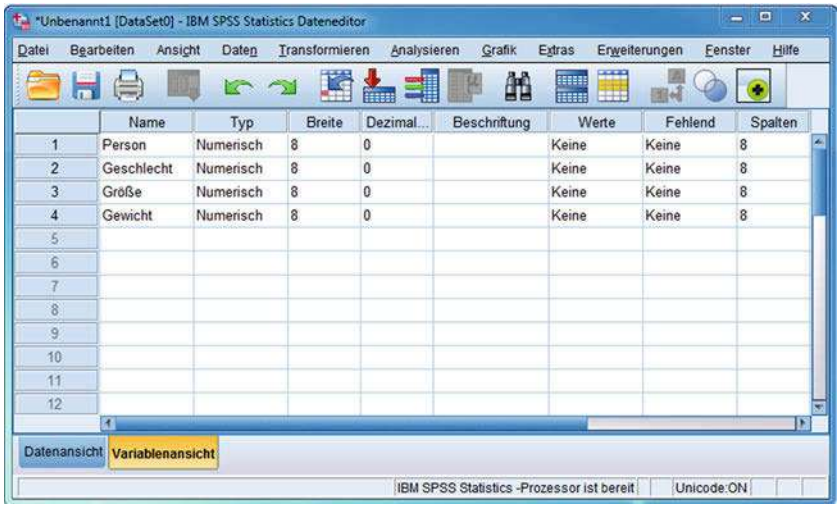


Abbildung 8: Variablenansicht mit Variablennamen

Bei der Benennung von Variablen sind einige Beschränkungen zu beachten:

- Variablennamen müssen eindeutig sein (dürfen nicht doppelt vorkommen). Variablenname
- Der Name muss mit einem Buchstaben beginnen.
- Der Name darf maximal 64 Zeichen umfassen.
- Der Name kann aus Buchstaben und Ziffern sowie einigen Sonderzeichen (_, ., $, @, #) gebildet werden. Er darf keine Leerzeichen enthalten.

Variablentyp

Für jede Variable lässt sich durch Klick auf die Schaltfläche in der Spalte „Typ“ das Menü „Variablentyp definieren“ öffnen, in dem diverse Spezifikationen vorgenommen werden können (vgl. Abbildung 9). Standardmäßig wird der Variablentyp „numerisch“ festgelegt (default). Alternativ können aber auch z. B. Datums- und Währungsformate oder auch Stringformate[10] gewählt werden. Variablentyp

Je nach gewähltem Typ können für eine Variable zusätzlich unterschiedliche Spezifikationen vorgenommen werden. So kann z. B. beim Typ „numerisch“ die Breite (maximal 40 Zeichen) und die Anzahl der Dezimalstellen (maximal 16) festgelegt werden. Voreingestellt sind die Werte 8 (Breite) und 0 (Dezimalstellen). Diese Spezifikationen werden auch in den Spalten „Spaltenformat“ und „Dezimalstellen“ der Variablenansicht angezeigt und können alternativ auch hier verändert werden.

Über die Schaltfläche „OK“ werden die vorgenommenen Einstellungen bezüglich der betreffenden Variablen aktiviert.

[10] Stringvariablen arbeiten mit Zeichenketten. Gültige Werte umfassen Buchstaben, Ziffern und Sonderzeichen.

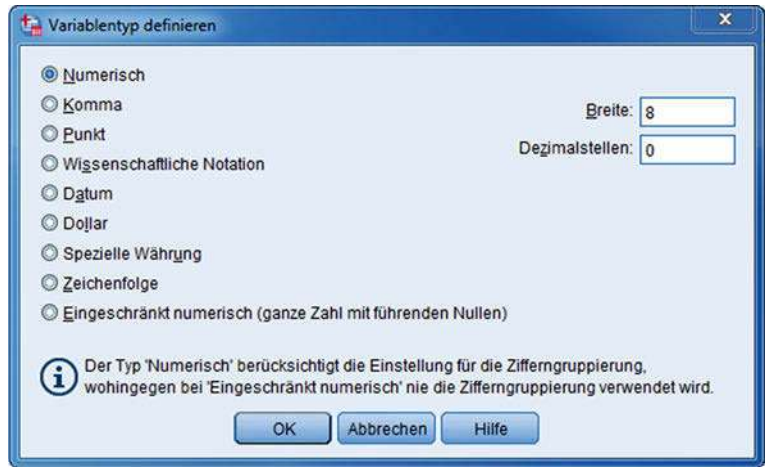


Abbildung 9: Dialogfenster „Variablentyp definieren:“

Variablenbeschriftung

Zusätzlich zum Variablennamen (der früher nur 8 Zeichen umfassen durfte gegenüber heute 64 Zeichen) lässt sich zur Beschreibung einer Variablen noch eine Beschriftung (Variablenlabel) angeben, welche maximal 256 Zeichen umfassen und auch Leerzeichen enthalten darf. So ließe sich der Variablen Größe die Beschriftung „Größe in cm“ und der Variablen Gewicht die Beschriftung „Gewicht in kg“ zuordnen. Statt dessen ließen sich aber unter Verwendung von Unterstrichen auch die Namen „Größe_in_cm“ und „Gewicht_in_kg“ vergeben, d. h. durch die Erweiterung von Namen auf 64 Zeichen sind zusätzliche Beschriftungen weitgehend entbehrlich geworden. Wenn Variablenbeschriftungen vorhanden sind, werden diese in den Dialogfeldern vor dem Variablennamen angezeigt.[11] Deshalb sollte man zwecks Übersichtlichkeit die Variablenbeschriftungen (wie auch die Namen) möglichst kurz und prägnant formulieren.

Wertebeschriftung

Auch einzelnen Ausprägungen einer Variablen lassen sich sog. „Wertebeschriftungen“ zuordnen. Dies ist insbesondere bei der numerisch kodierten Eingabe von nominalen Variablen nützlich, damit später nachvollzogen werden kann, wie die Kodierung erfolgte (z. B. Geschlecht: 0 = weiblich, 1 = männlich). Hierzu dient das Dialogfenster „Wertebeschriftungen“, das sich bei Anklicken des Feldes „Werte“ für die betreffende Variable öffnet (vgl. Abbildung 10). Wertebeschriftungen dürfen maximal 120 Zeichen (früher 20 Zeichen) umfassen.

[11] Über den Menüpunkt „Bearbeiten/ Optionen“ lässt sich unter der Registerkarte „Allgemein“ einstellen, ob die Variablenbeschriftungen oder nur die Namen angezeigt werden sollen (vgl. Abbildung 29). Über diesen Menüpunkt eröffnen sich dem Benutzer vielfältige Möglichkeiten zur Anpassung des Programms an individuelle Bedürfnisse.

Abbildung 10: Dialogfenster „Wertebeschriftungen“

Fehlende Werte

Missing Values

Ein Problem, das bei der praktischen Anwendung statistischer Methoden häufig auftritt, bilden fehlende Werte *(missing values)*. Hierbei handelt es sich z. B. um Variablenwerte, die bei einer Befragung von den Befragten entweder außerhalb des zulässigen Beantwortungsintervalls oder überhaupt nicht angegeben wurden. So kann zum Beispiel eine „0“ für das Gewicht einer Person bedeuten, dass der Wert nicht bekannt ist. Würde man jetzt den Mittelwert der Variablen berechnen, so erhielte man einen zu niedrigen Wert. Um eine derartige Fehlinterpretation zu vermeiden, kann der Benutzer dem IBM SPSS-Programm über die Schaltfläche „Fehlende Werte“ mitteilen, dass „0“ einen fehlenden Wert bedeutet, der dann bei der Berechnung von Statistiken übergangen wird. Man spricht in diesem Fall von *benutzerdefinierten fehlenden Werten*. Im Dialogfenster „Fehlende Werte“ (vgl. Abbildung 11) stehen für jede Variable drei Optionen zur Verfügung:

User-missing

Optionen

- keine fehlenden Werte (keine benutzerdefinierten fehlenden Werte),
- einzelne fehlende Werte (Eingabe von bis zu drei einzelnen Werten möglich, die als fehlende Werte behandelt werden sollen, z. B. „0“ für fehlende Angabe, „-1“ für falsche Angabe),
- Bereich und einzelner fehlender Wert (Eingabe eines Wertebereiches für fehlende Werte und eines einzelnen Wertes außerhalb dieses Bereiches). Diese Option ist nur für numerische Variablen verfügbar.

Listwise Deletion

Fälle mit fehlenden Werten werden bei Analysen mit IBM SPSS standardmäßig übergangen (Listenweiser Fallausschluss). Bei den einzelnen Analyseverfahren bestehen z. T. auch andere Optionen. So können z. B. fehlende Werte durch den Mittelwert der Variablen ersetzt werden oder innerhalb von Zeitreihen durch Interpolation geschätzt werden.[12]

[12] Die Ersetzung (Imputation) fehlender Werte in Zeitreihen kann mittels des Menüpunktes „Transformation/ fehlende Werte ersetzen“ erfolgen. Mit dem SPSS-Modul „Missing Value Analysis“ können Muster von fehlenden Daten ermittelt und fehlende Werte mittels unterschiedlicher Methoden geschätzt und ersetzt werden. Vgl. dazu SPSS Inc. (2007a).

System-missing

Neben den benutzerdefinierten fehlenden Werten kennt IBM SPSS noch eine zweite Art von fehlenden Werte, nämlich *systemdefinierte fehlende Werte*. Diese entstehen, wenn Felder im Daten-Editor leer gelassen wurden bzw. der Eintrag nicht dem Variablenformat entspricht. Für den Benutzer werden diese Werte durch einen Punkt in dem entsprechenden Feld kenntlich gemacht.[13] Systemdefinierte fehlende Werte werden standardmäßig, wenn nichts anderes vereinbart wird, wie benutzerdefinierte fehlende Werte behandelt.

Abbildung 11: Dialogfenster „Fehlende Werte“

Da in unserem kleinen Beispiel sämtliche Variablenwerte vorliegen, kann die Voreinstellung „Keine fehlenden Werte“ beibehalten werden.

Spaltenbreite und Ausrichtung

Voreinstellung

Die Schaltflächen „Spalten“ und „Ausrichtung“ betreffen die Anzeige der Daten im Dateneditor. Mittels „Spalten“ kann für jede Variable die Spaltenbreite (Anzahl der angezeigten Zeichen) und mittels „Ausrichtung“ die Textausrichtung (linksbündig, rechtsbündig oder zentriert) festgelegt werden. Voreingestellt sind eine Spaltenbreite von 8 Zeichen und rechtsbündige Ausrichtung für numerische Variablen. Die Spaltenbreite lässt sich auch in der Datenansicht ändern, indem man auf eine Spaltenbegrenzung klickt und sie an die gewünschte Stelle zieht.

Messniveau

Skalenniveau

Schließlich lässt sich noch das „Messniveau“ (Skaleniveau) der Variablen (metrisch, ordinal oder nominal) spezifizieren. Voreingestellt ist das Skalenniveau „metrisch“, das inzwischen umbenannt wurde in „Skala“. Dies trifft für die Variablen Größe und Gewicht zu. Die Variable Geschlecht hat dagegen das Skalenniveau „nominal“ und für die Variable Person lässt sich das Skalenniveau „ordinal“ angeben. Die drei Skalenniveaus werden durch Symbole (Icons) verdeutlicht (vgl. Abbildung 12, vorletzte Spalte).

[13] Allerdings gilt dies nicht für String-Variablen, da diese auch einen leeren Eintrag enthalten können.

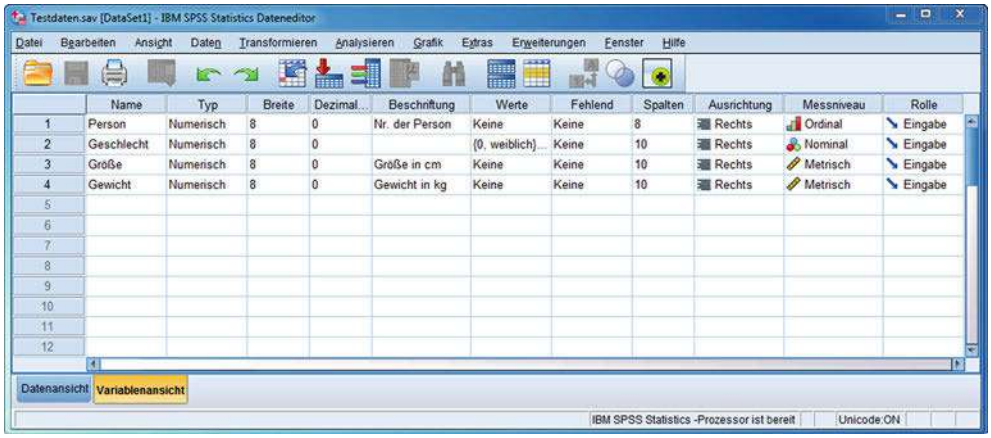


Abbildung 12: Variablenansicht nach Spezifikation der Variablen

4.1.2.2 Dateneingabe

Nachdem die Variablen definiert wurden, können die Daten direkt in den Daten-Editor eingegeben werden. Dabei kann man sowohl fall- als auch variablenweise vorgehen. Das jeweils aktive Feld, in das ein Wert eingegeben werden kann, ist durch eine starke Umrandung hervorgehoben. Die Korrektur von bereits vorhandenen Daten erfolgt im Feld-Editor der Bearbeitungszeile, die sich über den einzelnen Spalten befindet. Hier werden auch die entsprechende Zeilennummer und der Variablenname ausgewiesen (vgl. Abbildung 13). Mit dem Befehl „Ansicht/Wertelabels“ kann bei der Darstellung der Daten im Daten-Editor zwischen numerischer Kodierung und Wertelabels gewechselt werden. Bei der Eingabe der Daten ist zu beachten, dass nur Werte entsprechend dem definierten Variablentyp eingegeben werden können. Das heisst, dass beispielsweise beim Variablentyp „numerisch“ keine Buchstaben eingegeben werden können. Die Zulässigkeit überprüft IBM SPSS bereits während der Eingabe, indem unzulässige Zeichen gar nicht erst aufgenommen werden.

Dateneingabe

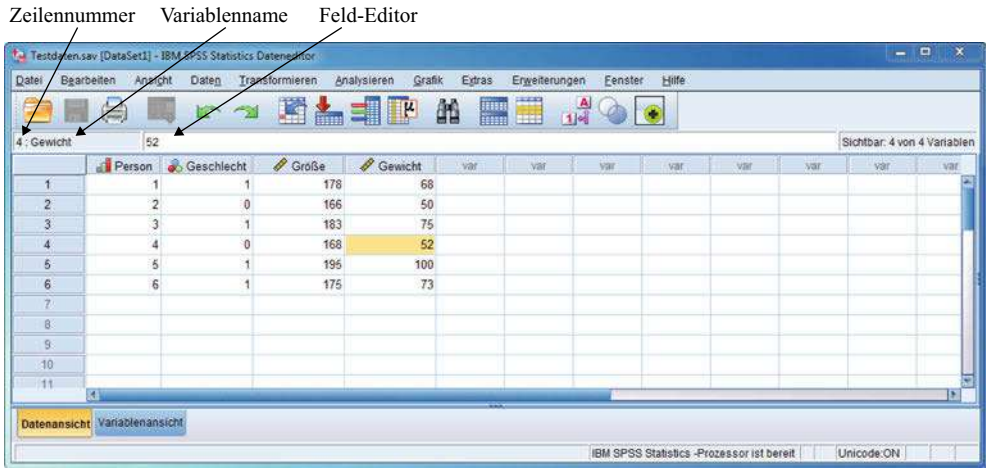


Abbildung 13: Aufbau des Daten-Editors

Nachdem die neuen Daten in den Daten-Editor eingegeben oder eine bereits bestehende Datei geändert wurde, muss die Datei vor dem Schließen bzw. dem Beenden von IBM SPSS gespeichert werden. Hierzu ist aus dem Menü der Befehl „Datei/ Speichern unter...“ auszuwählen. Es wird die Dialogbox „Daten speichern als“ geöffnet, über die die Datei unter Angabe eines Dateinamens, z. B. „Testdaten“, gespeichert werden kann. Die für Datendateien erforderliche Erweiterung .sav wird von IBM SPSS automatisch vergeben.

4.1.3 Einlesen einer vorhandenen Datendatei

Häufig sind die zu analysierenden Daten bereits in einer Datendatei vorhanden und müssen nicht erst mühselig, wenn es sich um große Datenmengen handelt, mit dem IBM SPSS-Dateneditor eingegeben werden. Datendateien können in sehr unterschiedlichen Formaten vorliegen, z. B.

Formate

- SPSS-Format,
- Tabellenkalkulationsblätter aus Excel oder Lotus,
- Datenbanktabellen,
- Datendateien anderer Statistikprogramme.

Einlesen einer Datei im SPSS-Format

Am einfachsten ist es, wenn die Daten bereits in einer SPSS-Datei vorliegen, wie sie im vorstehenden Abschnitt erzeugt wurde, da diese neben den Daten auch die Definition der Variablen enthält. Die erzeugte Beispieldatei „Testdaten.sav“ lässt sich mit dem Menüpunkt „Datei/ Öffnen/ Daten“ einlesen.[14] In dem Dialogfeld „Daten öffnen“ ist dabei das Verzeichnis, in dem die Datei abgelegt wurde, und der Name der Datei zu wählen. Nach dem Klick auf „Öffnen“ steht sie als Arbeitsdatei in IBM SPSS zur Verfügung. Es ergibt sich die Ansicht wie in Abbildung 13.

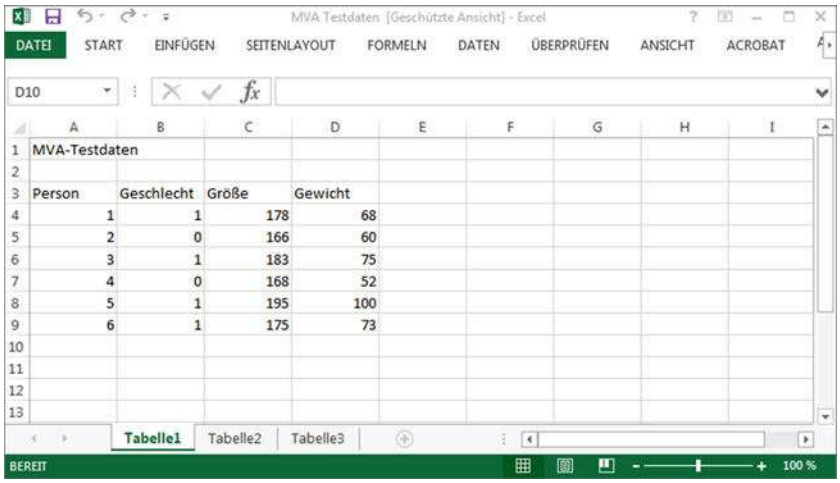
Import von Daten aus Microsoft Excel

Unter den zahlreichen Datenquellen soll hier insbesondere das Einlesen von Daten aus Microsoft Excel beschrieben werden. Abbildung 14 zeigt die Beispieldaten aus Abbildung 5 nach Eingabe in Excel. Die Datenmatrix belegt hier den Bereich A3:D9.

Am einfachsten importiert man numerische Daten aus einem Excel-Tableau nach SPSS mit Copy & Paste (Kopieren & Einfügen). Dazu sind zuvor der Daten-Editor von SPSS und die Excel-Datei zu öffnen. Nach Markieren des Bereichs A4:D9 im Excel-Tableau und Eingabe von „STRG+C“ ist in den Daten-Editor auf die erste Zelle zu springen und dann „STRG+V“ einzugeben. In gleicher Weise lassen sich auch Spalten (Variablen) und Zeilen (Beobachtungen) an eine Datenmatrix im SPSS-Dateneditor anhängen. Und in identischer Form lassen sich auch Daten von SPSS in ein Excel-Tableau übertragen.

Soll eine große Datenmatrix inklusive der Variablennamen aus Excel importiert werden, so empfiehlt sich eine andere Vorgehensweise: Erforderlich ist, dass die Variablennamen in der Zeile direkt über den Daten stehen. Die Datenmatrix umfasst

[14] Dazu ist in der Menüleiste zunächst auf die Option „Datei“ zu klicken, worauf sich ein neues Menü öffnet, in dem sodann auf die Option „Öffnen“ zu klicken ist, usw.



	A	B	C	D
1	MVA-Testdaten			
2				
3	Person	Geschlecht	Größe	Gewicht
4	1	1	178	68
5	2	0	166	60
6	3	1	183	75
7	4	0	168	52
8	5	1	195	100
9	6	1	175	73
10				
11				
12				
13				

Abbildung 14: Testdaten in Excel

damit den Bereich A3:D9. Diesen Bereich sollte man sich merken, bevor man die folgenden Schritte vornimmt:

- Den Menüpunkt „Datei/ Öffnen/ Daten“ wählen
- Im Dialogfeld „Daten öffnen“
  - Unter „Dateien vom Typ:“ den Dateityp „Excel (*.xls, *.xlsx, *.xlsm)“ wählen (voreingestellt ist „SPSS (*.sav)“),
  - Unter „Suchen in:“ das Verzeichnis und den Namen der Excel-Datei („Testdaten.xlsx“) wählen,
  - auf „Öffnen“ klicken.
- Im Dialogfeld „Öffnen einer Excel-Datenquelle“ (Abbildung 15)
  - den Bereich der Datenmatrix im Tabellenblatt angeben: „A3:D9“,
  - die Option „Variablen aus der ersten Datenzeile lesen“ anklicken,
  - auf den Button „OK“ klicken.

Danach stehen die Daten im IBM SPSS-Dateneditor zur Verfügung (vgl. Abbildung 16).

Aus der Excel-Datei können, wie erwähnt, Variablennamen eingelesen werden, wenn diese in der ersten Zeile des angegebenen Bereichs, also direkt über den Daten, stehen. Falls diese Namen nicht den Regeln von IBM SPSS (vgl. 4.1.2.1) entsprechen, werden sie in gültige Variablennamen umgewandelt (z. B. werden Leerzeichen durch Unterstriche ersetzt) und die Spaltenüberschrift von Excel wird als Variablenbeschriftung gespeichert. Falls keine Variablennamen aus Excel eingelesen werden, verwendet IBM SPSS Standardnamen. Leere Zellen in Excel werden bei numerischen Variablen als systemdefinierte fehlende Werte (vgl. 4.1.2.1) behandelt und im Dateneditor durch einen Punkt angezeigt.

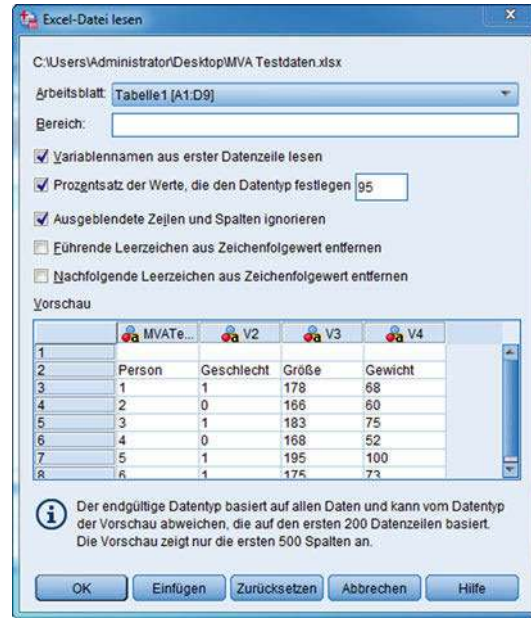


Abbildung 15: Öffnen der Excel-Tabelle mit den Testdaten

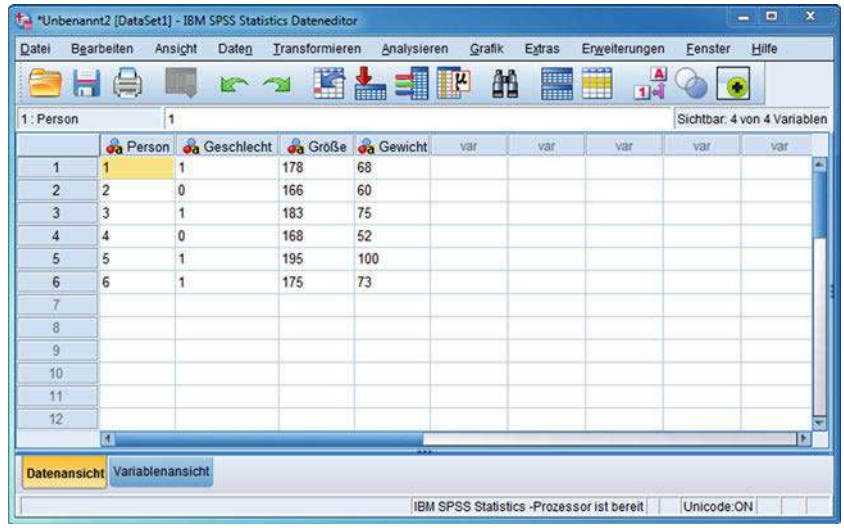


Abbildung 16: Testdaten nach Einlesen aus Excel in den SPSS-Dateneditor

Einlesen von Daten aus anderen Datenquellen

In ähnlicher Weise können Daten aus anderen Datenquellen eingelesen werden.[15] Zum Einlesen von Daten aus Text-Dateien bietet IBM SPSS einen Assistenten für Text-Import an (ähnlich dem Textkonvertierungs-Assistenten von Excel), der unter dem Menüpunkt „Datei/ Textdaten lesen“ nach Aufruf einer Datei gestartet wird.

[15] Vgl. dazu SPSS Inc. (2007): SPSS Base 16.0 - Benutzerhandbuch, S. 18 ff.

4.2 Einfache Statistiken und Grafiken

4.2.1 Einfache Statistiken

Nach der Eingabe oder dem Einlesen von Daten in den Daten-Editor lassen sich diese unter Anwendung der vielfältigen Methoden, die IBM SPSS anbietet, analysieren. Diese Methoden (statistischen Prozeduren) lassen sich über das Menü „Analysieren“ aufrufen (vgl. Abbildung 17).[16]

Statistische Prozeduren

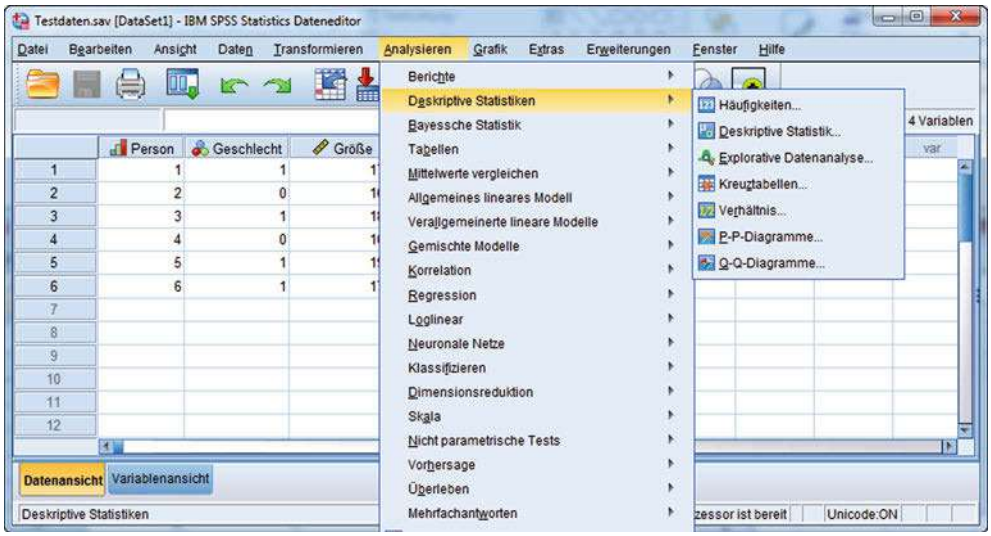


Abbildung 17: Daten-Editor mit Auswahl der Option „Analysieren/Deskriptive Statistiken/Häufigkeiten“

Bevor man mit der Durchführung komplexer Analysen beginnt, ist es in der Regel sinnvoll, zunächst die Daten selbst zu betrachten und sich ein Bild von deren Verlauf oder Verteilung zu machen. Damit werden umfangreiche Datenmengen überschaubar und man erlangt eventuell Hinweise, um Hypothesen über den Zusammenhang zwischen zwei oder mehreren Variablen aufzustellen. Um eine Variable zu beschreiben, lassen sich mit IBM SPSS sehr einfach

Datenbeschreibung

- Häufigkeitsauszählungen vornehmen (für nominale Variable), — Häufigkeiten
- Statistiken wie Lageparameter (Mittelwert, Median, Modalwert), Streuungsparameter (z. B. Standardabweichung, Spannweite) oder Schiefe und Kurtosis (Wölbung) einer Verteilung berechnen, — Statistiken
- grafische Darstellungen der Verteilung in Form eines Balkendiagrammes (für nominale Variablen) oder eines Histogrammes (für metrische Variablen) erstellen und überprüfen. — Diagramme

Derartige Analysen sind auch für die Aufdeckung etwaiger Eingabefehler hilfreich.[17] Weiterhin ist es mittels eines Streudiagrammes möglich, auch zwei oder drei Variablen

Streudiagramm

[16] Zur Beschreibung dieser Methoden siehe auch Norusis/SPSS Inc. (2008). Die den Methoden zugrunde liegenden Algorithmen sind beschrieben in SPSS Inc. (2007b).

[17] Hierfür eignen sich insbesondere auch Box-Plots, die unter dem Menüpunkt „Analysieren/ Deskriptive Statistiken/ Explorative Datenanalyse“ gewählt werden können.

Abbildung 18: Dialogfeld „Häufigkeiten“

gemeinsam in einem Diagramm darzustellen, um so eine erste Vermutung über einen möglichen Zusammenhang zu erhalten. Im Folgenden soll auf einige dieser einfachen Analysen eingegangen werden.

Zur Beschreibung von Variablen wählen wir den Menüpunkt „Analysieren/ Deskriptive Statistiken/ Häufigkeiten“ (vgl. Abbildung 17), mittels dessen sich Häufigkeitsverteilungen tabellarisch und grafisch darstellen und viele weitere beschreibende Statistiken berechnen lassen.[18] Im Dialogfeld „Häufigkeiten“ (vgl. Abbildung 18) sind zunächst eine oder mehrere Variable, die analysiert werden sollen, auszuwählen. Die vorhandenen Variablen sind im linken Fenster (Quellvariablenliste) aufgelistet.

Quellvariablenliste

Beispiel

Als Beispiel wurde hier die Variable „Größe“ ausgewählt. Man wählt eine Variable aus, indem man sie in der Quellvariablenliste anklickt und dann in das rechte Fenster „Variable(n):“ zieht. Alternativ kann man auch nach Markierung der Variablen den Pfeil zwischen den Fenstern (Variablen-Selektionsschalter) anklicken.

Im Dialogfeld „Häufigkeiten“ werden die Optionen „Statistiken“ und „Diagramme“ angeboten. Abbildung 19 zeigt das Dialogfeld für Statistiken. Hier können die gewünschten statistischen Kennzahlen ausgewählt werden. Anschließend ist auf „Weiter“ zu klicken. Unter der Option „Diagramme“ können Balkendiagramme, Kreisdiagramme und Histogramme gewählt werden. Da es sich bei der Variablen „Größe“ um eine metrische Variable handelt, wählen wir „Histogramme“ und zusätzlich die Option „Mit Normalverteilungskurve“. Nach Auswahl der gewünschten Optionen ist wiederum auf „Weiter“ und dann auf „OK“ zu klicken.

Häufigkeitsverteilung

Für eine metrische Variable macht die Auszählung von Häufigkeiten i. d. R. keinen Sinn und kann viel unnützen Output produzieren. Bei größerer Anzahl von Fällen (Personen) wäre es daher zweckmäßig, die Option „Häufigkeitstabellen anzeigen“ (vgl. Abbildung 18 unten links) zu deaktivieren.

SPSS-Viewer

Die Ergebnisse der Analysen von IBM SPSS werden im SPSS-Viewer gezeigt (vgl. Abbildung 20). Der Viewer ist in zwei Fenster unterteilt. Das rechte Fenster enthält den Output (sowie eventuelle Fehlermeldungen), den man bei Bedarf ausdrucken oder als Datei abspeichern kann (Die Ausgabedatei erhält ab SPSS 16.0 die Erweiterung .spv, früher .spo). Das linke Fenster enthält einen Überblick über die Inhalte des Outputs. Außerdem ist im Viewer, gleichermaßen wie im Dateneditor, die Menüleiste verfügbar.

[18] Zur Beschreibung von Variablen lassen sich auch die Menüoptionen „Deskriptive Statistiken“ und „Explorative Datenanalyse“ heranziehen. Der Menüpunkt „Explorative Datenanalyse“ ermöglicht auch die Erstellung von Boxplots und Stengel-Blatt-Diagrammen.

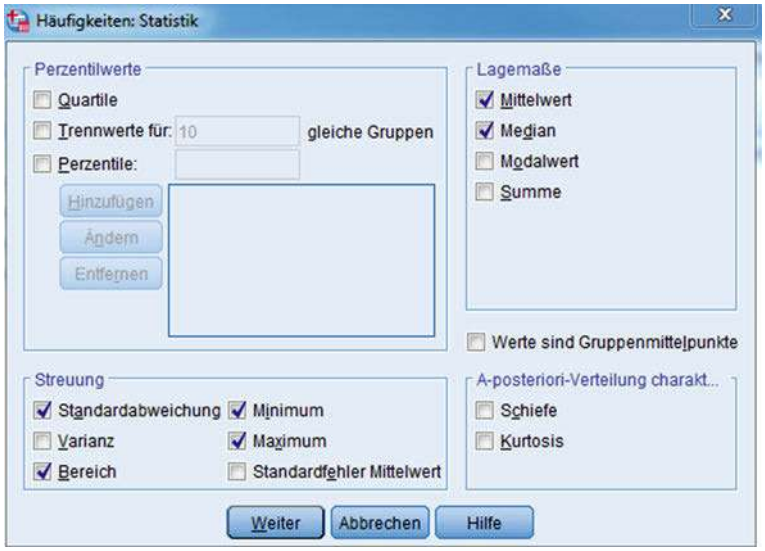


Abbildung 19: Dialogfeld „Häufigkeiten: Statistik“

Abbildung 20 zeigt einen Teil des Outputs, die Werte der statistischen Kennzahlen (Mittelwert, Median, Standardabweichung, Bereich (Spannweite), Minimum, Maximum), die für die Variable „Größe“ optional ausgewählt wurden. Abbildung 21 zeigt das Histogramm für die Variable „Größe“ mit der Normalverteilungskurve. Man ersieht, dass jeweils zwei Personen eine Größe von 160 bis <170 cm und 170 bis <180 cm aufweisen, und dass jeweils eine Person im Bereich 180 bis <190 cm und 190 bis <200 cm liegt. Sämtliche Tabellen und Grafiken lassen sich im Viewer bzw. in der Ausgabedatei auch weiter bearbeiten.

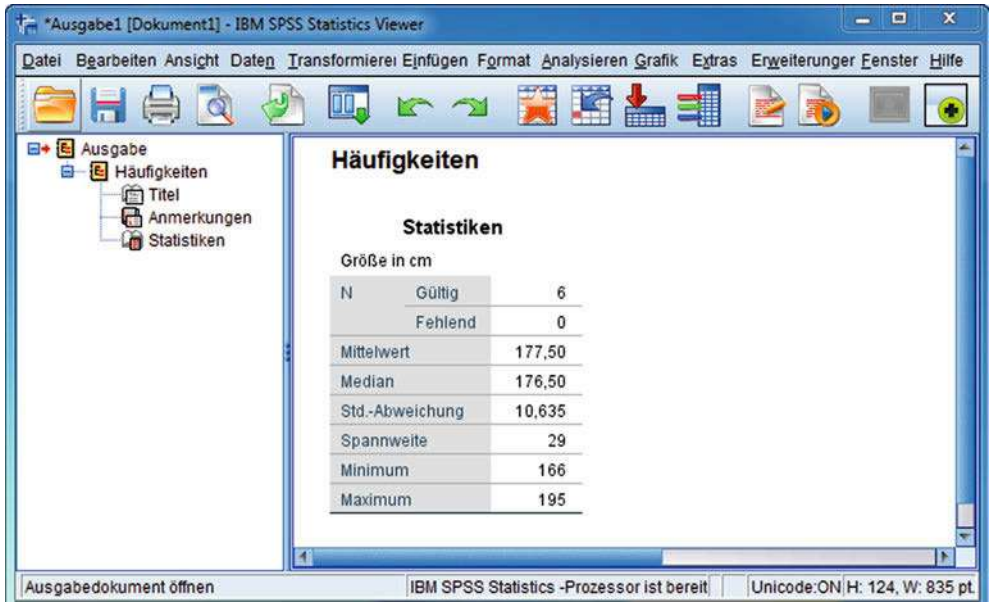


Abbildung 20: SPSS-Viewer mit Output

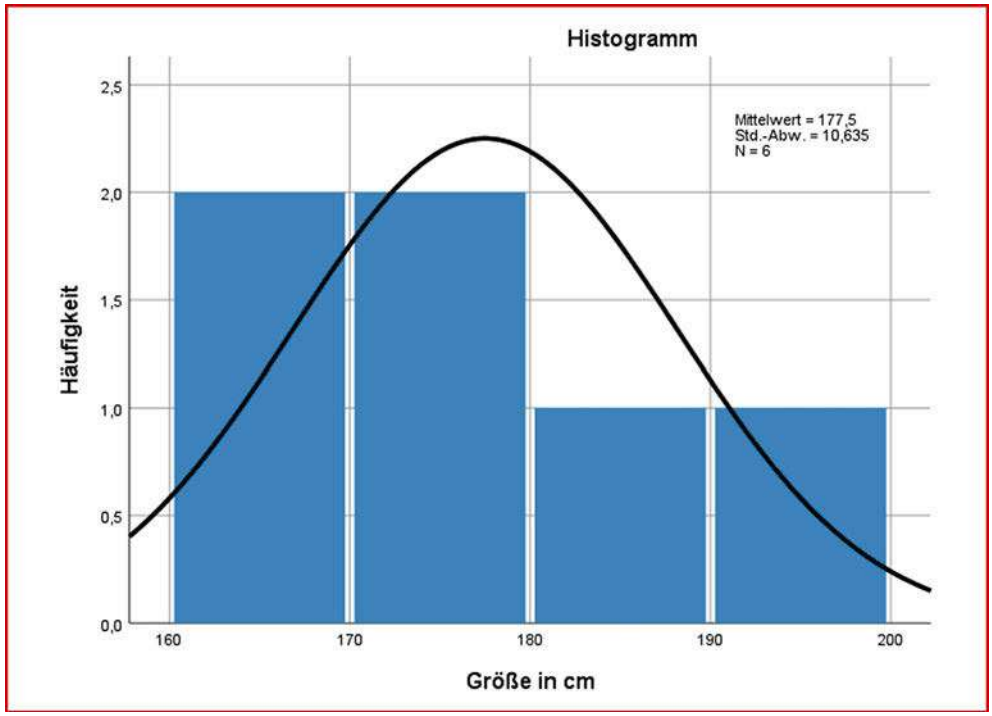


Abbildung 21: Histogramm mit Normalverteilungskurve

4.2.2 Erstellung von Diagrammen

Viele Statistikprozeduren von IBM SPSS enthalten Optionen für die Erstellung von Diagrammen, wie z. B. die vorstehend behandelte Prozedur „Häufigkeiten“. Darüber hinaus aber bietet IBM SPSS unter dem Menüpunkt „Grafik“ spezielle Funktionen zur Erstellung von Grafiken und Diagrammen.

Streudiagramm

Um für zwei (oder drei) metrische Variablen die gemeinsame Verteilung darzustellen und somit auch einen ersten Einblick in deren möglichen Zusammenhang zu erhalten, bietet es sich an, diese Variablen gemeinsam in einem Streudiagramm abzubilden. Hierzu kann die Funktion „Grafik/ Diagrammerstellung“ aufgerufen werden, wodurch das Dialogfeld „Diagrammerstellung“ (vgl. Abbildung 22) geöffnet wird.

Am einfachsten ist die Diagrammerstellung, wenn man im Dialogfeld „Diagrammerstellung“ einen Diagrammtyp wählt. Das geschieht wie folgt:

- Registerkarte „Galerie“ (links in der mittleren Leiste) anklicken,
- eine *Diagrammkategorie* auswählen in der darunter befindlichen Liste „Auswählen aus:“ (jede Kategorie umfasst mehrere Diagrammtypen, die bildlich dargestellt werden),
- einen *Diagrammtyp* auswählen durch Doppelklick auf das Bild oder indem das Bild in die *Zeichenfläche* (die große Fläche oben rechts) gezogen wird.

In Abbildung 22 wurde bereits aus der Kategorie „Streu-/Punktdiagramm“ der erste Typ („einfaches Streudiagramm“) gewählt.

Im nächsten Schritt sind die Achsen zu spezifizieren, was in Abbildung 22 ebenfalls bereits erfolgt ist. Hierzu sind die betreffenden Variablen aus der Liste „Variablen:“ auf die Achsen zu ziehen. Es wurde hier die Variable Größe der x-Achse und die Variable Gewicht der y-Achse zugeordnet. Die in der Zeichenfläche dargestellten Punkte

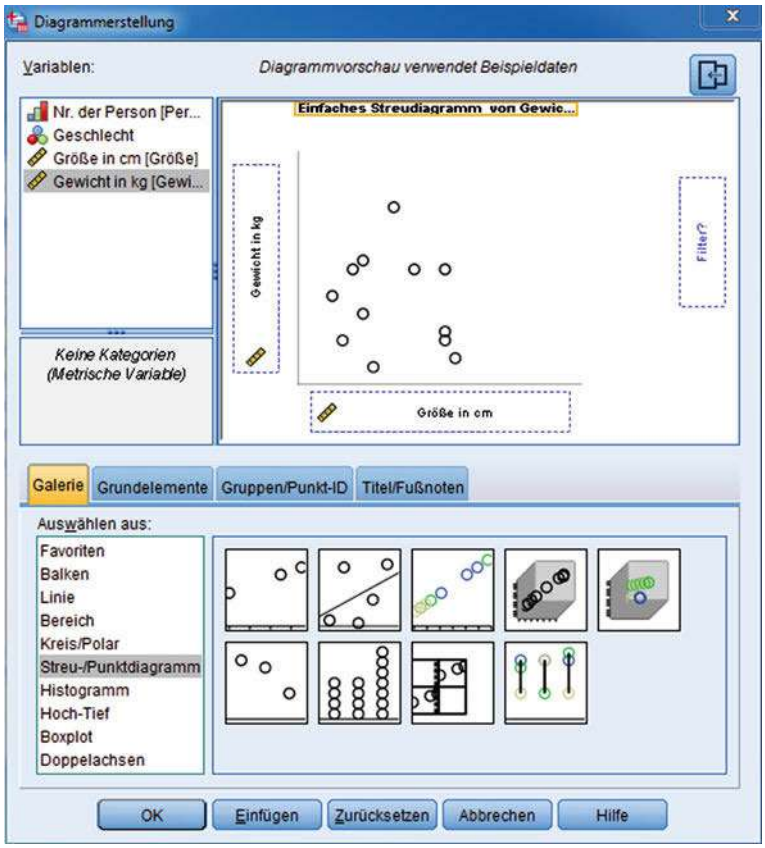


Abbildung 22: Dialogfeld „Diagrammerstellung“

bilden noch nicht die Daten ab, sondern dienen nur der Veranschaulichung des Diagrammtyps. Um das Diagramm zu erstellen, ist auf „OK“ zu klicken.

Das Ergebnis zeigt Abbildung 23. In dem erhaltenen Streudiagramm wird jedes Wertepaar der sechs Personen durch eine Markierung (hier ein roter Punkt) angezeigt.[19] Wie das Streudiagramm verdeutlicht, besteht zwischen den Variablen Gewicht und Größe offenbar ein positiver Zusammenhang. Das heisst, dass mit zunehmender Größe auch das Gewicht zunimmt. Gestützt wird dieser Eindruck auch durch den Korrelationskoeffizienten, der sich durch IBM SPSS ebenfalls leicht berechnen lässt (Menü: „Analysieren/ Korrelation/ Bivariat“). Er beträgt hier 0,975, liegt also nahe bei Eins.

Zusammenhang

Dieser Zusammenhang lässt sich noch deutlicher erkennen, wenn in das Streudiagramm eine Regressionsgerade (siehe zur Regression ausführlich Kapitel 1) eingefügt wird. Dabei ist wie folgt vorzugehen: Durch einen Doppelklick auf das Streudiagramm

[19] Diese Markierungen können nach Form, Größe und Farbe verändert werden. Dazu ist zunächst durch Doppelklick auf das Streudiagramm der Diagramm-Editor zu öffnen (vgl. Abbildung 24) und dann durch Doppelklick auf eine der Markierungen das Dialogfeld „Eigenschaften“. Zusätzlich wäre es möglich, diesen Markierungen zum Beispiel die Variable Geschlecht zuzuordnen. Im Streudiagramm können dann die Markierungen je nach Ausprägung des Geschlechts einer Person unterschiedlich gewählt werden.

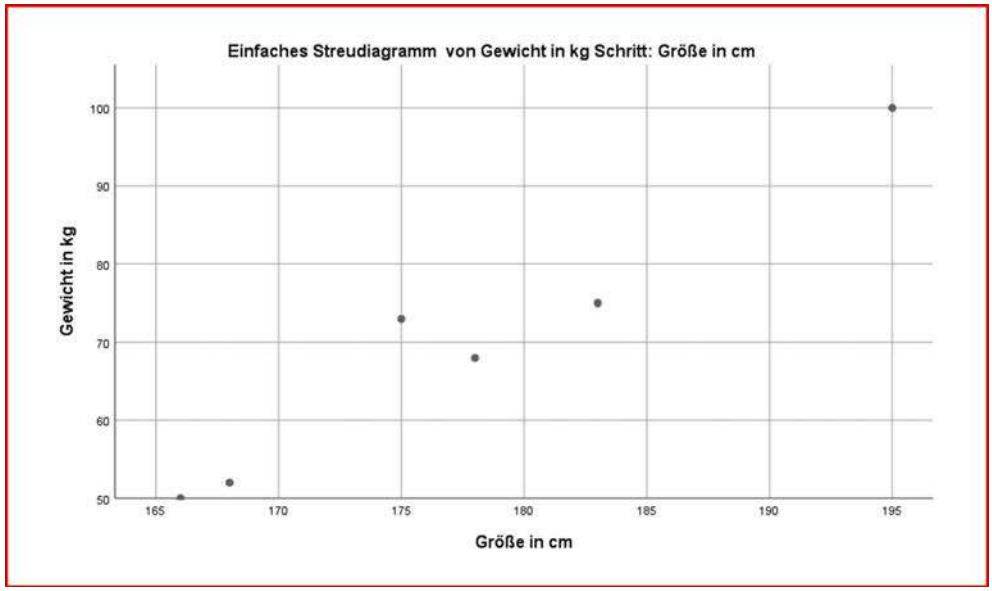


Abbildung 23: Einfaches Streudiagramm für die Variablen Größe und Gewicht

wird ein neues Fenster geöffnet, der Diagramm-Editor (vgl. Abbildung 24). In diesem Editor ist es möglich, das Diagramm weiter zu bearbeiten.

Regressionsgerade

Zum Einfügen der Regressionsgeraden ist im Diagramm-Editor auf das Symbol „Anpassungslinie hinzufügen“ (fünftes Symbol von links direkt über dem Diagrammfeld) zu klicken. Die dadurch in die Grafik eingefügte Regressionsgerade lautet:

$$y = -227 + 1,67x$$

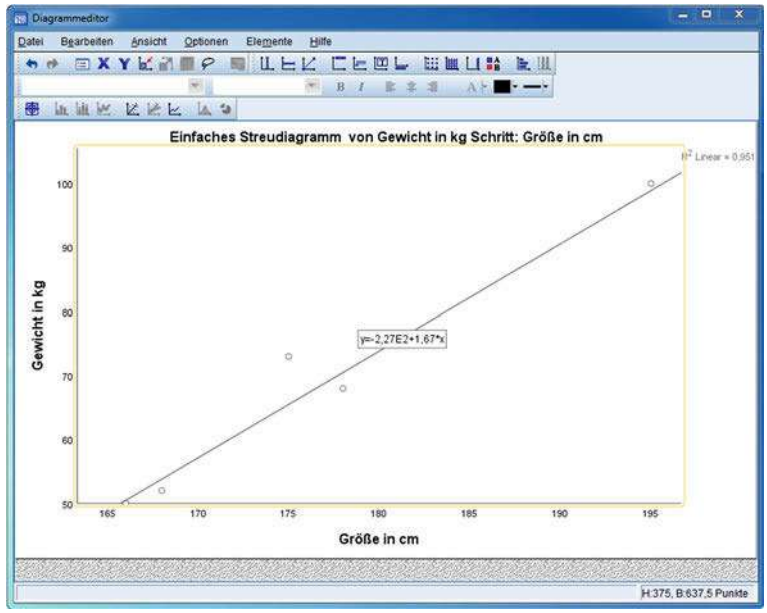


Abbildung 24: Diagramm-Editor

(vgl. Abbildung 24). Sie bestätigt die Vermutung des engen Zusammenhangs zwischen den beiden Variablen denn die einzelnen Wertepaare weisen nur sehr geringe Abweichungen von der Geraden auf. Oben rechts wird das Quadrat des Korrelationskoeffizienten (Bestimmtheitsmaß der Regression) angezeigt, das im Beispiel 0,951 beträgt. Neben einer linearen Anpassungslinie können auch nichtlineare Anpassungslinien (z. B. quadratisch, kubisch) gewählt werden.

4.3 Die Kommandosprache

Das Programmsystem IBM SPSS existiert in unterschiedlichen Versionen für PC und Großrechner. Allen Versionen liegt eine gemeinsame Kommandosprache zugrunde. Auf diese wird auch von der grafischen Benutzeroberfläche von IBM SPSS zugegriffen, d. h. wenn der Benutzer über die Dialogfelder des Programmes Befehle auswählt, werden diese automatisch in die Kommandosprache übersetzt und in eine Syntaxdatei (Kommandodatei) geschrieben. Es handelt sich dabei um eine einfache Textdatei, die gelesen und bearbeitet werden kann. Alternativ kann man aber auch direkt eine Syntaxdatei erstellen und damit den Programmablauf steuern.

Wenngleich sich mit SPSS auch ohne Kenntnis der Kommandosprache arbeiten lässt, so ist es doch vorteilhaft, einige Grundkenntnisse hierüber zu haben. Zum einen sind einige Funktionen von IBM SPSS nur über die Kommandosprache zugänglich und zum anderen ist es bei komplexeren oder häufig wiederkehrenden Analysen von Vorteil, mit Syntaxdateien zu arbeiten. Die Erstellung einer Syntaxdatei wird dem Anwender sehr erleichtert, indem ihm die beim Dialogbetrieb intern erzeugte Kommandosequenz über ein Dialogfenster, dem IBM SPSS Syntax-Editor, zugänglich gemacht wird. Dort kann er sie wie einen Text weiterbearbeiten und sodann erneut starten. Bei Bedarf kann er sie in einer Datei abspeichern, auf die sich später wieder zugreifen lässt. Hierauf wird unten ausführlicher eingegangen.

4.3.1 Aufbau einer Syntaxdatei

Abbildung 25 zeigt ein Beispiel für eine Syntaxdatei. Neben den Syntaxkommandos enthält diese Datei auch den Datensatz aus Abbildung 5.

Die Syntaxdatei gliedert sich in zwei Teile:

- Datendefinition
- Prozeduren (Datenanalyse).

Syntaxdatei

Die Datendefinition beschreibt die Daten und kann auch, wie in der Syntaxdatei in Abbildung 25, die Daten selbst enthalten. Bei größeren Datensätzen wird es dagegen vorteilhaft sein, diese in einer separaten Datei abzulegen. Es ist dann in der Datendefinition der Name der betreffenden Datendatei anzugeben.

Der Prozedurteil instruiert IBM SPSS, welche Analysen mit den Daten vorzunehmen sind. Das Kommando DESCRIPTIVES im Beispiel weist IBM SPSS an, für die drei Variablen Geschlecht, Größe und Gewicht einfache Statistiken wie den arithmetischen Mittelwert und die Standardabweichung zu berechnen. Es können beliebig viele Prozedurkommandos folgen. Mittels FREQUENCIES werden hier die Häufigkeiten der Geschlechter ausgezählt.

4.3.2 Syntax der Kommandos

Die Kommandos entsprechen den Sätzen einer Sprache. Sie sind nach einfachen syntaktischen Regeln aufgebaut.

```
* MVA: Einführung.
* DATENDEFINITION.
DATA LIST FREE / person geschlecht groesse gewicht.
VARIABLE LABELS person "Nr. der Person"
  /geschlecht "Geschlecht"
  /groesse "Groesse in cm"
  /gewicht "Gewicht in kg".
VALUE LABELS
  /geschlecht 0 "weiblich" 1 "männlich".

BEGIN DATA
1 1 178 68
2 0 166 50
3 1 183 75
.....
6 1 175 73
END DATA.

* PROZEDUR.
* Berechnung deskriptiver Statistiken.
DESCRIPTIVES VARIABLES = geschlecht groesse gewicht.

* Berechnung von Häufigkeiten.
FREQUENCIES VARIABLES = geschlecht
  /HISTOGRAM.
```

Abbildung 25: Beispiel einer Syntaxdatei für SPSS

Ein *Kommando* besteht aus einem

- *Schlüsselwort* (keyword), das gleichzeitig auch den Namen des Kommandos bildet (z. B. TITLE, DATA LIST oder DESCRIPTIVES) und
- *Spezifikationen*, die zusätzliche Informationen enthalten.

Spezifikationen

Spezifikationen können folgende Elemente enthalten:

- Schlüsselwörter, z. B. FREE oder VARIABLES,
- Namen, z. B. Person oder Geschlecht,
- Zahlen, z. B. Daten oder Parameter,
- sonstige Zeichenketten (Strings), die durch Hochkommata oder Anführungszeichen eingeschlossen sein müssen, z. B. Titel oder Beschriftungen.

Kommando-Struktur

Beispiel: DATA LIST-Kommando

$$\overbrace{\text{DATA LIST}}^{\text{Kommando}} \qquad \overbrace{\text{FREE / Person Geschlecht Größe Gewicht.}}^{\text{Spezifikation}}$$

Schlüsselwörter sind hier DATA LIST und FREE.

Spezifikationen bilden hier die Formatangabe FREE und die Variablenliste mit den Namen der Variablen. Mehrere Spezifikationen sind durch Schrägstrich (/) zu trennen.

Zur Unterscheidung von Namen und Strings werden hier Schlüsselwörter mit Großbuchstaben geschrieben. IBM SPSS unterscheidet dagegen nicht zwischen Klein- und Großbuchstaben.

Ein Kommando kann auch *Unterkommandos* enthalten, die ebenso aufgebaut sind. Wie alle Kommandos beginnen auch Unterkommandos mit einem Schlüsselwort, das gleichzeitig dessen Namen bildet. Kommandos wie Unterkommandos können Spezifikationen enthalten, müssen es aber nicht. z. B. ist HISTOGRAM ein Unterkommando des Kommandos FREQUENCIES. Es erzeugt eine Darstellung der Häufigkeitsverteilung, die durch FREQUENCIES ermittelt wird. Mehrere Unterkommandos sind

Unterkommando

durch Schrägstrich (/) zu trennen. Falls das Unterkommando Spezifikationen umfasst, so sind diese durch das Gleichheitszeichen (=) vom Kommando-Schlüsselwort zu trennen (z. B. VARIABLES = Geschlecht).

Ein Kommando kann beliebig viele Zeilen umfassen. Es muss aber immer in einer neuen Zeile begonnen und durch einen Punkt (.) abgeschlossen werden. Alternativ kann auch eine Leerzeile angehängt werden. Leerzeichen innerhalb eines Kommandos werden vom Programm überlesen.

Kommentar

Neben den Kommandos kann eine Syntaxdatei auch Kommentarzeilen enthalten, die durch einen Stern (*) einzuleiten sind. Sie dienen der besseren Lesbarkeit der Syntaxdatei. Ein Kommentar kann auch mehrere Zeilen umfassen, wobei Fortsetzungszeilen ebenfalls durch einen Stern einzuleiten oder um wenigstens eine Spalte einzurücken sind.

Die *IBM SPSS-Kommandos* lassen sich grob in drei Gruppen einteilen:

- Kommandos zur Datendefinition (z. B. DATA LIST, VALUE LABELS),
- Prozedurkommandos (z. B. DESCRIPTIVES, REGRESSION),
- Hilfskommandos (z. B. TITLE).

4.3.3 Kommandos zur Datendefinition

Durch das Kommando DATA LIST wird dem IBM SPSS-Programm mitgeteilt, wo die Eingabedaten stehen und wie sie formatiert sind. Falls die Eingabedaten nicht, wie hier im Beispiel, in der Syntaxdatei stehen, könnte hier der Name der Datendatei angegeben werden.

Der Parameter FREE besagt, dass die Eingabedaten formatfrei (freefield) zu lesen sind. Erforderlich ist hierfür, dass die Zahlen durch Leerzeichen (blanks) oder Kommata voneinander getrennt stehen. Wenn den Variablen feste Spalten zugewiesen werden sollen, ist der Parameter FIXED zu verwenden. In diesem Fall ist kein Trennzeichen zwischen den Variablenwerten erforderlich.

Mittels der folgenden Liste von Variablennamen wird angezeigt, wieviele Variablen der Datensatz enthält. Ein Variablenname darf maximal 64 Zeichen umfassen, von denen das erste Zeichen ein Buchstabe sein muss (vgl. hierzu die Ausführungen unter 4.1.2.1). Falls das Datenformat FIXED spezifiziert wurde, muss hinter jedem Namen angegeben werden, welche Spalten die betreffende Variable belegt.

Mit dem Kommando VALUE LABELS können den Werten einer Variablen Beschreibungen zugeordnet werden, um so den Ausdruck besser lesbar zu machen. Die Labels sollten nicht mehr als 120 Zeichen umfassen und müssen durch Hochkommata oder Anführungsstriche eingeschlossen sein.

Ein ähnliches Kommando ist VARIABLE LABELS, mit dem den Variablen bei Bedarf erweiterte Beschriftungen oder Beschreibungen (bis zu 256 Zeichen) zugeordnet werden können.

Die Kommandos BEGIN DATA und END DATA zeigen Beginn und Ende der Daten an. Sie müssen unmittelbar vor der ersten und nach der letzten Datenzeile stehen. Die Daten lassen sich auch als eine Spezifikation von BEGIN DATA auffassen.

Ein Problem, das bei der praktischen Anwendung statistischer Methoden häufig auftaucht, bilden *fehlende Werte*. So kann z.B. der Wert „0“ für das Gewicht einer Person bedeuten, dass der Wert nicht bekannt ist. Um eine Fehlinterpretation zu vermeiden, muss dies dem Programm durch das folgende Kommando angezeigt werden:

```
MISSING VALUE Gewicht (0).
```

Der fehlende Wert, für den hier die „0“ steht, wird dann bei den Durchführungen von Rechenoperationen gesondert behandelt.

Fehlende Werte

Neben derartigen *benutzerdefinierten fehlenden Werten* (user-missing values) setzt IBM SPSS auch automatisch sog. *systemdefinierte fehlende Werte* (system-missing values) ein, wenn im Datensatz anstelle einer Zahl ein Leerfeld oder eine sonstige Zeichenfolge steht. Diese werden bei der Ausgabe durch einen Punkt (.) gekennzeichnet. Generell aber ist es von Vorteil, wenn der Benutzer fehlende Werte durch das MISSING VALUE-Kommando spezifiziert.

4.3.4 Prozedurkommandos

Prozedurkommandos sind im Sprachgebrauch von IBM SPSS alle Kommandos, die „etwas mit den Daten machen“, z. B. sie einlesen, verarbeiten oder ausgeben. Die Kommandos zur Datendefinition (oder auch Transformationen) werden erst dann wirksam, wenn ein Prozedurkommando das Einlesen der Daten auslöst. Der Großteil der Prozedurkommandos betrifft die statistischen Prozeduren von IBM SPSS. Eine Ausnahme ist z. B. das Kommando LIST, mit dem sich die Daten in das Ausgabeprotokoll schreiben lassen.

Durch Prozedurkommandos wird IBM SPSS mitgeteilt, welche statistischen Analysen mit den zuvor definierten Daten durchgeführt werden sollen. So lassen sich z. B. mit dem Kommando DESCRIPTIVES einfache Statistiken wie Mittelwert und Standardabweichung berechnen oder mit dem Kommando REGRESSION eine multiple Regressionsanalyse durchführen. Weitere Kommandos zur Durchführung multivariater Analysen sind z. B. ANOVA, DISCRIMINANT, FACTOR oder CLUSTER. Sie werden im Zusammenhang mit der Darstellung der Verfahren in den jeweiligen Kapiteln dieses Buches erläutert.

Eine Syntaxdatei kann beliebig viele Prozedurkommandos enthalten. Die Prozedurkommandos sind z. T. sehr komplex und können eine große Zahl von Unterkommandos (subcommands) umfassen.

Defaults

Viele Kommandos wie auch Unterkommandos besitzen hinsichtlich ihrer möglichen Spezifikationen *Voreinstellungen (defaults)*, die zur Anwendung kommen, wenn durch den Benutzer keine Spezifikation erfolgt. Die Voreinstellungen von Unterkommandos treten z. T. auch in Kraft, wenn das Unterkommando selbst nicht angegeben wird. So wurde hier bei den Prozeduren DESCRIPTIVES und FREQUENCIES jeweils auf Angabe des Unterkommandos STATISTICS verzichtet, mit Hilfe dessen sich steuern lässt, welche statistischen Maße berechnet und ausgegeben werden sollen.

4.3.5 Hilfskommandos

IBM SPSS kennt eine Vielzahl weiterer Kommandos, die weder die Datendefinition noch die Datenanalyse betreffen und die hier der Einfachheit halber als Hilfskommandos bezeichnet werden. Hierunter fallen z. B. die Kommandos TITLE und SUBTITLE, mit denen sich Seitenüberschriften spezifizieren lassen. Weitere Hilfskommandos, die IBM SPSS anbietet, dienen z. B. zur Steuerung der Ausgabe oder zur Selektion, Gewichtung, Sortierung und Transformation von Daten.

4.3.6 Einlesen einer Syntaxdatei

Eine vorhandene Syntaxdatei kann über den Menüpunkt „Datei/ Öffnen/ Syntax“ eingelesen werden (vgl. Abbildung 26). Standardmäßig haben IBM SPSS-Syntaxdateien

die Extension „.sps“. Es lassen sich aber auch Textdateien mit anderen Extensions einlesen. Nach dem Einlesen der in Abbildung 25 dargestellten Syntax aus einer Textdatei erscheint diese im IBM SPSS Syntax-Editor (vgl. Abbildung 27).

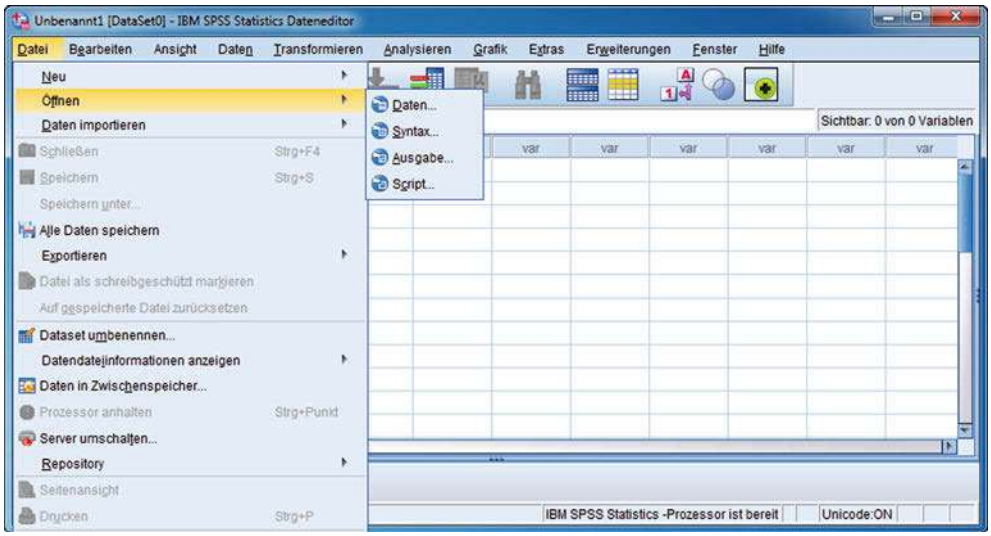


Abbildung 26: Einlesen einer Syntaxdatei

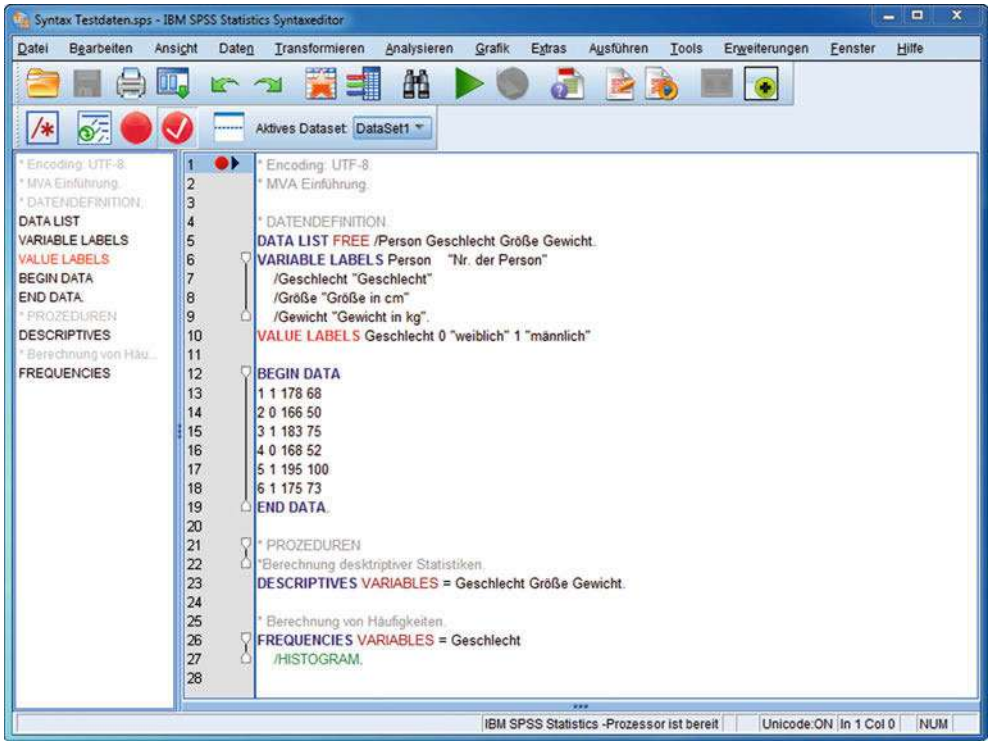


Abbildung 27: SPSS Syntax-Editor

4.3.7 Ausführen der Syntaxdatei

Nachdem eine Syntaxdatei eingelesen wurde, lassen sich sodann entweder sämtliche Befehle der Datei oder einzelne, unmittelbar aufeinander folgende Befehle ausführen. Hierzu ist aus dem Menü des Syntax-Editors der Befehl „Ausführen“ zu wählen, wobei dieser wie folgt spezifiziert werden kann (vgl. Abbildung 28):

- *Alle:* Alle Kommandos der Syntaxdatei werden ausgeführt.
- *Auswahl:* Nur die markierten Kommandos werden ausgeführt.
- *Bis Ende Durchgehen:* Die Kommandos von der aktuellen Cursorposition bis zum Ende der Syntaxdatei werden ausgeführt.

Abbildung 28: Auswahl der Option „Ausführen“ im Syntax-Editor

4.3.8 Erstellen einer Syntaxdatei

Text-Editor

Eine Syntaxdatei lässt sich mit einem beliebigen Text-Editor erstellen. Vorteilhaft ist es jedoch, hierfür den IBM SPSS Syntax-Editor zu verwenden. Ist er nicht bereits geöffnet, so lässt er sich mit dem Menüpunkt „Datei/ Neu/ Syntax“ aufrufen (vgl. Abbildung 26). Es ist aber auch möglich, IBM SPSS so einzustellen, dass automatisch bei jedem Programmstart auch der Syntax-Editor gestartet wird. Hierzu ist

Optionen

der Menüpunkt „Bearbeiten/ Optionen“ aufzurufen. Unter der Registerkarte „Allgemein“ ist sodann die Option „Syntax-Fenster beim Start öffnen“ zu aktivieren (vgl. Abbildung 29).

Die manuelle Erstellung einer Syntaxdatei ist recht mühselig, da zum einen die Kommandosprache beherrscht werden muss[20] und zum anderen leicht Fehler unterlaufen.

Erleichterungen

Erleichtern lässt sich dies, indem man Analysen im Dialog über die Benutzeroberfläche durchführt und die entsprechende Syntax durch IBM SPSS generieren lässt.

[20] Vgl. hierzu SPSS Inc. (2011) oder Sarstedt/Schütz/Raithel (2014).

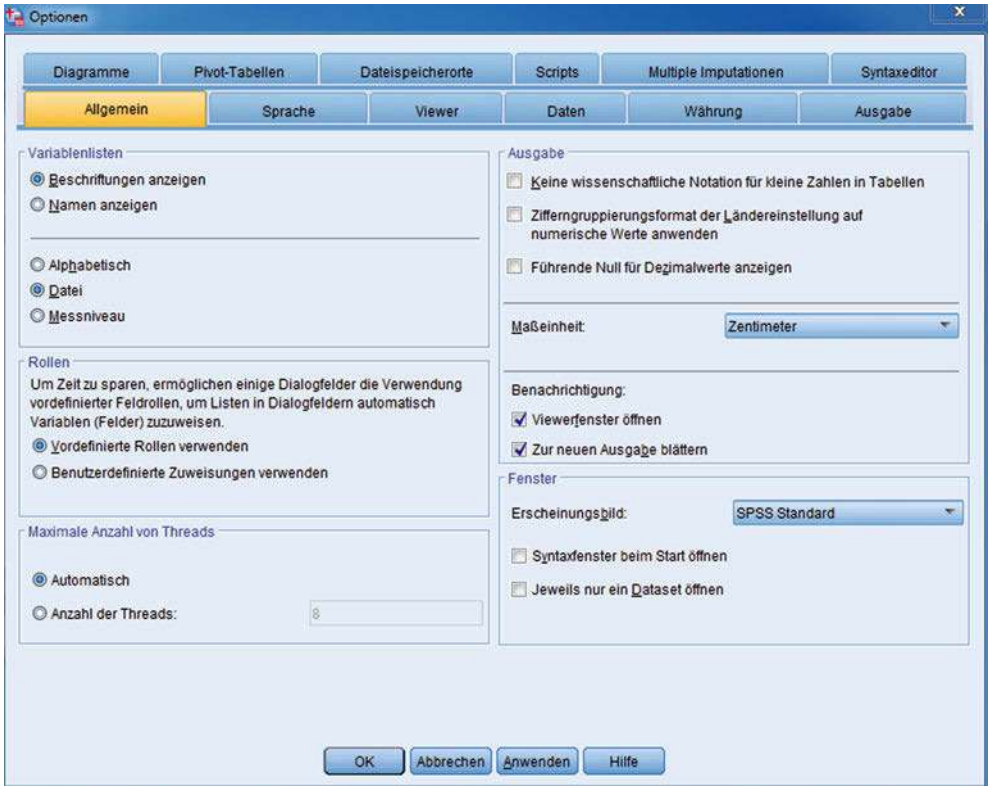


Abbildung 29: Dialogfenster „Bearbeiten/Optionen/Allgemein“

Hierfür stehen zwei Möglichkeiten zur Verfügung:

- Aktivierung der Option „Einfügen“ im Dialogfenster einer Analyse oder Diagrammerstellung (vgl. Abbildung 18 oder Abbildung 22). Die Syntax wird dann automatisch in den Syntax-Editor geschrieben. Syntax-Editor

- Eine zweite Möglichkeit besteht darin, die Syntax im Output-Fenster des Viewers protokollieren zu lassen, von wo sie sich dann manuell in den Syntax-Editor kopieren lässt. Hierzu ist im Menüpunkt „Bearbeiten/ Optionen“ die Registerkarte „Viewer“ zu wählen und dort die Option „Befehle im Log anzeigen“ zu aktivieren (vgl. Abbildung 30 unten links). Abbildung 31 zeigt das Protokoll (Log) des Aufrufs unserer Testdatei und der Analyse der Variablen „Größe“ mittels der Prozedur „Häufigkeiten“ (Frequencies). Viewer

- Eine dritte Möglichkeit, die hier nur erwähnt werden soll, besteht darin, die Syntax in einer Journaldatei zu speichern. Sie hat die Extension „.jnl“, kann aber als Syntaxdatei (.sps) gespeichert werden. Zur Erstellung des Journals ist im Menüpunkt „Bearbeiten/ Optionen“ die Registerkarte „Dateispeicherorte“ zu wählen und dort die Option „Syntax in Journal aufzeichnen“ zu aktivieren. Hier kann auch festgelegt werden, in welchem Verzeichnis das Journal gespeichert werden soll und ob das Journal bei jeder IBM SPSS-Sitzung erweitert („Anhängen“) oder überschrieben („Überschreiben“) werden soll. Journal

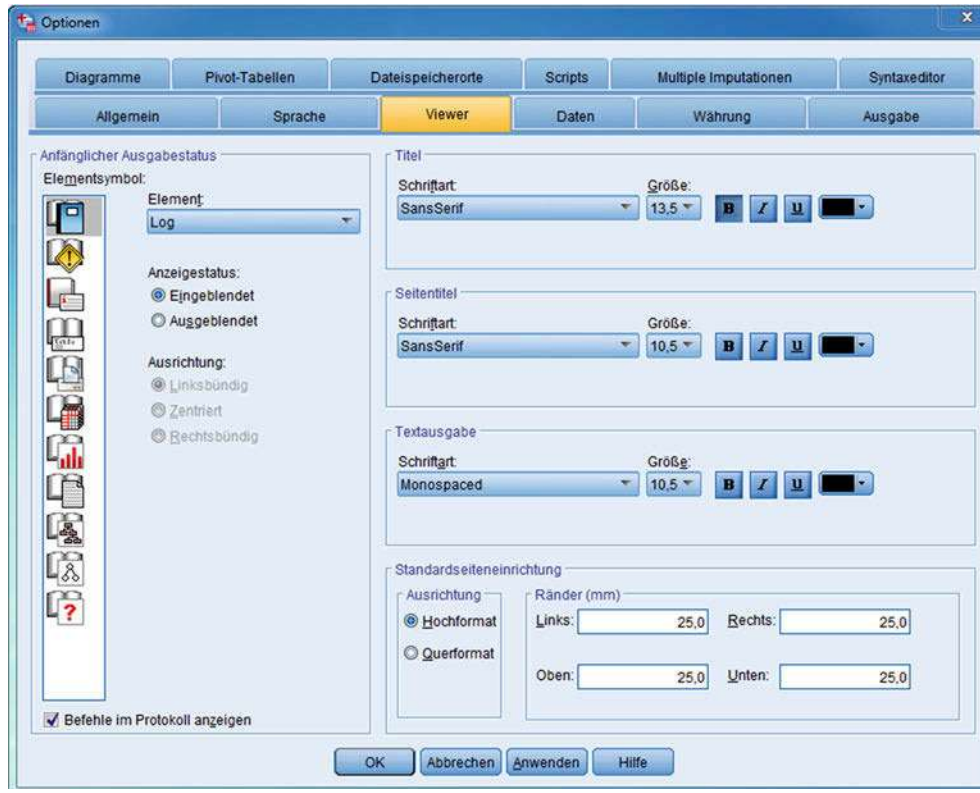


Abbildung 30: Dialogfenster „Bearbeiten/ Optionen/ Viewer“

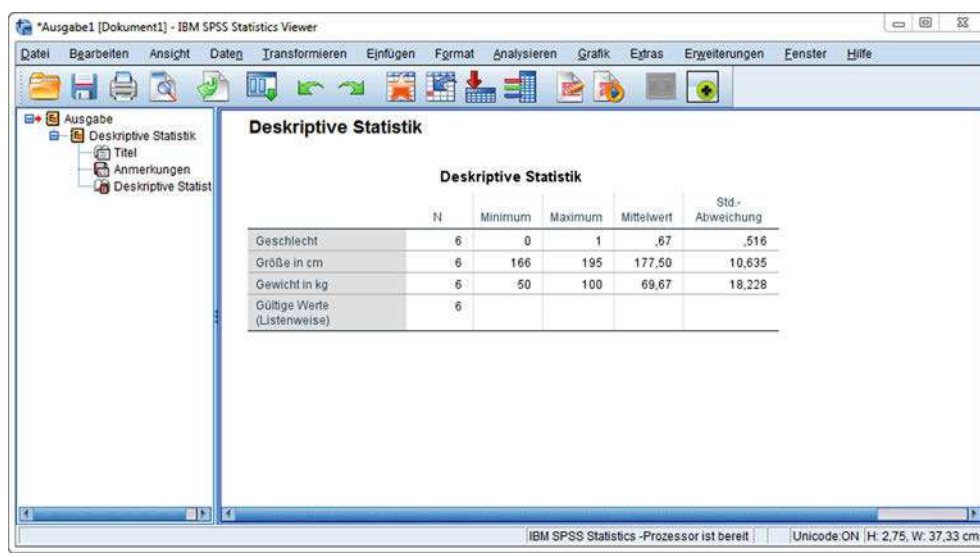


Abbildung 31: Syntaxprotokoll im Viewer

Das Abspeichern einer Syntaxdatei erfolgt über den Menüpunkt „Datei/ Speichern“ oder „Datei/ Speichern unter...“. Sie kann dann zu einem späteren Zeitpunkt wieder aufgerufen werden.

Durch das Erstellen von Syntaxdateien lassen sich komplexe Analysen bequem wiederholen oder auf neue Datensätze anwenden. Bei Bedarf kann die Syntax im IBM SPSS Syntax-Editor bearbeitet und modifiziert werden oder es lässt sich auf Basis einer vorhandenen Syntaxdatei eine neue Syntaxdatei erstellen, was einfacher ist, als bei Null zu beginnen.

4.4 Pakete und Module von IBM SPSS

Um IBM SPSS nutzen zu können, ist auf jeden Fall der Erwerb des Basispaketes „*IBM SPSS Statistics Base*“ erforderlich, das grundlegende statistische Analysen enthält und die Voraussetzung für den Zukauf von weiteren Pakten bzw. Modulen darstellt. Die vielfältigen Erweiterungsmodule, die einzeln oder in Modulpaketen (Bundles) gekauft werden können, haben meist einen Analyseschwerpunkt (z.B. SPSS Regression (Regressionsanalysen); SPSS Conjoint (Conjoint-Analysen); SPSS Neural Networks (Neuronal Netze) und sind an den Belangen der jeweiligen Anwendungsfelder orientiert.[21]

Eine alternative Möglichkeit bietet die Nutzung des Paketes „*IBM SPSS Statistics Premium*“, das alle Verfahren aus dem Basis- und dem Advanced-Paket beinhaltet und an den meisten Universitäten unter der Bezeichnung „*IBM SPSS Statistics*“ verfügbar und damit für Studierende zugänglich ist.

Abbildung 32 gibt eine Übersicht über die in diesem Buch sowie die in unserem Lehrbuch „Fortgeschrittene Multivariate Analysemethoden“ behandelten Analysemethoden und den zugehörigen SPSS-Prozeduren. Mit Ausnahme der Strukturgleichungsanalyse und der konfirmatorischen Faktorenanalyse, die das Spezialpaket *IBM SPSS AMOS* benötigen, sind alle anderen in unseren beiden Lehrbüchern behandelten Verfahren in dem SPSS-Premiumpaket enthalten und laufen unter der gemeinsamen Benutzeroberfläche von „SPSS Statistics“. Für Leser, die nicht das SPSS Premiumpaket nutzen, haben wir in der Spalte „Zusatzmodule“ noch diejenigen Module bzw. Pakete von SPSS aufgeführt, in denen die entsprechenden Verfahren ebenfalls enthalten sind.

4.5 Ergänzende Verwendung von MS Excel

Neben der Verwendung von IBM SPSS zur praktischen Durchführung von multivariaten Analysen bildet die Verwendung des Tabellenkalkulationsprogramms Microsoft Excel eine nützliche Ergänzung. Das Durchrechnen von einfachen Beispielen kann zum besseren Verständnis der Methoden sehr hilfreich sein. Und die Ergebnisse einer Datenanalyse, die mit SPSS erzielt wurden, lassen sich in Excel unmittelbar weiterverarbeiten. Die Analyse kann damit Bestandteil eines übergeordneten Planungssystem werden, das sich mit Excel aufbauen lässt. Für die ergänzende Verwendung von MS Excel sprechen unter anderem folgende Vorteile:

[21] Aktuelle Informationen zu den verschiedenen Modulangeboten finden sich unter: https://www.ibm.com/analytics/de/de/technology/spss/index.html.

Analysemethode	SPSS-Prozeduren	SPSS-Zusatzmodul
Lehrbuch: Multivariate Analysemethoden (15. Auflage)		
1 Regressionsanalyse	REGRESSION	Statistic Base
2 Zeitreihenanalyse	REGRESSION CURVEFIT NLR	Statistics Base Statistics Base SPSS Regression
3 Varianzanalyse	UNIANOVA ONEWAY GLM	Statistics Base
4 Diskriminanzanalyse	DISCRIMINANT	Statistics Base
5 Logistische Regression	LOGISTIC REGRESSION NOMREG	Advanced Statistics oder: SPSS Regression
6 Kreuztabellierung und Kontingenzanalyse	CROSSTABS LOGLINEAR HILOGLINEAR	Statistics Base Advanced Statistics Advanced Statistics
7 Faktorenanalyse	FACTOR	Statistics Base
8 Clusteranalyse	CLUSTER QUICK CLUSTER	Statistics Base Statistics Base
9 Conjoint-Analyse	CONJOINT ORTHOPLAN PLANCARDS	SPSS Conjoint
Lehrbuch: Fortgeschrittene Multivariate Analysemethoden (3. Auflage)		
1 Nichtlineare Regression	NLR	SPSS Regression
2 Strukturgleichungsanalyse		SPSS Amos*
3 Konfirmatorische Faktorenanalyse		SPSS Amos*
4 Auswahlbasierte Conjoint-Analyse	COXREG	Advanced Statistics
5 Neuronale Netze	MLP, RBF	Advanced Statistics bzw. SPSS Neural Networks oder: SPSS Modeler*
6 Multidimensionale Skalierung	ALSCAL PROXSCAL	Statistics Base SPSS Categories
7 Korrespondenzanalyse	CORRESPONDENCE	Statistics Categories

Abbildung 32: Behandelte Analysemethoden und SPSS-Prozeduren
* Eigenständiges Programmpaket, das nicht unter der gemeinsamen Benutzeroberfläche von „SPSS Statistics“ läuft.

- MS Excel ist heute auf nahezu jedem PC verfügbar.[22]
- In MS Excel lassen sich die Daten eines Problems, deren Verarbeitung und die resultierenden Ergebnisse, eventuell ergänzt durch grafische Darstellungen, in einem einzigen Tableau (bzw. Arbeits- oder Tabellenblatt) in übersichtlicher Form zusammenzufassen.
- MS Excel bietet maximale Flexibilität bei Gestaltung von komplexen Berechnungen mittels selbst erstellter Formeln und/oder unter Verwendung der zahlreichen mathematischen und statistischen Funktionen.

[22] Detaillierte Informationen über Excel vermitteln z.B. Arendt-Theilen et al. (2014): Microsoft Excel – Das Handbuch; Jeschke et al. (2013): Microsoft Excel – Formeln & Funktionen.

- Ein Tableau lässt sich dabei so gestalten, dass die Ergebnisse einschließlich der grafischen Darstellungen automatisch an Änderungen der Daten angepasst werden, ohne das erneut eine Funktion aufgerufen werden muss.
- Die Einbettung von Excel in Microsoft Office macht es leicht, die erzielten Ergebnisse in andere Office-Programme zu transferieren, z.B. in Word oder PowerPoint, zwecks Erstellung von Präsentationen oder Dokumentationen.

In Abschnitt 4.1.3 wurde bereits der Import von Daten aus Excel in den SPSS-Daten-editor beschrieben. Umgekehrt lassen sich auch Daten aus dem SPSS-Dateneditor oder Ergebnisse aus dem SPSS-Viewer in eine Excel-Datei über tragen. Am einfachsten erfolgt dies mittels Copy & Paste (Kopieren & Einfügen), womit sich der Inhalt von einzelnen Zellen, aber auch ganzen Spalten, Zeilen oder Matrizen, von SPSS nach Excel oder in umgekehrter Richtung transportieren lassen. Auch der komplette Inhalt des SPSS-Dateneditors, also die Datenmatrix inklusive der Variablennamen, lässt sich in eine Excel-Datei übertragen.

Dazu ist wie folgt vorzugehen:

- Den Menüpunkt „Datei/ Speichern unter“ wählen.
- Im Dialogfeld „Daten speichern als“ unter Punkt „Speichern als Typ:“ den Dateityp wählen, z.B. „Excel 2007 bis 2010 (*.xlsx)“ . Diese Option gilt auch für Excel 2013.
- Bei Bedarf den Dateinamen ändern.
- Bei Bedarf die Option „Variablennamen im Arbeitsblatt speichern“ wählen (default).
- Auf den Button „Speichern“ klicken.

In der so erzeugten Excel-Datei stehen die Variablennamen in der ersten Zeile und die Datenmatrix beginnt in der zweiten Zeile.

Mit MS Excel lassen sich in begrenztem Umfang auch multivariate Analysen durchführen, wenngleich dies weitaus mühseliger ist als mit SPSS. Dabei sind z. B. die in MS Excel enthaltenen Matrixfunktionen sehr nützlich. Durch die Einbeziehung von Makros, deren Erstellung allerding Kenntnisse in der VBA-Programmierung (*Visual Basic for Applications*) erfordert, lassen sich die Möglichkeiten nahezu beliebig erweitern. Zum Leistungsumfang von MS Excel gehört auch das Paket *Analyse-Funktionen*, das vornehmlich zur Durchführung von statistischen Analysen dient. Es wird mit Excel ausgeliefert, muss aber als Add-In installiert werden. In Bezug auf multivariate Analysen ist der Leistungsumfang allerdings bislang recht beschränkt. Enthalten sind Funktionen für Korrelationen, multiple Regression sowie ein- und zweifaktorielle Varianzanalyse. Inzwischen existieren auch leistungsfähige Statistikpakete, die sich als Add-In in Excel integrieren lassen.[23] Da sie auf Excel aufbauen, sind sie recht preisgünstig zu erwerben. MS Excel kann daher nicht nur komplementär, sondern in zunehmendem Maße auch alternativ für multivariate Analysen verwendet werden. Mit allen diesen Möglichkeiten aber lässt sich der Leistungsumfang, den ein so mächtiges Programmsystem wie IBM SPSS bietet, nicht erreichen.

[23] Derartige Statistikpakete sind z. B. XLSTAT und WinSTAT.

Literaturhinweise

A. Basisliteratur zur Verwendung des Buches

Bühl, A. (2014), SPSS 22: Einführung in die moderne Datenanalyse, 14. Auflage, München.

Hair, J./Black, W./Babin, B./Anderson, R. (2010), Multivariate Data Analysis, 7. Auflage, Englewood Cliffs (N.J.).

Herrmann, A./Homburg, C./Klarmann, M. (Hrsg.) (2008), Handbuch Marktforschung, 3. Auflage, Wiesbaden.

Janssen, J./Laatz, W. (2017), Statistische Datenanalyse mit SPSS, 9. Auflage, Berlin u. a.

Norusis, M./SPSS Inc. (2008), SPSS 16.0 Statistical Procedures Companion, Upper Saddle River (N.J.).

IBM Software Group (o.J.), IBM SPSS Statistics Base 22, Chicago.

B. Zitierte Literatur

Agresti, A. (1990), Categorical Data Analysis, John Wiley, New York et al.

Arendt-Theilen, F./Gieringer, D./Hügemann, H./Pfeifer, E./Schiecke, D./Schuster, H. (2014), Microsoft Excel – Das Handbuch, Köln.

Bleymüller, J./Weißbach, R. (2015), Statistik für Wirtschaftswissenschaftler, 17. Auflage, München.

Bühl, A. (2014), SPSS 22: Einführung in die moderne Datenanalyse, 14. Auflage, Pearson, Hallbergmoos.

Fahrmeir, L./Heumann, C./Künstler, R./Pigeot, I./Tutz, G. (2016), Statistik – Der Weg zur Datenanalyse, 8. Auflage, Berlin u. a.

Hair, J./Black, W./Babin, B./Anderson, R. (2010), Multivariate Data Analysis, 7. Auflage, Englewood Cliffs (N.J.).

Härdle, W. K./Simar, L. (2015), Applied Multivariate Statistical Analysis, 4. Auflage, Heidelberg.

Herrmann, A./Homburg, C./Klarmann, M. (2008), Handbuch Marktforschung, 3. Auflage, Wiesbaden.

Janssen, J./Laatz, W. (2017), Statistische Datenanalyse mit SPSS, 9. Auflage, Berlin u. a.

Jeschke, E./Pfeifer, E./Reinke, H./Unverhau, S./Fienitz, B./Bocki, J. (2013), Microsoft Excel – Formeln und Funktionen – Das Maxibuch, 3. Auflage, Unterschleißheim.

Norusis, M./SPSS Inc. (2008), Statistical Procedures Companion, Upper Saddle River (N.J.).

Sarstedt, M./Schütz, T./Raithel, S. (2014), IBM SPSS Syntax: Eine anwendungsorientierte Einführung, 2. Auflage, München.

Schlittgen, R. (2009), Multivariate Statistik, München.

SPSS Inc. (2007a), SPSS Base 16.0 User's Guide, Chicago.

SPSS Inc. (2007b), SPSS 16.0 Command Syntax Reference, Chicago.

SPSS Inc. (2011), IBM SPSS 20 Command Syntax Reference, IBM Corporation.

Tutz, G. (2000), Die Analyse kategorialer Daten, München.

Teil II

Grundlegende Verfahren der multivariaten Analyse

1 Regressionsanalyse

K. Backhaus, B. Erichson, W. Plinke, R. Weiber, *Multivariate Analysemethoden*,
https://doi.org/10.1007/978-3-662-56655-8_2

1.1 Problemstellung

Die Regressionsanalyse ist eines der flexibelsten und am häufigsten eingesetzten statistischen Analyseverfahren. Sie dient der Analyse von Beziehungen zwischen einer abhängigen Variablen und einer oder mehreren unabhängigen Variablen (Abbildung 1.1). Insbesondere wird sie eingesetzt, um

Einsatzbereiche

- Zusammenhänge quantitativ zu beschreiben und sie zu erklären,
- Werte der abhängigen Variablen zu schätzen bzw. zu prognostizieren.

Regressionsanalyse	
eine	eine oder mehrere
ABHÄNGIGE VARIABLE	UNABHÄNGIGE VARIABLE(N)
metrisch	metrisch
Y	$X_1, X_2, \ldots, X_j, \ldots, X_J$

Abbildung 1.1: Die Variablen der Regressionsanalyse

Beispiel

Beispiel: Untersucht wird der Zusammenhang zwischen dem Absatz eines Produktes und seinem Preis sowie anderen den Absatz beeinflussenden Variablen, wie Werbung, Verkaufsförderung etc. Die Regressionsanalyse bietet in einem solchen Fall Hilfe bei z. B. folgenden Fragen: Wie wirkt der Preis auf die Absatzmenge? Welche Absatzmenge ist zu erwarten, wenn der Preis und gleichzeitig auch die Werbeausgaben um bestimmte Größen verändert werden? (Abbildung 1.2)

Regressionsanalyse	
Absatzmenge eines Produktes	Preis, Werbung, Verkaufsförderung, etc.
Y	$X_1, X_2, \ldots, X_j, \ldots, X_J$

Abbildung 1.2: Beispiel zur Regressionsanalyse

Kausalbeziehungen

Der primäre Anwendungsbereich der Regressionsanalyse ist die Untersuchung von *Kausalbeziehungen* (Ursache-Wirkungs-Beziehungen), die wir auch als *Je-Desto-Beziehungen* bezeichnen können. Im einfachsten Fall lässt sich eine solche Beziehung zwischen zwei Variablen, der abhängigen Variablen Y und der unabhängigen Variablen X, wie folgt ausdrücken:

$$Y = f(X) \tag{1.1}$$

Einfach-Regression

Beispiel: Absatzmenge = f(Preis). Je niedriger der Preis, desto größer die abgesetzte Menge. Die Änderungen von Y sind Wirkungen der Änderungen von X (Ursache). Mit Hilfe der Regressionsanalyse lässt sich diese Beziehung quantifizieren und damit angeben, wie groß die Änderung der Absatzmenge bei einer bestimmten Preisänderung ist.

In der Regressionsanalyse wird allerdings kein deterministisches Modell zugrunde gelegt, wie es die funktionale Beziehung (1.1) zum Ausdruck bringt, sondern ein stochastisches Modell der Form

$$Y = f(X) + u$$

wobei u eine Zufallsvariable (Störgröße) ist, in der vielfältige und im einzelnen nicht beobachtbare Einflussgrößen zum Ausdruck kommen. Zunächst aber werden wir uns der Einfachheit halber auf das deterministische Modell beschränken. Das stochastische Modell wird benötigt, um die Ergebnisse der Schätzung des Modells beurteilen zu können, z.B. mittels statistischer Tests.

Mehrfachregression

Bei vielen Problemstellungen liegt keine monokausale Beziehung vor, sondern die zu untersuchende Variable Y wird durch zahlreiche Größen beeinflusst. So wirken neben dem Preis auch andere Maßnahmen wie Werbung, Verkaufsförderung etc. auf die Absatzmenge. Dies lässt sich formal wie folgt ausdrücken:

$$Y = f(X_1, X_2, \ldots, X_j, \ldots, X_J) \tag{1.2}$$

Probleme der Form (1.1) lassen sich mittels *einfacher Regressionsanalyse* behandeln und Probleme der Form (1.2) mittels *multipler Regressionsanalyse*. In jedem Fall muss der Untersucher vor Durchführung einer Regressionsanalyse entscheiden, welches die abhängige und welches die unabhängige(n) Variable(n) ist (sind). Diese Entscheidung liegt oft auf der Hand. So ist sicherlich der Absatz eines Eisverkäufers abhängig vom Wetter und nicht umgekehrt. Manchmal jedoch ist diese Entscheidung schwierig.

Vermutungen

Beispiel: Zu untersuchen sind die Beziehungen zwischen dem Absatz eines Produktes und seinem Bekanntheitsgrad. Welche der beiden Variablen ist die abhängige, welche die unabhängige? Eine Erhöhung des Bekanntheitsgrades eines Produktes bewirkt i. d. R. auch eine Erhöhung der Absatzmenge. Umgekehrt aber wird der Absatz und die damit verbundene Verbreitung des Produktes auch eine Erhöhung des Bekanntheitsgrades bewirken. Ähnlich verhält es sich z. B. im Bereich der Volkswirtschaft zwischen Angebot und Nachfrage.

Interdependenzen

Derartige *interdependente Beziehungen* lassen sich nicht mehr mit einer einzigen Gleichung erfassen. Vielmehr sind hierfür Mehrgleichungsmodelle (simultane Gleichungssysteme) erforderlich, deren Behandlung den hier gegebenen Rahmen allerdings sprengen würde.[1] Wir beschränken uns hier auf Fragestellungen, in denen eine einseitige Wirkungsbeziehung unterstellt werden kann.

Die Bezeichnungen „abhängige“ und „unabhängige“ Variable dürfen nicht darüber hinwegtäuschen, dass es sich bei der in einer Regressionsanalyse unterstellten Kausalbeziehung oft nur um eine Hypothese handelt, d. h. eine Vermutung des Untersuchers. Eine derartige Hypothese muss immer auf ihre Plausibilität geprüft werden, und dazu bedarf es außerstatistischen Wissens, d. h. theoretischer und sachlogischer Überlegungen oder auch der Durchführung von Experimenten.[2]

Plausibilität

[1] Mehrgleichungssysteme behandeln wir im Kapitel Strukturgleichungsanalyse unseres Buches Backhaus/Erichson/Weiber (2015). Siehe dazu auch z. B. Schneeweiß (1990), S. 242 ff.; Kmenta (1997), S. 651 ff.; Greene (2018), S. 326 ff.; Weiber/Mühlhaus (2014), S. 21 ff.

[2] Siehe hierzu z. B. Hammann/Erichson (2000), S. 180 ff.

Kausalität

Es soll hier betont werden, dass sich weder mittels Regressionsanalyse noch sonstiger statistischer Verfahren Kausalitäten zweifelsfrei nachweisen lassen. Vielmehr vermag die Regressionsanalyse nur Korrelationen zwischen Variablen nachzuweisen. Dies ist zwar eine notwendige, aber noch keine hinreichende Bedingung für Kausalität. Im Gegensatz zu einer einfachen Korrelationsanalyse vermag die Regressionsanalyse allerdings sehr viel mehr zu leisten.

Typische Fragestellungen

Typische Fragestellungen, die mit Hilfe der Regressionsanalyse untersucht werden, sowie mögliche Definitionen der jeweils abhängigen und unabhängigen Variablen zeigt Abbildung 1.3. Der Fall Nr. 4 in Abbildung 1.3 stellt einen Spezialfall der Regressionsanalyse dar, die *Zeitreihenanalyse*. Sie untersucht die Abhängigkeit einer Variablen von der Zeit. Formal beinhaltet sie die Schätzung einer Funktion $Y = f(t)$, wobei t einen Zeitindex bezeichnet. Bei Kenntnis dieser Funktion ist es möglich, die Werte der Variablen Y für zukünftige Perioden zu schätzen (prognostizieren). Die Zeitreihenanalyse wird im nachfolgenden Kapitel 2 dieses Buches behandelt. Abbildung 1.4 fasst die in Abbildung 1.3 beispielhaft aufgeführten Fragestellungen zu den drei zentralen Anwendungsbereichen der Regressionsanalyse zusammen.

Fragestellung	Abhängige Variable	Unabhängige Variable
1. Hängt die Höhe des Verkäuferumsatzes von der Zahl der Kundenbesuche ab?	Umsatz pro Verkäufer pro Periode	Zahl der Kundenbesuche pro Verkäufer pro Periode
2. Wie wird sich der Absatz ändern, wenn die Werbung verdoppelt wird?	Absatzmenge pro Periode	Ausgaben für Werbung pro Periode oder Sekunden Werbefunk oder Zahl der Inserate etc.
3. Reicht es aus, die Beziehung zwischen Absatz und Werbung zu untersuchen oder haben auch Preis und Zahl der Vertreterbesuche eine Bedeutung für den Absatz?	Absatzmenge pro Periode	Zahl der Vertreterbesuche, Preis pro Packung, Ausgaben für Werbung pro Periode
4. Wie lässt sich die Entwicklung des Absatzes in den nächsten Monaten schätzen?	Absatzmenge pro Monat t	Menge pro Monat t − k (k = 1, 2, . . . , K)
5. Wie erfasst man die Wirkungsverzögerung der Werbung?	Absatzmenge in Periode t	Werbung in Periode t, Werbung in Periode $t-1$, Werbung in Periode $t-2$ etc.
6. Wie wirkt eine Preiserhöhung von 10 % auf den Absatz, wenn gleichzeitig die Werbeausgaben um 10 % erhöht werden?	Absatzmenge pro Periode	Ausgaben für Werbung, Preis, Einstellung und kognitive Dissonanz
7. Sind das wahrgenommene Risiko, die Einstellung zu einer Marke und die Abneigung gegen kognitive Dissonanzen Faktoren, die die Markentreue von Konsumenten beeinflussen?	Anteile der Wiederholungskäufe einer Marke an allen Käufen eines bestimmten Produktes durch einen Käufer	Rating-Werte für empfundenes Risiko, Einstellung und kognitive Dissonanz

Abbildung 1.3: Typische Fragestellungen der Regressionsanalyse

Ursachenanalyse	Welche Variablen (Faktoren) beeinflussen eine bestimmte interessierende Variable bzw. von welchen Variablen ist sie abhängig?
Wirkungsprognosen	Wie verändert sich die abhängige Variable bei einer Änderung der unabhängigen Variablen?
Zeitreihenanalysen	Wie verändert sich die abhängige Variable im Zeitablauf und somit ceteris paribus auch in der Zukunft?

Abbildung 1.4: Anwendungsbereiche der Regressionsanalyse

Bezeichnungen

Für die Variablen der Regressionsanalyse werden unterschiedliche Bezeichnungen verwendet, was oft verwirrend wirkt. Die Bezeichnungen „abhängige“ und „unabhängige“ Variable sind zwar die gebräuchlichsten, können aber, wie oben dargelegt, Anlass zu Missverständnissen geben. In Abbildung 1.5 finden sich vier weitere Bezeichnungen. Die Benennung der Variablen als Regressanden und Regressoren erscheint am neutralsten und ist somit zur Vermeidung von Missverständnissen besonders geeignet.

Der Begriff der „Regression“ stammt von dem genialen englischen Wissenschaftler Sir Francis Galton (1822–1911), der die Abhängigkeit der Körpergröße von Söhnen in Abhängigkeit von der Körpergröße ihrer Väter untersuchte und dabei die Tendenz einer Rückkehr (regress) zur durchschnittlichen Körpergröße feststellte. D. h. z. B., dass die Söhne von extrem großen Vätern tendenziell weniger groß und die von extrem kleinen Vätern tendenziell weniger klein sind.

Y	$X_1, X_2, \ldots, X_j, \ldots, X_J$
Regressand	Regressoren
abhängige Variable	unabhängige Variable
endogene Variable	exogene Variable
erklärte Variable	erklärende Variable
Prognosevariable	Prädiktorvariable

Abbildung 1.5: Alternative Bezeichnungen der Variablen in der Regressionsanalyse

Dummy-Variablen-Regression

Anwendbarkeit

Die Regressionsanalyse ist immer anwendbar, wenn sowohl die abhängige als auch die unabhängige(n) Variable(n) metrisches Skalenniveau besitzen, es sich also um quantitative Variablen handelt. Dies ist der klassische Fall. Wir hatten aber bereits im einleitenden Kapitel dieses Buches (Abschnitt 3.1) darauf hingewiesen, dass sich durch Anwendung der Dummy-Variablen-Technik qualitative (nominalskalierte) Variablen in binäre Variablen umwandeln lassen, die dann wie metrische Variablen behandelt werden können. Der Anwendungsbereich der Regressionsanalyse lässt sich damit ganz erheblich erweitern.

Allerdings steigt durch Umwandlung die Anzahl der Variablen (eine nominale Variable mit n Ausprägungen ist durch n-1 Dummy-Variablen zu ersetzen). Die Technik kann daher nur für die unabhängigen Variablen, deren Zahl zumindest prinzipiell nicht begrenzt ist, genutzt werden.

Es ist somit grundsätzlich möglich, alle Problemstellungen der Varianzanalyse mit Hilfe der Regressionsanalyse zu behandeln (wenngleich dies nicht immer zweckmäßig ist). Auch eine einzelne binäre Variable kann in der Regressionsanalyse als abhängige

Variable fungieren, und es lassen sich so in beschränktem Umfang auch Probleme der Diskriminanzanalyse (Zwei-Gruppen-Fall) mittels der Regressionsanalyse behandeln. Eine Erweiterung der Regressionsanalyse für nominalskalierte abhängige Variablen ist die Logistische Regression, die im 5. Kapitel dieses Buches behandelt wird.

Anwendungsbeispiel

Beispiel

Wir wollen die Grundgedanken der Regressionsanalyse zunächst an einem kleinen Beispiel demonstrieren. Der Manager eines Margarineherstellers ist mit dem mengenmäßigen Absatz seiner Marke nicht zufrieden. Er stellt zunächst fest, dass der Absatz zwischen seinen Verkaufsgebieten stark differiert. Er möchte wissen, warum die Werte so stark differieren und deshalb prüfen, von welchen Faktoren, die er beeinflussen kann, im wesentlichen die Absatzmenge abhängt. Zu diesem Zweck nimmt er eine Stichprobe von Beobachtungen aus zehn etwa gleich großen Verkaufsgebieten. Insbesondere erhebt er Daten über die abgesetzte Menge, den Preis, die Ausgaben für Verkaufsförderung sowie die Zahl der Vertreterbesuche. Das Ergebnis zeigt die Tabelle in Abbildung 1.6. Die Rohdaten dieses Beispiels enthalten die Werte von vier Variablen, unter denen MENGE als abhängige und PREIS, AUSGABEN (für Verkaufsförderung) sowie (Zahl der Vertreter-) BESUCHE als unabhängige Variablen in Frage kommen. Der Manager hält diese Einflussgrößen für relevant.

Nr.	Menge Kartons pro Periode (MENGE)	Preis pro Karton (PREIS)	Ausgaben für Verkaufsförderung (AUSGABEN)	Zahl der Verterterbesuche (BESUCHE)
1	2.585	12,50	2.000	109
2	1.819	10,00	550	107
3	1.647	9,95	1.000	99
4	1.496	11,50	800	70
5	921	12,00	0	81
6	2.278	10,00	1.500	102
7	1.810	8,00	800	110
8	1.987	9,00	1.200	92
9	1.612	9,50	1.100	87
10	1.913	12,50	1.300	79

Abbildung 1.6: Ausgangsdaten des Rechenbeispiels

Die Untersuchung soll nun Antwort auf die Frage geben, ob und wie die genannten Einflussgrößen sich auf die Absatzmenge auswirken. Wenn z. B. ein ursächlicher Zusammenhang zwischen der Anzahl der Vertreterbesuche und der Absatzmenge besteht, dann müssen Änderungen in der Zahl der Besuche sich auch in Änderungen der Absatzmenge niederschlagen.

1.2 Vorgehensweise

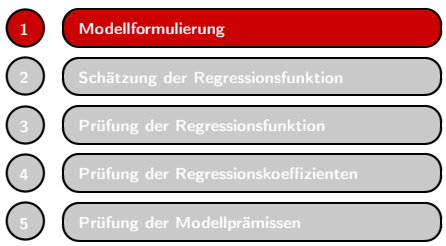
Ursache-Wirkungs-Modell

Bei der Regressionsanalyse geht man regelmäßig in einer bestimmten, der Methode entsprechenden Schrittfolge vor. Zunächst geht es darum, das sachlich zugrunde liegende Ursache-Wirkungs-Problem in Form einer linearen Regressionsfunktion abzubilden. Diese Regressionsfunktion ist sodann auf Basis von Daten empirisch zu

schätzen. In den folgenden Schritten muss die so geschätzte Funktion im Hinblick auf ihre Güte überprüft werden. Den Ablauf zeigt Abbildung 1.7.

Abbildung 1.7: Ablaufschritte der Regressionsanalyse

1.2.1 Modellformulierung



Ein Modell ist eine vereinfachte Abbildung der Realität, oder genauer, eines Ausschnitts oder Aspekts der Realität. Es soll eine strukturelle oder funktionale Ähnlichkeit mit dem Original aufweisen. So bildet z. B. ein Stadtplan das vereinfachte Abbild einer Stadt, das insbesondere den Verlauf der Straßen widerspiegelt. Infolge der Vereinfachung lässt sich ein Stadtplan auf eine handliche Größe reduzieren, sodass man ihn in eine Tasche stecken kann, was mit einer Stadt nicht möglich ist. Modellbildung ist immer eine Gratwanderung zwischen Einfachheit und Komplexität (Vollständigkeit). Ein Modell muss in der Lage sein, einen oder mehrere relevante Aspekte, die den Untersucher interessieren, zu erfassen. Je vollständiger aber ein Modell die Realität abbildet, desto komplexer wird es, und seine Handhabung wird damit zunehmend schwieriger oder gar unmöglich.[3] Der zweckmäßige Detaillierungsgrad hängt vom Verwendungszweck ab, aber auch von der Erfahrung des Verwenders und den verfügbaren Daten. Zweckmäßig ist oft eine evolutorische Vorgehensweise, indem man mit einem einfachen Modell beginnt, welches dann mit zunehmender Erfahrung und Expertise erweitert wird.[4]

Unser Manager beschränkt sich daher im ersten Schritt auf die Untersuchung des folgenden Zusammenhangs:

$$\text{Absatzmenge} = f(\text{Zahl der Vertreterbesuche})$$

Diese Modellformulierung impliziert die Hypothese, dass zwischen Absatzmenge und Zahl der Vertreterbesuche ein kausaler Zusammenhang besteht, wobei die Absatzmenge die abhängige Variable ist und die Zahl der Vertreterbesuche die unabhängige Variable. Der Manager vermutet weitergehend, dass die Wirkung der Vertreterbesuche positiv ist, also dass mit steigender Zahl der Vertreterbesuche auch die Absatzmenge

[3] Dies tangiert in der Regressionsanalyse das Problem der Multikollinearität, das in Abschnitt 1.2.5.6 behandelt wird.
[4] Siehe dazu Little (1970).

steigt. Dieser bislang hypothetisch formulierte Zusammenhang ist im Folgenden zu überprüfen.

Eine wichtige Klasse von Modellen sind mathematische Modelle, da für ihre Handhabung die Mathematik verwendet werden kann. Ein sehr einfaches, aber vielseitig verwendbares mathematisches Modell bildet die folgende lineare Regressionsfunktion:[5]

Regressionsfunktion

$$\hat{Y} = b_0 + b_1 X \tag{1.3}$$

mit

$\hat{Y}$ = Schätzung der abhängigen Variablen Y
b_0 = konstantes Glied
b_1 = Regressionskoeffizient
X = unabhängige Variable

Die Funktion (1.3) ist in Abbildung 1.8 dargestellt. Sie bildet eine Gerade und wird daher auch als Regressionsgerade bezeichnet.

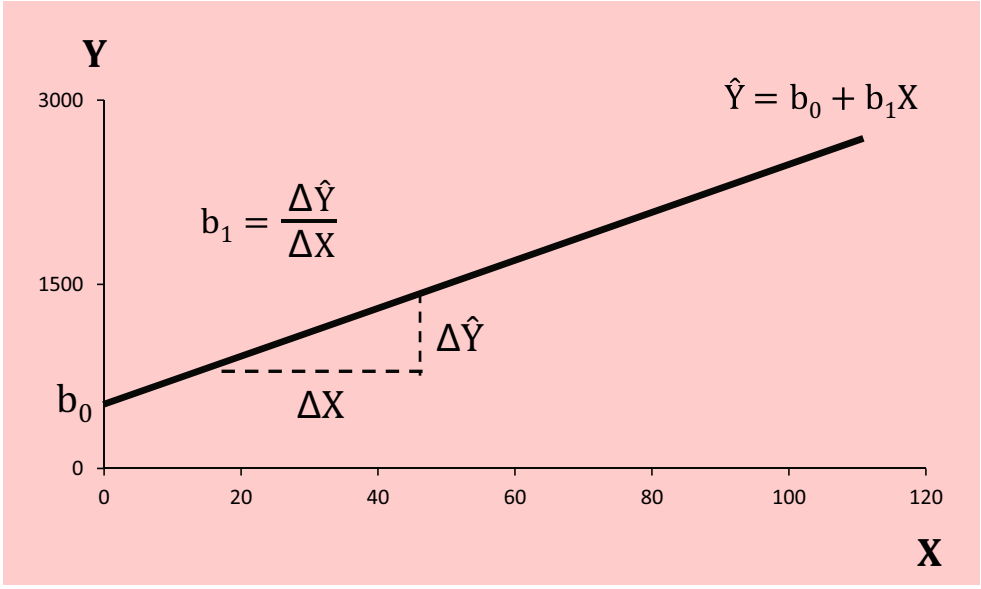


Abbildung 1.8: Lineare Regressionsfunktion

Bezogen auf unser Beispiel lässt sich die Regressionsfunktion (1.3) wie folgt schreiben:

$$\text{Geschätzte Absatzmenge} = b_0 + b_1 \cdot \text{Zahl der Vertreterbesuche}$$

[5] Zu unterscheiden ist zwischen einer linearen Regressionsfunktion, deren Schätzung das Ziel einer Regressionsanalyse bildet, und dem linearen Regressionsmodell, welches den theoretischen Unterbau der Regressionsanalyse liefert. Es handelt sich dabei um ein stochastisches Modell, auf das in Abschnitt 1.2.3.2 eingegangen wird.

Die Parameter der Funktion haben folgende Bedeutung:

- b_0 gibt den Schnittpunkt der Regressionsgeraden mit der Y-Achse (vertikale Achse bzw. Ordinate) des Koordinatensystems an (vgl. Abbildung 1.8). Inhaltlich bedeutet b_0 diejenige Absatzmenge, die zu erwarten ist, wenn die Zahl der Vertreterbesuche gleich Null ist. b_0 wird als *konstantes Glied* der Regressionsfunktion bezeichnet.

- b_1 ist der Koeffizient der unabhängigen Variablen X, der Vertreterbesuche. Er wird als *Regressionskoeffizient* bezeichnet. Geometrisch gesehen gibt er die Steigung oder Neigung der Regressionsgeraden an. Es gilt:

$$b_1 = \frac{\Delta \hat{Y}}{\Delta X} \tag{1.4}$$

Inhaltlich bildet der Regressionskoeffizient ein Maß für die Stärke der Wirkung von X auf Y. Er gibt an, um wieviel Einheiten sich Y vermutlich ändern wird, wenn sich X um eine Einheit ändert. Gilt z. B. $b_1 = 20$, so heißt das, dass jeder Vertreterbesuch im Durchschnitt einen Mehrabsatz in Höhe von 20 Kartons erbringt. Dies ist eine für Marketingentscheidungen wichtige Information, an der unser Manager ganz besonders interessiert ist.

Nehmen wir weiterhin an, dass $b_0 = 500$ gilt, dann erhalten wir anstelle von (1.3) die folgende Funktion:

$$\hat{Y} = 500 + 20X$$

Das ist genau die in Abbildung 1.8 dargestellte Regressionsfunktion. Damit könnte der Manager für jede mögliche Anzahl von Vertreterbesuchen die resultierende Absatzmenge berechnen. Für einzelne Werte von $\hat{Y}$ und X verwenden wir kleine Buchstaben. Für x = 100 Besuche erhält man z. B. für die Absatzmenge den folgenden Schätzwert:

$$\hat{y} = 500 + 20 \cdot 100 = 2500$$

Aber noch sind uns die Werte der Parameter b_0 und b_1 nicht bekannt, sondern sie müssen erst mittels Regressionsanalyse auf Basis der beobachteten Daten geschätzt werden. Dazu ist es zweckmäßig, sich zunächst ein Bild von den vorhandenen Daten zu verschaffen. Hierfür eignet sich ein *Streudiagramm* (auch Punktediagramm, X,Y-Diagramm), wie es Abbildung 1.9 zeigt.

Jede Beobachtung von Absatzmenge und Zahl der Vertreterbesuche für ein Verkaufsgebiet ist als Punkt (x_k, y_k) dargestellt $(k = 1, 2, ..., 10)$. Der Punkt rechts oben ist der Punkt (x_1, y_1), also die erste Beobachtung mit den Werten (109, 2.585) aus Abbildung 1.6. Mittels Excel oder SPSS lassen sich derartige Streudiagramme auch für große Datenmengen leicht erstellen.

Das Streudiagramm lässt erkennen, dass die Absatzmenge tendenziell mit der Zahl der Vertreterbesuche ansteigt. Dies bestätigt die Hypothese des Verkaufsleiters, dass eine Wirkungsbeziehung zwischen Vertreterbesuchen und Absatzmenge besteht und das diese Beziehung annähernd linear verläuft. Und weiterhin bestätigt es die Vermutung des Managers, dass die Wirkung der Vertreterbesuche positiv ist und damit die Beziehung einen ansteigenden Verlauf aufweist. Das hypothetisch formulierte lineare Regressionsmodell erscheint daher zur Abbildung dieser Beziehung geeignet zu sein.

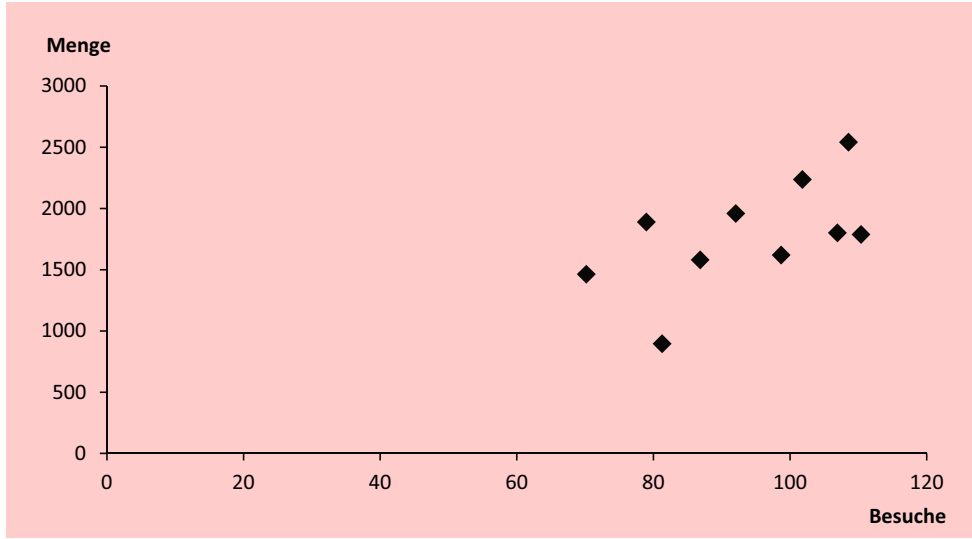


Abbildung 1.9: Streudiagramm der Beobachtungswerte

Anders sähe es aus, wenn das in Abbildung 1.10 dargestellte Streudiagramm vorliegen würde. Dieses lässt einen nichtlinearen Verlauf erkennen, der eventuell dadurch entstanden sein könnte, dass zu viele Vertreterbesuche den Käufern lästig werden und sie daher aus Verärgerung weniger kaufen. Ein derartiger nichtlinearer Zusammenhang lässt sich zwar ebenfalls mittels linearer Regressionsanalyse behandeln, macht aber eine andere Modellspezifikation erforderlich.[6]

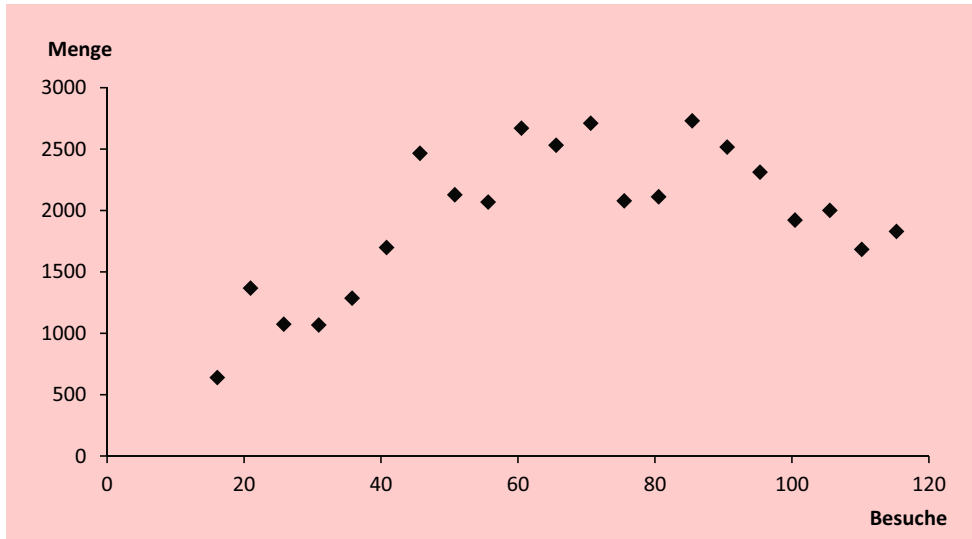


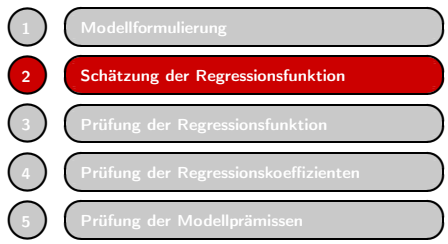
Abbildung 1.10: Streudiagramm mit nichtlinearem Verlauf

[6] In Abschnitt 1.2.5.1 wird darauf eingegangen, wie sich auch nichtlineare Zusammenhänge mittels linearer Regressionsanalyse behandeln lassen. Zur Nichtlinearen Regressionsanalyse siehe Backhaus/ Erichson/Weiber (2015).

Die Modellbildung kann damit abgeschlossen werden und es müssen im nächsten Schritt die Parameter des Modells (1.3), die Werte von b_0 und b_1, auf Basis der vorhandenen Daten geschätzt werden.

1.2.2 Die Schätzung der Regressionsfunktion

1.2.2.1 Einfache Regression



Bei Betrachtung der vorliegenden Daten im Streudiagramm (Abbildung 1.9) wird deutlich, dass keine Regressionsgerade existiert, die durch alle Punkte des Streudiagramms verläuft. Vielmehr muss es bei der Regressionsanalyse darum gehen, eine Regressionsgerade zu finden, die sich der empirischen Punkteverteilung möglichst gut anpasst, oder anders ausgedrückt, die die Abweichungen minimiert.

Wir nehmen hier das Ergebnis der Regressionsanalyse vorweg. Die optimale Regressionsgerade lautet:

$$\hat{Y} = 39,5 + 18,881X$$

und sie ist in Abbildung 1.11 dargestellt.

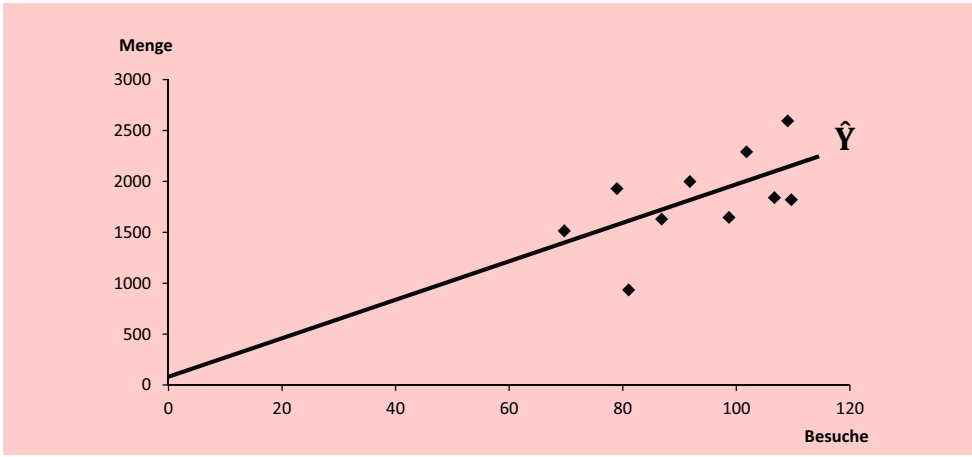


Abbildung 1.11: Streudiagramm mit Regressionsgerade

Abbildung 1.12 zeigt einen vergrößerten Ausschnitt aus dem Streudiagramm in Abbildung 1.11. Durch die Verschiebung der Y-Achse nach rechts verschiebt sich auch der Schnittpunkt der Regressionsgeraden mit der Y-Achse nach oben und liegt jetzt nicht mehr bei b_0. Zusätzlich ist in Abbildung 1.12 der Mittelpunkt der Punktewolke $(\overline{x}, \overline{y}) = (93,6, 1.806,8)$ markiert, durch den die Regressionsgerade verläuft.

Ein Grund dafür, dass die Punkte für die beobachteten Absatzmengen nicht auf einer Geraden liegen, sondern um diese streuen, liegt darin, dass neben den Vertreterbesuchen natürlich noch andere Einflussgrößen auf die Absatzmenge einwirken. Dabei

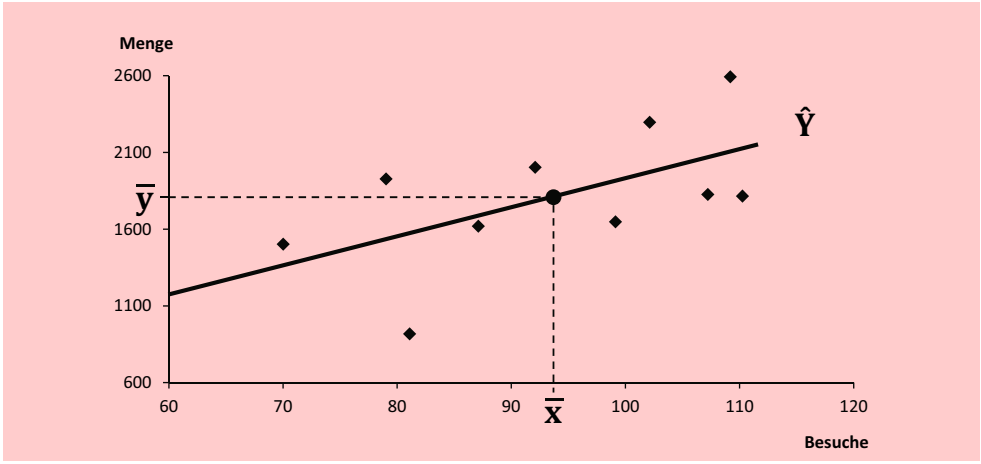


Abbildung 1.12: Ausschnitt aus dem Streudiagramm mit Regressionsgerade

sind zwei Arten von Einflussgrößen zu unterscheiden:

- *Systematische Einflussgrößen*, wie z. B. der Preis, die Ausgaben für Verkaufsförderungen oder Maßnahmen der Konkurrenz, die in der Regressionsgleichung bislang nicht berücksichtigt wurden.
- *Zufällige Einflussgrößen*, die sich nicht erfassen lassen.

Weitere Gründe für die Streuung der empirischen Werte können z. B. Beobachtungsfehler bzw. Messfehler sein.

Die Differenzen zwischen den beobachteten und den geschätzten Y-Werten werden als *Residuen* bzw. Residualgrößen bezeichnet und gewöhnlich durch e (wie „error“) symbolisiert.

Residualgröße

$$e_k = y_k - \hat{y}_k \qquad (k = 1, 2, \ldots, K) \tag{1.5}$$

mit

$$\begin{array}{lcl} y_k & = & \text{Beobachtungswert der abhängigen Variablen } Y \text{ für } x_k \\ \hat{y}_k & = & \text{ermittelter Schätzwert von } Y \text{ für } x_k \\ K & = & \text{Zahl der Beobachtungen} \end{array}$$

Geometrisch handelt es sich bei der Residualgröße e_k um den senkrechten Abstand eines Beobachtungspunktes k von der Regressionsgeraden (siehe Abbildung 1.13). Liegt ein Beobachtungspunkt unterhalb der Regressionsgeraden, so nimmt die Residualgröße einen negativen Wert an. Durch Umformung von (1.5) und unter Einbeziehung von (1.3) lässt sich folgende Regressionsfunktion bilden:

$$\begin{aligned} Y &= \hat{Y} + e \\ &= b_0 + b_1 X + e \end{aligned} \tag{1.6}$$

Die Regressionsfunktion Y setzt sich damit additiv zusammen aus einer systematischen Komponente $\hat{Y}$ und der Residualgröße e. Bezogen auf unser Beispiel heißt das: Die beobachteten Absatzmengen setzen sich zusammen aus den (durch die Regressionsgerade) geschätzten Absatzmengen und einer Restgröße, deren Variationen sich durch die Zahl der Vertreterbesuche nicht erklären lassen.

Für eine einzelne Beobachtung gilt:

$$y_k = b_0 + b_1 x_k + e_k \qquad (k = 1, 2, \ldots, K)$$

Abbildung 1.13 veranschaulicht dieses grafisch.

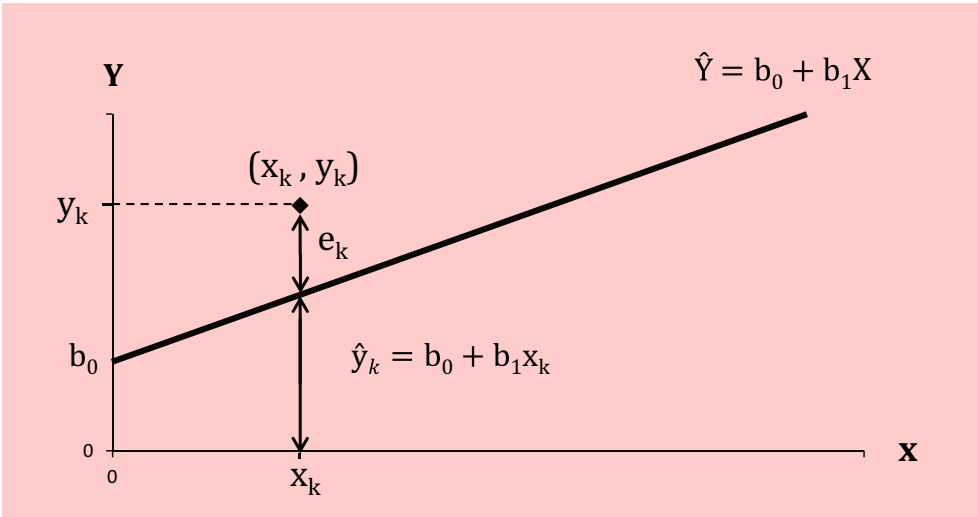


Abbildung 1.13: Systematische Komponente und Residualgröße

Die Residualgrößen bilden den Schlüssel zur Schätzung der Regressionsgeraden, also zur Bestimmung der Parameter b_0 und b_1. Die Regressionsgerade erlangt dann eine gute Anpassung an die empirische Punkteverteilung, wenn die Residualgrößen möglichst klein werden. Die Schätzung der Regressionsfunktion bildet damit mathematisch gesehen die Lösung eines Optimierungsproblems. Benötigt wird dazu ein geeignetes Optimierungskriterium.

Intuitiv erscheint es plausibel, dass die optimale Gerade durch den Mittelpunkt der Punktwolke verlaufen muss. Man könnte daher versucht sein, das Problem zu lösen, indem man eine Gerade durch den Mittelpunkt legt und diese dann dreht, bis die Summe der Residuen möglichst klein wird (vgl. Abbildung 1.14). Dabei aber wird man feststellen, dass für jede Gerade durch den Mittelpunkt gilt:

$$\sum_{k=1}^{K} e_k = 0$$

Das liegt daran, dass sich die positiven und negativen Residuen jeweils kompensieren. Die Summe der Residuen bildet daher kein geeignetes Optimierungskriterium.

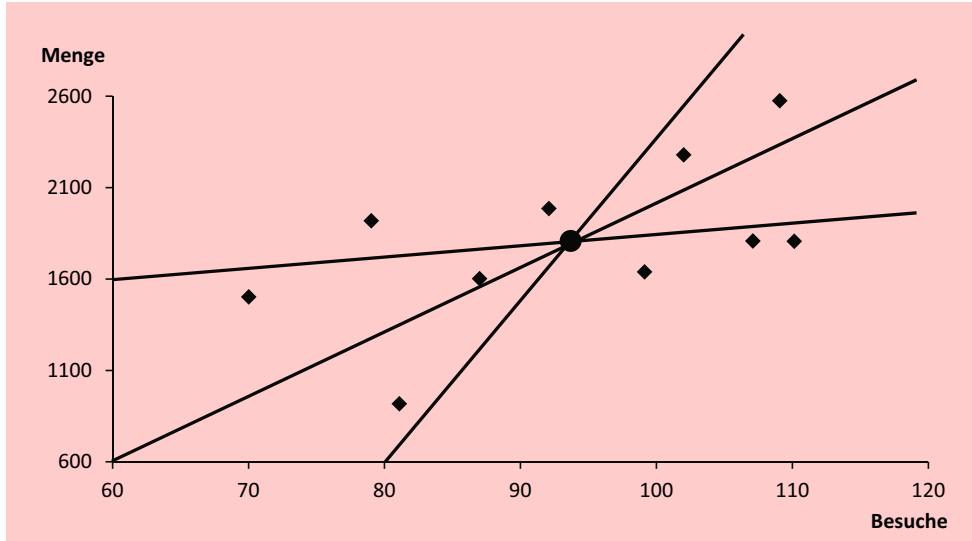


Abbildung 1.14: Alternative Geraden durch den Mittelpunkt der Punktwolke

Um geeignete Optimierungskriterien zu erhalten, muss man die negativen Vorzeichen ausschalten. Das kann geschehen, indem man entweder die Absolutwerte der Residuen verwendet oder indem man die Residuen quadriert. Damit ergeben sich die folgenden zwei alternativen Optimierungskriterien:

$$\text{a)} \qquad \sum_{k=1}^{K} |e_k| \rightarrow min! \qquad\qquad \text{b)} \qquad \sum_{k=1}^{K} e_k^2 \rightarrow min!$$

Mathematisch ist das zweite Kriterium leichter zu handhaben. Es wird daher vorherrschend in der Regressionsanalyse verwendet. Man bezeichnet es als Kleinst-Quadrate-Kriterium oder kurz KQ-Kriterium. Unter Verwendung von (1.6) lässt sich damit die Zielfunktion für das zu lösende Optimierungsproblem wie folgt schreiben:

Kleinst-Quadrate-Kriterium der Regressionsanalyse

$$\sum_{k=1}^{K} e_k^2 = \sum_{k=1}^{K} [y_k - (b_0 + b_1 x_k)]^2 \rightarrow min! \tag{1.7}$$

Die obige Summe ist eine Funktion der unbekannten Regressionsparameter b_0 und b_1. Gesucht werden also diejenigen Werte der Parameter, für die die Summe der quadrierten Residuen minimal wird und damit die Regressionsgerade eine optimale Anpassung an die Beobachtungswerte erhält. Automatisch erhält man damit eine Regressionsgerade, die durch den Nullpunkt der Punkteverteilung verläuft und für die die Summe der Residuen somit gleich Null ist. Abbildung 1.11 und Abbildung 1.12 zeigen die optimale Regressionsgerade im Streudiagramm. Diese Art der Schätzung wird auch als „*Methode der kleinsten Quadrate*“ oder kurz KQ-Methode bezeichnet.

Die KQ-Methode gehört zu den wichtigsten statistischen Schätzverfahren. Durch die Quadrierung der Residuen wird vermieden, dass sich positive und negative Abweichungen kompensieren, und es werden größere Abweichungen stärker gewichtet.[7]

Analytisch erhält man die gesuchten Schätzwerte für b_0 und b_1 durch partielle Differentiation von (1.7) nach b_0 und b_1. Dadurch ergeben sich folgende Formeln:

$$b_1 = \frac{K(\sum x_k y_k) - (\sum x_k)(\sum y_k)}{K(\sum x_k^2) - (\sum x_k)^2} \qquad \text{Regressionskoeffizient} \tag{1.8}$$

$$b_0 = \bar{y} - b_1 \bar{x} \qquad \text{Konstantes Glied} \tag{1.9}$$

Die Herleitung dieser Formeln ist im Anhang dieses Kapitels dargestellt. Mit den beiden Parametern b_0 und b_1 ist die Regressionsgleichung vollständig bestimmt.

Für die rechnerische Auswertung ist es zweckmäßig, eine Arbeitstabelle anzulegen, wie sie Abbildung 1.15 für unser Beispiel zeigt.

Beobachtung	MENGE	BESUCHE		
k	y_k	x_k	$x \cdot y$	x^2
1	2.585	109	281.765	11.881
2	1.819	107	194.633	11.449
3	1.647	99	163.053	9.801
4	1.496	70	104.720	4.900
5	921	81	74.601	6.561
6	2.278	102	232.356	10.404
7	1.810	110	199.100	12.100
8	1.987	92	182.804	8.464
9	1.612	87	140.244	7.569
10	1.913	79	151.127	6.241
$\sum$	18.068	936	1.724.403	89.370
	$\overline{y}$=1.806,8	$\overline{x}$=93,6		

Abbildung 1.15: Arbeitstabelle

Nach Einsetzen der Werte aus der Arbeitstabelle in die Formeln (1.8) und (1.9) erhält man:

$$b_1 = \frac{10 \cdot 1.724.403 - 936 \cdot 18.068}{10 \cdot 89.370 - (936)^2} = 18,881$$

$$b_0 = 1.806,8 - 18,881 \cdot 93,6 = 39,5$$

Damit ergibt sich die bereits gezeigte Regressionsfunktion

$$\hat{Y} = 39,5 + 18,881$$

die in Abbildung 1.11 und Abbildung 1.12 dargestellt ist.

Der geschätzte Regressionskoeffizient $b_1 = 18,9$ besagt, dass eine Erhöhung der Absatzmenge um 18,9 Einheiten zu erwarten ist, wenn ein zusätzlicher Vertreterbesuch durchgeführt wird. Auf diese Weise kann der Regressionskoeffizient wichtige Hinweise für eine optimale Vertriebsgestaltung geben.

[7] Letzteres ist nicht unbedingt ein Vorteil und kann zu Problemen bei Ausreißern führen. Siehe dazu Belsley/Kuh/Welsch (1980), S. 16 ff.; Fox (2008), S. 79 ff.; Greene (2018), S. 56 ff.

Mit Hilfe der gefundenen Regressionsgleichung ist man, wie schon gezeigt, in der Lage, für beliebige X-Werte die $\hat{Y}$-Werte zu schätzen.

Beispiel: Die Zahl der Vertreterbesuche für Beobachtung Nr. 6 beträgt 102. Wie hoch ist die geschätzte Absatzmenge?

$$\begin{aligned}\hat{y}_6 &= 39,5 + 18,881 \cdot 102 \\ &= 1.965\end{aligned}$$

Beobachtet wurde dagegen eine Absatzmenge von 2.278 Kartons. Das Residuum beträgt demnach $2.278 - 1.965 = 313$.

1.2.2.2 Multiple Regression

Mehrere unabhängige Variable

Für die meisten Untersuchungszwecke ist es erforderlich, mehr als eine unabhängige Variable in das Modell aufzunehmen. Der Regressionsansatz hat dann folgende Form:

$$\hat{Y} = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_j x_j + \ldots + b_J x_J \tag{1.10}$$

Die Ermittlung der Regressionsparameter b_0, b_1, b_2, …, b_J erfolgt wie bei der einfachen Regressionsanalyse durch Minimierung der Summe der Abweichungsquadrate (KQ-Kriterium).

Zielfunktion der multiplen Regressionsanalyse

$$\sum_{k=1}^{K} e_k^2 = \sum_{k=1}^{K} [y_k - (b_0 + b_1 x_{1k} + b_2 x_{2k} + \ldots + b_j x_{jk} + \ldots + b_j x_{Jk})]^2 \rightarrow min \tag{1.11}$$

mit

e_k	=	Werte der Residualgröße $(k = 1, 2, \ldots, K)$
y_k	=	Werte der abhängigen Variablen $(k =, 1, 2, \ldots, K)$
b_0	=	konstantes Glied
b_j	=	Regressionskoeffizienten $(j = 1, 2, \ldots, J)$
x_{jk}	=	Werte der unabhängigen Variablen $(j = 1, 2, \ldots, J; k = 1, 2, \ldots, K)$
J	=	Zahl der unabhängigen Variablen
K	=	Zahl der Beobachtungen

Die Auffindung von Regressionsparametern, die das Zielkriterium (1.11) minimieren, erfordert die Lösung eines linearen Gleichungssystems, die mit erheblichem Rechenaufwand verbunden sein kann.[8]

Wir kommen zurück auf unser Beispiel mit den Daten in Abbildung 1.6. Angenommen, der Manager misst allen drei unabhängigen Variablen (PREIS, AUSGABEN und BESUCHE) eine Relevanz für die Erklärung der Absatzmenge zu. Ihre Berücksichtigung führt dann zu einer multiplen Regressionsanalyse folgender Form:

$$\hat{Y} = b_0 + b_1 \cdot \text{PREIS} + b_2 \cdot \text{AUSGABEN} + b_3 \cdot \text{BESUCHE}$$

[8] Siehe hierzu die Ausführungen im Anhang dieses Kapitels oder die einschlägige Literatur, z. B. Bleymüller/Weißbach (2015), S. 171-180; Greene (2018), S. 13 ff.; Kmenta (1997), S. 395-399; Schneeweiß (1990), S. 94-97.

Die Durchführung der multiplen Regressionsanalyse unter Anwendung des KQ-Kriteriums in Formel (1.11) liefert dann folgende Regressionsfunktion:[9]

$$\hat{Y} = -6,9 + 9,927 \cdot \text{PREIS} + 0,655 \cdot \text{AUSGABEN} + 11,085 \cdot \text{BESUCHE}$$

Betrachten wir beispielsweise den Fall Nr. 6, indem wir die Daten aus Abbildung 1.6 in die erhaltene Regressionsfunktion einsetzen. Man erhält damit als Schätzung für die Absatzmenge:

$$\hat{Y} = -6,9 + 9,927 \cdot 10 + 0,655 \cdot 1500 + 11,085 \cdot 102 = 2.206$$

Da der beobachtete Wert 2.278 ist, beträgt die Residualgröße jetzt nur noch 72. Die Übereinstimmung zwischen beobachtetem und geschätztem Wert hat sich demnach gegenüber der einfachen Regression (Residuum = 313) deutlich verbessert. Die Tatsache, dass sich der Regressionskoeffizient b_1 für die erste unabhängige Variable (BESUCHE) erheblich verringert hat, ist auf die Einbeziehung weiterer unabhängiger Variablen zurückzuführen.

Bedeutung der Regressionskoeffizienten

Interpretation der Regressionskoeffizienten

Die Regressionskoeffizienten besitzen eine wichtige inhaltliche Bedeutung, da sie den marginalen Effekt der Änderung einer unabhängigen Variablen auf die abhängige Variable Y angeben. Für den Manager in unserem Beispiel liefern sie damit wichtige Informationen für seine Maßnahmenplanung. So sagt ihm z. B. der Regressionskoeffizient b_3 = 0,655 für die Variable AUSGABEN, dass er 65,5 Kartons mehr absetzen wird, wenn er die Ausgaben für Verkaufsförderung um 100 erhöht. Bei einem Preis von 10 ergibt dies einen Mehrerlös von 655. Unter Berücksichtigung seiner sonstigen Kosten kann er damit feststellen, ob sich eine Erhöhung der Ausgaben für Verkaufsförderung lohnt.

Skalenabhängigkeit der Regressionskoeffizienten

Die Größe eines Regressionskoeffizienten darf allerdings nicht als Maß für die Wichtigkeit der betreffenden Variablen angesehen werden. Die Werte verschiedener Regressionskoeffizienten lassen sich nur vergleichen, wenn die Variablen in gleichen Einheiten gemessen wurden, denn der numerische Wert b_j ist abhängig von der Skala, auf der die Variable X_j gemessen wurde. So verringert sich z. B. der Regressionskoeffizient für den Preis um den Faktor 100, wenn der Preis anstatt in Euro in Cent gemessen wird. Und die Skala für die Variable BESUCHE ist eine völlig andere als die für den Preis. Um sie vergleichbar zu machen, müsste man sie mit den Kosten pro Besuch in eine monetäre Skala umwandeln und könnte dann mit den so erhaltenen Werten eine erneute Regressionsanalyse durchführen.

Standardisierte Regressionskoeffizienenten

Eine andere Möglichkeit, die Regressionskoeffizienten miteinander vergleichbar zu machen besteht darin, sie zu standardisieren. Die standardisierten Regressionskoeffizienenten, die auch als *Beta-Werte* bezeichnet werden, errechnen sich wie folgt:

$$\hat{b}_j = b_j \cdot \frac{\text{Standardabweichung von } X_j}{\text{Standardabweichung von } Y} \tag{1.12}$$

Durch die Standardisierung werden die unterschiedlichen Messdimensionen der Variablen, die sich in den Regressionskoeffizienten niederschlagen, eliminiert. Letztere

[9] Der Datensatz in Abbildung 1.6 zeigt, dass die Ausgaben im Fall 5 den Wert Null haben. Es handelt sich hier nicht um einen fehlenden Wert, sondern im betreffenden Verkaufsgebiet erfolgten keine Ausgaben. Ansonsten würde sich das Ergebnis verändern.
Die praktische Durchführung der Regressionsanalyse wird nachfolgend für ein etwas umfangreicheres Fallbeispiel unter Anwendung des Computer-Programms IBM SPSS demonstriert.

sind daher unabhängig von linearen Transformationen der Variablen und können so als Maß für deren Wichtigkeit verwendet werden. Bei Durchführung einer Regressionsanalyse mit standardisierten Variablen würde man die Beta-Werte als Regressionskoeffizienten erhalten.

In unserem Beispiel betragen die Standardabweichungen der Variablen Y und X_1 (BESUCHE):[10]

$$\begin{aligned} S_{MENGE} &= 449,23 \\ S_{BESUCHE} &= 13,99 \end{aligned}$$

Damit erhält man den standardisierten Regressionskoeffizienten

$$\hat{b}_1 = 11,085 \cdot \frac{13,99}{449,23} = 0,345$$

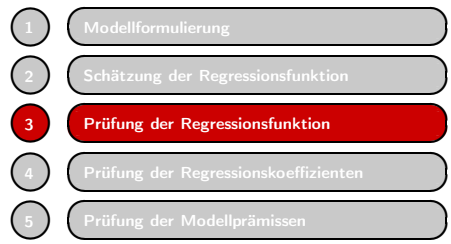
Analog ergeben sich für die Variablen PREIS und AUSGABEN die folgenden Werte:

$$\begin{aligned} S_{PREIS} &= 1,55 & \hat{b}_2 &= 0,034 \\ S_{AUSGABEN} &= 544,29 & \hat{b}_3 &= 0,794 \end{aligned}$$

Es zeigt sich hier, dass die Variable AUSGABEN, die den kleinsten Regressionskoeffizienten hat, den höchsten standardisierten Regressionskoeffizienten aufweist und somit am stärksten auf die Absatzmenge wirkt.[11]

Durch Ermittlung der standardisierten Regressionskoeffizienten werden die nicht standardisierten Regressionskoeffizienten allerdings nicht überflüssig. Da sie den marginalen Effekt der Änderung einer unabhängigen Variablen angeben, haben sie eine wichtige inhaltliche Bedeutung. Zur Durchführung von Wirkungsprognosen sind also weiterhin die unstandardisierten Regressionskoeffizienten zu verwenden.

1.2.3 Prüfung der Regressionsfunktion



Nachdem die Regressionsfunktion geschätzt wurde, ist deren Güte zu überprüfen, d. h. es ist zu klären, wie gut sie als Modell der Realität geeignet ist. Die Überprüfung lässt sich in zwei Bereiche gliedern:

1. Globale Prüfung der Regressionsfunktion
   Hier geht es um die Prüfung der Regressionsfunktion als Ganzes, d. h. ob und wie gut die abhängige Variable Y durch das Regressionsmodell erklärt wird (goodness of fit).

[10] Die Standardabweichung einer Variablen X berechnet sich durch:

$$s_X = \sqrt{\frac{\sum_{k=1}^{K} (x_k - \overline{x})^2}{K-1}}$$

[11] Bei der Beurteilung der Wichtigkeit von unabhängigen Variablen mit Hilfe der Beta-Werte ist allerdings Vorsicht geboten, da ihre Aussagekraft durch Multikollinearität (Korrelation zwischen den unabhängigen Variablen) stark beeinträchtigt werden kann.

2. Prüfung der Regressionskoeffizienten
   Hier geht es um die Frage, ob und wie gut einzelne Variablen des Regressionsmodells zur Erklärung der abhängigen Variablen Y beitragen.

Wenn sich aufgrund der Prüfung der Regressionskoeffizienten zeigt, dass eine Variable keinen Beitrag zur Erklärung leistet, so ist diese aus der Regressionsfunktion zu entfernen. Zuvor aber ist die globale Güte zu überprüfen. Erweist sich das Modell insgesamt als unbrauchbar, so erübrigt sich eine Überprüfung der einzelnen Regressionskoeffizienten.

Globale Gütemaße zur Prüfung der Regressionsfunktion sind

- das Bestimmtheitsmaß (R^2),
- die F-Statistik,
- der Standardfehler.

Gütemaße

Maße zur Prüfung der Regressionskoeffizienten sind

- der t-Wert,
- der Beta-Wert.

Nachfolgend soll zunächst auf die globalen Gütemaße und in Abschnitt 1.2.4 sodann auf die Maße zur Prüfung der Regressionskoeffizienten eingegangen werden.

1.2.3.1 Streuungszerlegung und Bestimmtheitsmaß

Wie gezeigt wurde, wird die optimale Regressionsfunktion durch Minimierung der Summe der quadrierten Residuen gefunden, die in Anlehnung an englische Begriffe auch mit SSR („sum of squared residuals“) abgekürzt wird:

$$\text{SSR} = \sum_{k=1}^{K} e_k^2$$

Man könnte diese Summe als Maß für die Güte der Anpassung einer Regressionsfunktion an die beobachteten Daten („goodness of fit“) verwenden: Je kleiner SSR, desto besser die Anpassung. Das gilt allerdings nur für einen gegebenen Datensatz, denn SSR variiert neben der Anpassungsgüte auch mit der Anzahl und Größe der Y-Werte. Für unser Beispiel gilt für die einfache Regression SSR = 1.188.685 (vgl. Abbildung 1.17 und für die multiple Regression SSR = 135.227. Durch die Erweiterung des Modells hat sich also die Anpassungsgüte erheblich verbessert.

Wenn auch die Verringerung von SSR hier deutlich ist, so kann man doch nicht sagen, wie gut oder schlecht die Werte sind. SSR findet zwar Eingang in alle oben genannten globalen Gütemaße, isoliert gesehen aber ist diese Größe als Gütemaß nicht geeignet. Vielmehr ist es erforderlich, SSR mit anderen Größen in Beziehung zu setzen. Eine geeignete Basis hierfür liefert das in der Statistik wichtige Prinzip der Streuungszerlegung (decomposition of variation)[12]. Zu seiner Erläuterung gehen wir auf die einfache Regressionsanalyse zurück, die Beziehung zwischen Absatzmenge und Zahl der Vertreterbesuche.

[12] Dieses Prinzip ist auch grundlegend für die Varianzanalyse (Kapitel 3) und die Diskriminanzanalyse (Kapitel 4).

Zerlegung der Gesamtabweichung

Wir betrachten zunächst eine einzelne Beobachtung x_k, die Absatzmenge y_k für die Zahl der Vertreterbesuche x_k (vgl. Abbildung 1.16). Der zugehörige Punkt (x_k, y_k) liegt oberhalb des Mittelwertes $\overline{y}$. Die Abweichung vom Mittelwert, $y_k - \overline{y}$, bezeichnen wir als Gesamtabweichung.

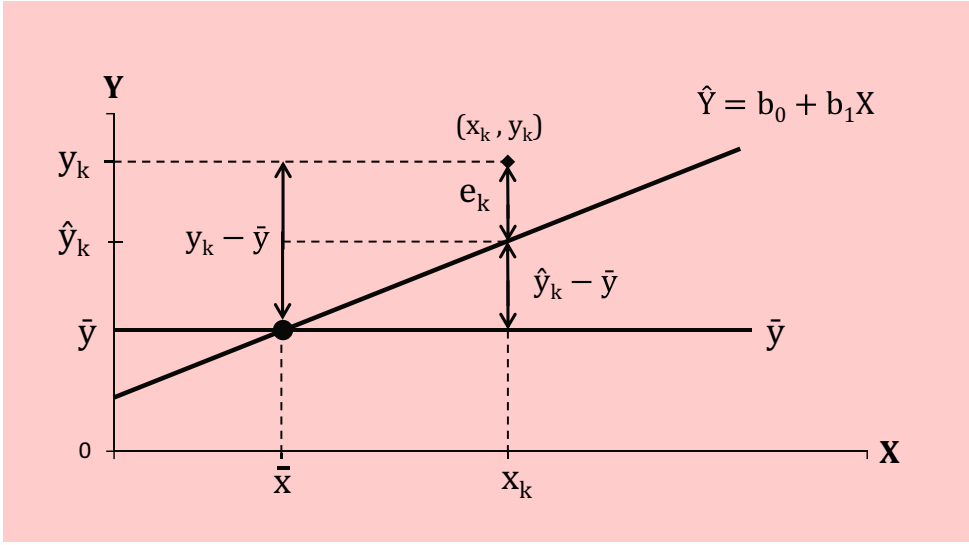


Abbildung 1.16: Zerlegung der Gesamtabweichungen vom Mittelwert

Ein Teil dieser Abweichung, nämlich $\hat{y}_k - \overline{y}$, lässt sich durch die Regressionsgerade erklären, da die Zahl der Vertreterbesuche x_k über dem Durchschnitt $\overline{x}$ liegt.
Es gilt daher:

$$\hat{y}_k - \overline{y} = b_1(x_k - \overline{x})$$

Der Rest, das Residuum e_k, kann dagegen nicht erklärt werden.

Die Gesamtabweichung einer einzelnen Beobachtung y_k lässt sich damit wie folgt zerlegen:

$$\begin{aligned} &\text{Gesamtabweichung} = \text{Erklärte Abweichung} + \text{Residuum} \\ &y_k - \overline{y} \qquad\qquad = (\hat{y}_k - \overline{y}) \qquad\qquad + (y_k - \hat{y}_k) \end{aligned}$$

Dies ist recht trivial. Nicht trivial ist dagegen, dass die Gleichung noch gilt, wenn man die Elemente quadriert und über die Beobachtungen summiert.[13] Man erhält damit die folgende Zerlegung der *Gesamtstreuung*.

Zerlegung der Gesamtstreuung

$$\begin{array}{lcllcll} \text{Gesamtstreuung} & = & \text{erklärte Streuung} & + & \text{nicht erklärte Streuung} & & \\ \sum_{k=1}^{K}(y_k - \overline{y})^2 & = & \sum_{k=1}^{K}(\hat{y}_k - \overline{y})^2 & + & \sum_{k=1}^{K}(y_k - \hat{y}_k)^2 & & (1.13) \\ \text{SST} & = & \text{SSE} & + & \text{SSR} & & \end{array}$$

[13] Dies ergibt sich aufgrund der KQ-Schätzung und gilt nur für lineare Modelle. Zum Beweis siehe z. B. Kmenta (1997), S. 239.

Dabei steht SST für „total sum of squares“ und SSE für „explained sum of squares“.[14]
Auf Basis der Sreuungszerlegung lässt sich als Maß für die Güte der Anpassung das *Bestimmtheitsmaß* berechnen. Es wird mit R^2 (R-Quadrat, R squared) bezeichnet und ergibt sich aus dem Verhältnis von erklärter Streuung zur Gesamtstreuung:

Bestimmtheitsmaß

$$R^2 = \frac{\text{SSE}}{\text{SST}} = \frac{\sum_{k=1}^{K} (\hat{y}_k - \overline{y})^2}{\sum_{k=1}^{K} (y_k - \overline{y})^2} = \frac{\text{erklärte Streuung}}{\text{Gesamtstreuung}} \tag{1.14}$$

Das Bestimmtheitsmaß ist eine normierte Größe, dessen Wertebereich zwischen Null und Eins liegt. Es ist um so größer, je höher der Anteil der erklärten Streuung an der Gesamtstreuung ist. Im Extremfall, wenn die gesamte Streuung erklärt wird, ist $R^2 = 1$, im anderen Extremfall entsprechend $R^2 = 0$.

Man kann das Bestimmtheitsmaß auch durch Subtraktion des Verhältnisses der nicht erklärten Streuung zur Gesamtstreuung vom Maximalwert 1 ermitteln, was rechentechnisch von Vorteil ist, da die nicht erklärte Streuung leicht zu berechnen ist und meist ohnehin vorliegt:

Bestimmtheitsmaß

$$R^2 = 1 - \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\sum_{k=1}^{K} e_k^2}{\sum_{k=1}^{K} (y_k - \overline{y})^2} \tag{1.15}$$

Aus Formel (1.15) wird deutlich, dass das Kleinstquadrate-Kriterium, das zur Schätzung der Regressionsbeziehung angewendet wird, gleichbedeutend mit der Maximierung des Bestimmtheitsmaßes ist. Je kleiner SSR, desto größer wird R^2. Mit den Werten für unser Beispiel, die in Abbildung 1.17 zusammengestellt sind, erhält man mit Formel (1.15) für das Bestimmtheitsmaß:

$$R^2 = 1 - \frac{1.188.683,49}{1.816.255,60} = 0,346$$

Das Ergebnis besagt, dass rund 35 % der gesamten Streuung durch die Variable BESUCHE erklärt werden kann, während 65 % unerklärt bleiben. Die Schwankungen der Absatzmenge Y sind also zu einem großen Anteil auf andere Einflüsse, die in der Regressionsgleichung nicht erfasst wurden, zurückzuführen.

Das Bestimmtheitsmaß lässt sich alternativ auch als Quadrat der Korrelation $R = r_{Y\hat{Y}}$ zwischen den beobachteten und den geschätzten Y-Werten berechnen (hieraus resultiert die Bezeichnung R-Quadrat). Es besteht in dieser Hinsicht kein Unterschied zwischen einfacher und multipler Regressionsanalyse. Da aber bei der multiplen Regression $\hat{Y}$ durch lineare Verknüpfung von mehreren unabhängigen Variablen gebildet wird, bezeichnet man in diesem Fall R als *multiplen Korrelationskoeffizienten.*

[14] Es sei darauf hingewiesen, dass in der Literatur die Abkürzungen SSE und SSR manchmal mit umgekehrter Bedeutung gebraucht werden. Es steht dann E für „error“ und R für „regression“.

k	y_k	$\hat{y}_k$	$y_k - \hat{y}_k$	$(y_k - \hat{y}_k)^2$	$y_k - \overline{y}$	$(y_k - \overline{y})^2$
1	2.585	2.097,57	487,43	237.588,00	778,20	605.595,24
2	1.819	2.059,81	-240,81	57.989,46	12,20	148,84
3	1.647	1.908,76	-261,76	68.518,30	-159,80	25.536,04
4	1.496	1.361,21	134,79	18.168,34	-310,80	96.596,64
5	921	1.568,90	-647,90	419.774,41	-885,80	784.641,64
6	2.278	1.965,40	312,60	97.718,76	471,20	222.029,44
7	1.810	2.116,45	-306,45	93.911,60	3,20	10,24
8	1.987	1.776,59	210,41	44.272,37	180,20	32.472,04
9	1.612	1.682,19	-70,19	4.926,64	-194,80	37.947,04
10	1.913	1.531,14	381,86	145.817,06	106,20	11.278,44
$\overline{y}$	1.806,8					
$\sum$				SSR = 1.188.683,49		SST = 1.816.255,60

Abbildung 1.17: Aufbereitung der Daten für die Ermittlung des Bestimmtheitsmaßes

Das Bestimmtheitsmaß als Anteil der erklärten Streuung ist ein sehr anschauliches Gütemaß. Als alleiniges Kriterium zur Beurteilung der Güte eines Modell ohne weitere Informationen aber ist es nicht ausreichend. Im Extremfall bei nur zwei Beobachtungen würde eine einfache Regression immer ein Bestimmtheitsmaß von Eins liefern, da sich eine Gerade immer ohne Abweichungen durch zwei Punkte legen lässt. Zu einem so „geschätzten“ Modell aber könnte man kein Vertrauen haben. Das Bestimmtheitsmaß berücksichtigt nicht die Anzahl K der Beobachtungen (Größe der Stichprobe).

Ein weiterer Nachteil ist, dass das Bestimmtheitsmaß auch nicht die Komplexität eines Modells berücksichtigt. Ein komplexeres Modell mit vielen Variablen (großem J) wird immer eine bessere Anpassung an die Daten erzielen, aber nicht unbedingt bessere Schätzwerte. Denn mit steigender Anzahl J der unabhängigen Variablen sinkt die Zahl der *Freiheitsgrade* $df = K - J - 1$ für die Schätzung. Jeder zu schätzende Parameter verbraucht einen Freiheitsgrad.

In unserem Beispiel liefert die multiple Regression mit drei Regressoren für das Bestimmtheitsmaß den Wert 0,926 gegenüber 0,346 für die einfache Regression mit der Variable Besuche. Das ist eine erhebliche Vergrößerung. Mit jedem hinzukommenden Regressor wird ein mehr oder weniger großer Erklärungsanteil hinzugefügt, der möglicherweise nur zufällig bedingt ist. Der Wert des Bestimmtheitsmaßes kann daher mit der Aufnahme von weiteren Regressoren, auch wenn diese möglicherweise irrelevant sind, nur zunehmen, aber nicht abnehmen. Bei kleiner Zahl von Beobachtungen ist dabei die relative Verkleinerung der Freiheitsgrade erheblich. Im Beispiel verringert sich die Zahl der Freiheitsgrade von 8 auf 6, also um 25 %.

Maße, die die Größe der Stichprobe und die Zahl der Regressoren berücksichtigen, sind das korrigierte Bestimmtheitsmaß und die F-Statistik, die im folgenden Abschnitt behandelt werden. Beim korrigierten Bestimmtheitsmaß gemäß Formel (1.16) wird das einfache Bestimmtheitsmaß um eine Korrekturgröße verringert, die um so größer ist, je größer die Zahl der Regressoren und je kleiner die Zahl der Freiheitsgrade ist. Das korrigierte Bestimmtheitsmaß kann daher im Gegensatz zum einfachen Bestimmtheitsmaß durch die Aufnahme weiterer Regressoren auch abnehmen.[15]

[15] Vgl. z. B. Wooldridge (2016), S. 181; Greene (2018), S. 47 .

Korrigiertes Bestimmtheitsmaß

$$R^2_{korr} = R^2 - \frac{J \cdot (1 - R^2)}{K - J - 1} \tag{1.16}$$

mit

$$\begin{aligned} K &= \text{Zahl der Beobachtungswerte} \\ J &= \text{Zahl der Regressoren} \\ K - J - 1 &= \text{Zahl der Freiheitsgrade} \end{aligned}$$

Für unser Beispiel liefert das korrigierte Bestimmtheitsmaß im Fall der multiplen Regression den Wert 0,888 gegenüber 0,926 für das einfache Bestimmtheitsmaß.

Das korrigierte Bestimmtheitsmaß ist immer kleiner oder maximal gleich dem einfachen Bestimmtheitsmaß und es kann auch negativ werden. Übergroße Modellkomplexität soll durch die Korrektur bestraft werden. Manche Autoren argumentieren allerdings, dass dieser Bestrafungseffekt noch zu gering ist.[16]

1.2.3.2 Stochastisches Modell und F-Test

Das Bestimmtheitsmaß drückt aus, wie gut sich die Regressionsfunktion an die beobachteten Daten anpasst. In empirischen Untersuchungen wird die Regressionsanalyse aber nicht nur deskriptiv zur Beschreibung vorliegender Daten eingesetzt. Vielmehr handelt es sich i. d. R. um Daten einer Stichprobe und es stellt sich die Frage, ob das geschätzte Modell auch über die Stichprobe hinaus für die Grundgesamtheit Gültigkeit besitzt. Eine notwendige Bedingung hierfür bildet neben der Repräsentanz der Stichprobe die Signifikanz des geschätzten Modells (bzw. seines Bestimmtheitsmaßes). Zur Signifikanzprüfung verwendet man die F-Statistik, in deren Berechnung neben der obigen Streuungszerlegung zusätzlich auch der Umfang der Stichprobe und die Zahl der Regressoren eingeht.

Die geschätzte Regressionsfunktion (Regressionsfunktion der Stichprobe)

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_j X_j + \ldots + b_J X_J + e$$

lässt sich als Realisation einer „wahren“ Funktion mit den unbekannten Parametern $\beta_0, \beta_1, \beta_2, \ldots, \beta_J$ auffassen, die den Wirkungszusammenhang in der Grundgesamtheit wiedergibt. Da diese Funktion neben dem systematischen Einfluss der Variablen $X_1, X_2, \ldots, X_J$, die auf Y wirken, auch eine Zufallsgröße u (stochastische Komponente) enthält, bezeichnet man sie als das stochastische Modell der Regressionsanalyse.

Stochastisches Modell der Regressionsanalyse

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_j X_j + \ldots + \beta_J X_J + u \tag{1.17}$$

mit

$$\begin{aligned} Y &= \text{Abhängige Variable} \\ \beta_0 &= \text{Konstantes Glied der Regressionsfunktion} \\ \beta_j &= \text{Regressionskoeffizient } (j = 1, 2, \ldots, J) \\ X_j &= \text{Unabhängige Variable } (j = 1, 2, \ldots, J) \\ u &= \text{Störgröße} \end{aligned}$$

[16] Siehe dazu Fahrmeir/Kneib/Lang (2009), S. 161.

In der Größe u ist die Vielzahl zufälliger Einflüsse, die neben dem systematischen Einfluss der Variablen $X_1, X_2, \ldots, X_J$ auf Y wirken, zusammengefasst. Sie ist eine Zufallsvariable und wird als *Störgröße* bezeichnet, da sie den systematischen Einfluss überlagert und damit verschleiert. Die Störgröße u ist nicht beobachtbar, manifestiert sich aber in den Residuen e_k.

Störgröße

Da in der abhängigen Variablen Y die Störgröße u enthalten ist, bildet Y ebenfalls eine Zufallsvariable, und auch die Schätzwerte b_j für die Regressionsparameter, die aus Beobachtungen von Y gewonnen wurden, sind Realisationen von Zufallsvariablen. Bei wiederholten Stichproben schwanken diese um die wahren Werte β_j.

Wenn zwischen der abhängigen Variablen Y und den unabhängigen Variablen X_j ein kausaler Zusammenhang besteht, wie es hypothetisch postuliert wurde, so müssen die wahren Regressionskoeffizienten β_j ungleich Null sein. Zur Prüfung des Modells wird jetzt die Hypothese H_0 („Nullhypothese“) formuliert, die besagt, dass kein Zusammenhang besteht und somit in der Grundgesamtheit die Regressionskoeffizienten alle Null sind:

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_J = 0$$

F-Test

Zur Prüfung dieser Nullhypothese kann ein *F-Test* verwendet werden. Er besteht im Kern darin, dass ein empirischer F-Wert (F-Statistik) berechnet und mit einem kritischen Wert verglichen wird. Bei Gültigkeit der Nullhypothese ist zu erwarten, dass der F-Wert Null ist. Weicht er dagegen stark von Null ab und überschreitet einen kritischen Wert, so ist es unwahrscheinlich, dass die Nullhypothese richtig ist. Folglich ist diese zu verwerfen und zu folgern, dass in der Grundgesamtheit ein Zusammenhang existiert und somit nicht alle β_j Null sind.

In die Berechnung der F-Statistik gehen die Streuungskomponenten ein (wie in das Bestimmtheitsmaß) und zusätzlich der Stichprobenumfang K und die Zahl der Regressoren J. Sie berechnet sich wie folgt:

F-Statistik

$$F_{emp} = \frac{\sum_{k=1}^{K} (\hat{y}_k - \overline{y})^2 / J}{\sum_{k=1}^{K} (y_k - \hat{y}_k)^2 / (K - J - 1)} = \frac{\text{erklärte Streuung}/J}{\text{nicht erklärte Streuung}/(K - J - 1)} \tag{1.18}$$

Zur Berechnung sind die erklärte und die nicht erklärte Streuung jeweils durch die Zahl ihrer *Freiheitsgrade* zu dividieren und ins Verhältnis zu setzen. Die Zahl der Freiheitsgrade der

Freiheitsgrade

- erklärten Streuung ist gleich der Zahl der unabhängigen Variablen: J
- nicht erklärten Streuung ist gleich der Zahl der Beobachtungen vermindert um die zu schätzenden Parameter in der Regressionsbeziehung: K-J-1.

Mit Hilfe von (1.14) lässt sich die F-Statistik auch als Funktion des Bestimmtheitsmaßes formulieren:

$$F_{emp} = \frac{R^2/J}{(1 - R^2)/(K - J - 1)} \tag{1.19}$$

Der **F-Test** läuft in folgenden Schritten ab:

1. Berechnung des empirischen F-Wertes
   Empirischer F-Wert
   Im Beispiel hatten wir für das Bestimmtheitsmaß den Wert $R^2 = 0{,}3455$ errechnet. Mittels Formel (1.19) erhält man:

$$F_{emp} = \frac{0,3455/1}{(1-0,3455)/(10-1-1)} = 4,224$$

   Der Leser möge alternativ die Berechnung mittels Formel (1.18) durchführen.

2. Vorgabe eines Signifikanzniveaus
   Es ist, wie bei allen statistischen Tests, eine Wahrscheinlichkeit vorzugeben, die das Vertrauen in die Verlässlichkeit des Testergebnisses ausdrückt. Üblicherweise wird hierfür die *Vertrauenswahrscheinlichkeit* 0,95 (oder auch 0,99) gewählt. Vertrauens-wahrscheinlichkeit
   Das bedeutet: Mit einer Wahrscheinlichkeit von 95 Prozent kann man sich darauf verlassen, dass der Test zu einer Annahme der Nullhypothese führen wird, wenn diese korrekt ist, d. h. wenn kein Zusammenhang besteht.
   Entsprechend beträgt die Wahrscheinlichkeit, dass die Nullhypothese abgelehnt wird, obgleich sie richtig ist, $\alpha = 1 - 0,95 = 5$ Prozent. α ist die *Irrtumswahrscheinlichkeit* des Tests und wird als *Signifikanzniveau* bezeichnet. Die Irrtumswahrscheinlichkeit bildet das Komplement der Vertrauenswahrscheinlichkeit $1 - \alpha$. Irrtums-wahrscheinlichkeit

3. Auffinden des theoretischen F-Wertes
   Als kritischer Wert zur Prüfung der Nullhypothese dient ein theoretischer F-Wert, mit dem der empirische F-Wert zu vergleichen ist. Dieser ergibt sich für das gewählte Signifikanzniveau aus der F-Verteilung und kann aus einer *F-Tabelle* entnommen werden. Abbildung 1.18 zeigt einen Ausschnitt aus der F-Tabelle für die Vertrauenswahrscheinlichkeit 0,95 (vgl. Anhang A.3). Theoretischer F-Wert

   Der gesuchte Wert ergibt sich durch die Zahl der Freiheitsgrade im Zähler und im Nenner von Formel (1.18) oder (1.19). Die Zahl der Freiheitsgrade im Zähler ($J = 1$) bestimmt die Spalte und die der Freiheitsgrade im Nenner ($K - J - 1 = 8$) bestimmt die Zeile der Tabelle und man erhält den Wert 5,32. Der tabellierte Wert bildet das 95 %-Quantil der F-Verteilung mit der betreffenden Zahl von Freiheitsgraden, d. h. Werte dieser Verteilung sind mit 95 % Wahrscheinlichkeit kleiner als der tabellierte Wert.

4. Vergleich des empirischen mit dem theoretischen F-Wert
   Das Entscheidungskriterium für den F-Test lautet: Entscheidungs-kriterium

- Ist der empirische F-Wert (F_{emp}) größer als der aus der Tabelle abgelesene theoretische F-Wert (F_{tab}), dann kann die die Nullhypothese H_0 verworfen werden. Es ist also zu folgern, dass nicht alle β_j Null sind. Der durch die Regressionsbeziehung hypothetisch postulierte Zusammenhang wird damit als statistisch signifikant erachtet.
- Ist dagegen der empirische F-Wert klein und übersteigt nicht den theoretischen Wert, so kann die Nullhypothese nicht verworfen werden. Die Regressionsbeziehung ist damit nicht signifikant.

K-J-1	J=1	J=2	J=3	J=4	J=5	J=6	J=7	J=8	J=9
1	161,00	200,00	216,00	225,00	230,00	234,00	237,00	129,00	241,00
2	18,50	19,00	19,20	19,20	19,30	19,30	19,40	19,40	19,40
3	10,10	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68
8	**5,32**	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02

Legende:
J = Zahl der erklärenden Variablen (Freiheitsgerade des Zählers);
$K - J - 1$ = Zahl der Freiheitsgerade des Nenners (K = Zahl der Beobachtungen)

Abbildung 1.18: F-Tabelle (95 % Vertrauenswahrscheinlichkeit; Ausschnitt)

Zusammenfassend gilt für den F-Test:

$$F_{emp} > F_{tab} \rightarrow H_0 \quad \text{wird verworfen} \rightarrow \text{Zusammenhang ist signifikant}$$
$$F_{emp} \leq F_{tab} \rightarrow H_0 \quad \text{wird nicht verworfen}$$

Hier ergibt sich:

$$4,2 < 5,32 \qquad \rightarrow \qquad H_0 \text{ wird nicht verworfen!}$$

Da der empirische F-Wert hier kleiner ist als der Tabellenwert, kann die Nullhypothese nicht verworfen werden. Das bedeutet, dass der durch die Regressionsbeziehung postulierte Zusammenhang empirisch nicht bestätigt werden kann, d. h. er ist statistisch nicht signifikant.

Interpretation

Dies bedeutet allerdings nicht, dass kein Zusammenhang zwischen der Zahl der Vertreterbesuche und der Absatzmenge besteht. Möglicherweise ist dieser durch andere Einflüsse überlagert und wird damit infolge des geringen Stichprobenumfangs nicht deutlich. Oder er wird nicht deutlich, weil relevante Einflussgrößen (wie hier der Preis oder die Ausgaben für Verkaufsförderung) nicht berücksichtigt wurden und deshalb die nicht erklärte Streuung groß ist.

Prinzipiell kann die Annahme einer Nullhypothese nicht als Beweis für deren Richtigkeit angesehen werden. Sie ließe sich andernfalls immer beweisen, indem man den Stichprobenumfang klein macht und/oder die Vertrauenswahrscheinlichkeit hinreichend groß wählt. Nur umgekehrt kann die Ablehnung der Nullhypothese als Beweis dafür angesehen werden, dass diese falsch ist und somit ein Zusammenhang besteht. Damit wird auch deutlich, dass es keinen Sinn macht, die Vertrauenswahrscheinlichkeit zu groß (die Irrtumswahrscheinlichkeit zu klein) zu wählen, denn dies würde dazu führen, dass die Nullhypothese, auch wenn sie falsch ist, nicht abgelehnt wird und somit bestehende Zusammenhänge nicht erkannt werden. Man sagt dann, dass der Test an „Trennschärfe“ verliert.

Die zweckmäßige Wahl der Vertrauenswahrscheinlichkeit sollte berücksichtigen, welches Maß an Unsicherheit im Untersuchungsbereich besteht. Und sie sollte auch berücksichtigen, welche Risiken mit der fälschlichen An- oder Ablehnung der Nullhypothese verbunden sind. So wird man beim Bau einer Brücke eine andere Vertrauenswahrscheinlichkeit wählen als bei der Untersuchung von Kaufverhalten. Letztlich aber ist die Wahl der Vertrauenswahrscheinlichkeit immer mit einem gewissen Maß an Willkür behaftet.

Testdurchführung mittels p-Wert

Die Testdurchführung lässt sich erheblich vereinfachen, wenn man den p-Wert (prob value) der F-Statistik verwendet. Man kann dann auf Tabellen verzichten und gewinnt zusätzliche Information. Der *p-Wert der F-Statistik* ist definiert als die Wahrscheinlichkeit, dass eine F-verteilte Zufallsvariable F (mit df1 und df2 Freiheitsgraden) größer ist als der empirisch ermittelte F-Wert F_{emp}.

$$p = Pr(F > F_{emp})$$

Das Entscheidungskriterium für den F-Test lautet:

$$p < \alpha \quad \rightarrow \quad H_0 \text{ wird verworfen}$$

Andernfalls, also falls $p \geq \alpha$, muss die Nullhypothese beibehalten werden. Im Beispiel erhält man für $F_{emp} = 4,224$ den p-Wert $p = 0,074$. Da $p > \alpha$, kann H_0 nicht verworfen werden. Das Ergebnis ist natürlich identisch mit der oben gezeigten klassischen Testdurchführung.

Die statistischen Tabellen entstammen einer Zeit, als die Berechnung von p-Werten noch zu aufwendig war. Mittels heutiger Computer aber bildet sie kein Problem mehr.[17] Die heutigen Statistik-Programme weisen den p-Wert von wichtigen Statistiken i. d. R. automatisch aus.[18]

Wie beim klassischen Hypothesentest muss auch beim Test mittels p-Wert ein Signifikanzniveau α vom Untersucher vorgegeben werden, mit dem der p-Wert zu vergleichen ist. Die Wahl des „richtigen“ α ist letztlich ein ungelöstes Problem. Durch den p-Wert wird das Problem aber gemildert. Er wird im Unterschied zu α auch als *empirisches Signifikanzniveau* bezeichnet und ist im Gegensatz zum empirischen F-Wert gut interpretierbar. Letzterer kann beliebige positive Werte annehmen und ist auch von den Freiheitsgraden abhängig. Der p-Wert ist dagegen eine genormte Größe mit Werten zwischen 0 und 1. Man kann sich mit ihm sofort ein Urteil über die Glaubwürdigkeit der Nullhypothese bilden, ohne sich zuvor auf ein bestimmtes α festgelegt zu haben. Niedrige Werte sprechen gegen die Gültigkeit der Nullhypothese.

Im Beispiel kann der Untersucher bei Kenntnis von $p = 0,074$ z. B. folgern, dass die Regressionsbeziehung bei $\alpha = 0,10$ statistisch signifikant wäre, und auch bei $\alpha = 0,08$, nicht aber bei $\alpha = 0,07$. Also, wird α kleiner als $0,074$, dann kann H_0 hier nicht mehr abgelehnt werden. Der p-Wert lässt sich daher auch interpretieren als das minimale Signifikanzniveau, bei dem die Nullhypothese abgelehnt werden kann.

[17] In Excel erhält man den p-Wert der F-Verteilung durch die Funktion FVERT(x;df1;df2). Es gilt damit FVERT(4,224;1;8) = 0,074.
Umgekehrt könnte man mit Excel den Tabellenwert für $\alpha = 0,05$ durch FINV(0,05;1;8) = 5,318 berechnen.

[18] In SPSS wird der p-Wert unter der Bezeichnung „Signifikanz“ oder „Sig.“ ausgegeben.

Mit dem p-Wert erhält der Untersucher damit zusätzliche Information. Während der klassische Hypothesentest nur eine Schwarz-Weiß-Betrachtung für ein gegebenes α liefert (Annahme oder Ablehnung), ermöglicht der p-Wert eine differenziertere Betrachtung. Er zeigt, für welche Werte der Irrtumswahrscheinlichkeit α die Nullhypothese verworfen werden kann.

1.2.3.3 Standardfehler der Schätzung

Ein weiteres Gütemaß bildet der Standardfehler der Schätzung, der angibt, welcher mittlere Fehler bei Verwendung der Regressionsfunktion zur Schätzung der abhängigen Variablen Y gemacht wird. Er errechnet sich wie folgt:

$$s = \sqrt{\frac{\sum e_k^2}{K - J - 1}} \tag{1.20}$$

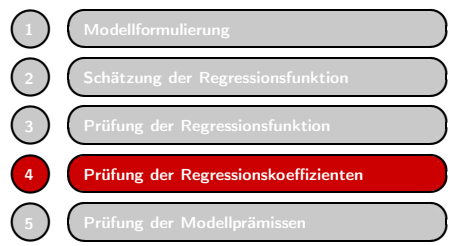
Im Beispiel ergibt sich mit dem Wert der nicht erklärten Streuung SSR aus Abbildung 1.17:

$$s = \sqrt{\frac{1.188.683}{10 - 1 - 1}} = 385$$

Bezogen auf den Mittelwert $\overline{y} = 1.806,8$ beträgt der Standardfehler der Schätzung damit 21 %, was wiederum nicht als gut beurteilt werden kann.

1.2.4 Prüfung der Regressionskoeffizienten

1.2.4.1 t-Test des Regressionskoeffizienten



Nullhypothese

Wenn die globale Prüfung der Regressionsfunktion durch den F-Test ergeben hat, dass nicht alle Regressionskoeffizienten β_j Null sind (und somit ein Zusammenhang in der Grundgesamtheit besteht), sind jetzt die Regressionskoeffizienten einzeln zu überprüfen. Üblicherweise wird auch hier wieder die Nullhypothese H_0: $\beta_j = 0$ getestet. Prinzipiell jedoch könnte auch jeder andere Wert getestet werden. Ein geeignetes Prüfkriterium hierfür ist die t-Statistik.

t-Statistik

$$t_{emp} = \frac{b_j - \beta_j}{s_{bj}} \tag{1.21}$$

mit

t_{emp}	=	Empirischer t-Wert für den j-ten Regressor
β_j	=	Wahrer Regressionskoeffizient (unbekannt)
b_j	=	Regressionskoeffizient des j-ten Regressors
s_{bj}	=	Standardfehler von b_j

Wird die Nullhypothese H_0: $\beta_j = 0$ getestet, so vereinfacht sich (1.21) zu

$$t_{emp} = \frac{b_j}{s_{bj}} \tag{1.22}$$

Der t-Wert einer unabhängigen Variablen errechnet sich also sehr einfach, indem man ihren Regressionskoeffizienten durch dessen Standardfehler dividiert. Diese Größe wird in den gängigen Computer-Programmen für Regressionsanalysen standardmäßig angegeben.[19] t-Test

Unter der Nullhypothese folgt die t-Statistik einer t-Verteilung (Student-Verteilung) um den Mittelwert Null, die in schematischer Form in Abbildung 1.20 dargestellt ist, und zwar mit den Quantilen für den zweiseitigen t-Test[20]. Die Werte für das rechte (positive) Quantil $t_{\alpha/2}$ sind in der t-Tabelle im Anhang für unterschiedliche Werte der Irrtumswahrscheinlichkeit α (Signifikanzniveau) und Freiheitsgrade wiedergegeben. Da die Verteilung symmetrisch ist, sind die Absolutwerte von linkem und rechtem Quantil gleich. Einen Ausschnitt aus der t-Tabelle zeigt Abbildung 1.20.

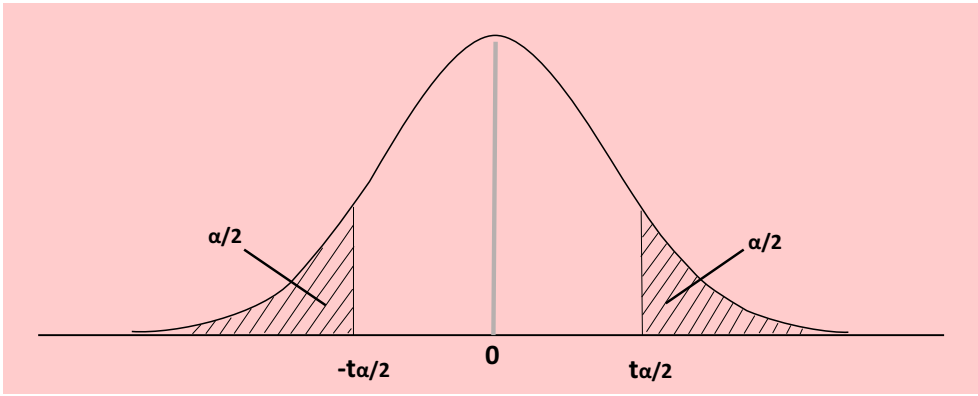


Abbildung 1.19: t-Verteilung und Quantile für Irrtumswahrscheinlichkeit α

Die Quantile markieren den Bereich, innerhalb dessen eine t-verteilte Zufallsvariable mit einer Wahrscheinlichkeit $1 - \alpha$ (Vertrauenswahrscheinlichkeit) variieren wird. Je größer dieser Bereich ist, desto größer wird auch die Wahrscheinlichkeit, dass eine Realisation der Zufallsvariablen in diesen Bereich fällt. Die Wahrscheinlichkeit entspricht der Fläche unter der Glockenkurve.

Analog gilt: Die Wahrscheinlichkeit dafür, dass ein Wert außerhalb der Quantile (unter dem schraffierten Bereich) realisiert wird, beträgt α (Irrtumswahrscheinlichkeit) und entspricht der schraffierten Fläche (links + rechts) unter der Glockenkurve. Je weiter die Quantile nach außen rücken, desto kleiner wird α. Die t-Tabelle liefert umgekehrt für ausgewählte Werte von α (z. B. 1%, 5%, 10%) die zugehörigen Quantile.

[19] Zur Berechnung des Standardfehlers des Regressionskoeffizienten siehe die Ausführungen im mathematischen Anhang dieses Kapitels, insbesondere die Formeln (A15) und (B10).

[20] In der Regressionsanalyse ist der zweiseitige t-Test üblich. Generell ist aber, soweit möglich, dem einseitigen Test der Vorzug zu geben. Siehe hierzu die Literatur zur Statistik, z. B. Fahrmeir et al. (2016), S. 437 ff.; Bortz/Schuster (2010), S. 183 ff.; Bleymüller/Weißbach (2015), S. 133 ff.

Wiederum gilt, dass bei Gültigkeit der Nullhypothese für die t-Statistik ein Wert von Null zu erwarten ist. Weicht der empirische t-Wert dagegen stark von Null ab, so ist die Wahrscheinlichkeit dafür, dass die Nullhypothese richtig ist, gering. Liegt der Wert außerhalb der Quantile, so ist diese Wahrscheinlichkeit kleiner als die akzeptierte Irrtumswahrscheinlichkeit α. Folglich ist dann die Nullhypothese zu verwerfen und zu folgern, dass ein Einfluss von X_j auf Y existiert und somit β_j ungleich Null ist.

Freiheitsgrade	Irrtumswahrscheinlichkeit α		
FG	0,10	0,05	0,01
1	6,314	12,706	63,657
2	2,920	4,303	9,925
3	2,353	3,182	5,841
4	2,132	2,776	4,604
5	2,015	2,571	4,032
6	1,943	2,447	3,707
7	1,895	2,365	3,499
8	1,860	**2,306**	3,355
9	1,833	2,262	3,250
10	1,812	2,228	3,169

Abbildung 1.20: t-Verteilung (Ausschnitt)

Der t-Test verläuft analog zum F-Test in folgenden Schritten:

Empirischer t-Wert

1. Berechnung des empirischen t-Wertes
   Für den Regressionskoeffizienten b_1 hatten wir den Wert 18,881 ermittelt und für den Standardfehler des Regressionskoeffizienten gilt hier $s_{bj} = 9,187$. Aus (1.22) folgt damit

$$t_{emp} = \frac{18,881}{9,187} = 2,055$$

2. Vorgabe eines Signifikanzniveaus
   Wir wählen wiederum eine Vertrauenswahrscheinlichkeit von 95 Prozent bzw. $\alpha = 0,05$.

Vertrauens-
wahrscheinlichkeit

3. Auffinden des theoretischen t-Wertes
   Für die vorgegebene Vertrauenswahrscheinlichkeit von 95 Prozent und die Zahl der Freiheitsgrade (der nicht erklärten Streuung) $K - J - 1 = 10 - 1 - 1 = 8$ erhält man aus Abbildung 1.20 den theoretischen t-Wert $t_{tab} = 2,306$.

Theoretischer t-Wert

4. Vergleich des empirischen mit dem theoretischen t-Wert
   Da der t-Wert auch negativ werden kann (im Gegensatz zum F-Wert), ist dessen Absolutbetrag mit dem theoretischen t-Wert zu vergleichen (zweiseitiger Test).
   - Ist der Absolutbetrag des empirischen t-Wertes (t_{emp}) größer als der aus der Tabelle abgelesene theoretische t-Wert (t_{tab}), dann ist die Nullhypothese H_0 zu verwerfen. Es ist also zu folgern, dass β_j ungleich Null ist. Der Einfluss von X_j auf Y wird damit als statistisch signifikant erachtet.
   - Ist dagegen der Absolutbetrag des empirischen t-Wertes klein und übersteigt nicht den theoretischen Wert, so kann die Nullhypothese nicht verworfen werden. Der Einfluss von X_j ist damit nicht signifikant.

Zusammenfassend gilt für den t-Test:

$$|t_{emp}| > t_{tab} \rightarrow \quad H_0 \text{ wird verworfen} \rightarrow \text{Einfluss ist signifikant}$$
$$|t_{emp}| \leq t_{tab} \rightarrow \quad H_0 \text{ wird nicht verworfen}$$

Der Verwerfungsbereich des Tests entspricht in Abbildung 1.19 dem Bereich außerhalb der Quantile (unter den schraffierten Flächen) und der Annahmebereich entspricht dem Bereich zwischen den Quantilen.
Hier ergibt sich:

$2,005 < 2,306 \rightarrow H_0$ wird nicht verworfen!

Der Einfluss der unabhängigen Variablen (Zahl der Vertreterbesuche) erweist sich damit als nicht signifikant (die Alternativhypothese $H_1 : \beta_i \neq 0$ kann nicht bewiesen werden). Dieses Ergebnis wurde schon durch den F-Test vorweggenommen.

Testdurchführung mittels p-Wert

Die Testdurchführung lässt sich wiederum (wie schon beim obigen F-Test) vereinfachen, wenn man hier den p-Wert der t-Statistik verwendet, da man dann auf den theoretischen t-Wert aus der t-Tabelle verzichten kann. Der *p-Wert für den zweiseitigen t-Test* (mit df Freiheitsgraden) ist die Wahrscheinlichkeit

$$p = Pr(|t| > |t_{emp}|)$$

Da die t-Statistik auch negative Werte annehmen kann, sind beim zweiseitigen t-Test die Absolutwerte zu betrachten. Das Entscheidungskriterium für den t-Test lautet wie für den F-Test:

$$p < \alpha \qquad \rightarrow \qquad H_0 \text{ wird verworfen}$$

Im Beispiel erhält man für $t_{emp} = 2,005$ den p-Wert $p = 0,074$.[21] Bei der einfachen Regression ist der p-Wert von t_{emp} identisch mit dem von F_{emp}. Es folgt somit wiederum, dass H_0 nicht verworfen werden kann.

Sehr einfach ist bei Verwendung des p-Wertes die Durchführung eines *einseitigen t-Tests*. Dazu ist lediglich der erhaltene p-Wert durch 2 zu dividieren. Der Test wird damit schärfer. In unserem Beispiel ergäbe sich $p = 0,037$. Die Nullhypothese $H_0 : \beta_j \leq 0$ könnte damit verworfen werden, da jetzt $p < \alpha$ gilt, und die Alternativhypothese ($H_1 : \beta_j > 0$), dass die Vertreterbesuche eine positive Wirkung auf die Absatzmenge haben, wäre somit bewiesen. Der F-Test ermöglicht keine Durchführung eines einseitigen Tests.

F-Test und t-Test

Bei nur einer unabhängigen Variablen ist der F-Test für das Modell (die Gesamtheit der Variablen) auch ein Test der einen Variablen, deren Einfluss hier durch den t-Test geprüft wurde. Im Fall der einfachen Regression reicht es daher aus, nur einen dieser beiden Tests durchzuführen, und wir haben hier nur aus didaktischen Gründen beide Tests durchgeführt.

Einsatz von F- und t-Test

[21] In Excel erhält man den p-Wert für den zweiseitigen t-Test durch die Funktion TVERT(t; df; 2) oder ab Excel 2013 auch durch T.VERT.2S(x; df). Es gilt damit T.VERT.2S(2,005;8) = 0,074.

Während der t-Test nur für die Prüfung einer einzelnen Variablen geeignet ist, kann der F-Test für die Prüfung einer Mehrzahl von Variablen verwendet werden. Wir behandeln hier nur den F-Test für die Gesamtheit der Variablen. Mit Hilfe des F-Tests kann jedoch in einem multiplen Regressionsmodell der Einfluss einer Untermenge der erklärenden Variablen getestet werden, was sehr nützlich sein kann.[22] Damit ist es natürlich auch immer möglich, mit dem F-Test eine einzelne Variable zu prüfen und ihn an Stelle eines t-Tests zu verwenden. In diesem Fall hat die F-Statistik nur einen Freiheitsgrad im Zähler und es gilt:

$$F = t^2$$

Man kann dies durch Vergleich der ersten Spalte einer F-Tabelle mit der t-Tabelle überprüfen. F-Test und t-Test kommen folglich in diesem Fall immer zu gleichen Aussagen.

Während also der F-Test für die Prüfung einer Mehrzahl von Variablen verwendet werden kann, ist für die Prüfung einer einzelnen Variablen die Anwendung des t-Tests einfacher. Überdies ermöglicht der t-Test auch die Durchführung von einseitigen Tests. Zur Prüfung eines multiplen Regressionsmodells sollten daher beide Tests zur Anwendung kommen.

1.2.4.2 Konfidenzintervall des Regressionskoeffizienten

Durch den t-Test wurde die Frage überprüft, ob die unbekannten, wahren Regressionskoeffizienten β_j (j = 1, 2, ..., J) sich von Null unterscheiden. Hierfür wurde ein *Annahmebereich* für b_j bzw. die Transformation von b_j in einen t-Wert konstruiert. Eine andere Frage ist jetzt, welchen Wert die unbekannten, wahren Regressionskoeffizienten β_j mutmaßlich haben. Dazu ist ein *Konfidenzintervall* für β_j zu bilden.

Konfidenzintervall

Die beste Schätzung für den unbekannten Regressionskoeffizienten β_j liefert der geschätzte Regressionskoeffizient b_j. Als Konfidenzintervall ist daher ein Bereich um b_j zu wählen, in dem der unbekannte Wert β_j mit einer bestimmten Wahrscheinlichkeit liegen wird. Dazu ist wiederum die Vorgabe einer Vertrauenswahrscheinlichkeit erforderlich.

Für diese Vertrauenswahrscheinlichkeit und die Zahl der Freiheitsgrade der nicht erklärten Streuung (K-J-1) ist sodann der betreffende t-Wert zu bestimmen (aus der t-Tabelle für den zweiseitigen t-Test entnehmen).

Konfidenzintervall für den Regressionskoeffizienten

$$b_j - t \cdot s_{b_j} \leq \beta_j \leq b_j + t \cdot s_{b_j} \tag{1.23}$$

mit

β_j	=	Wahrer Regressionskoeffizient (unbekannt)
b_j	=	geschätzter Regressionskoeffizient
t	=	t-Wert aus der Student-Verteilung
s_{b_j}	=	Standardfehler des Regressionskoeffizienten

Die benötigten Werte sind identisch mit denen, die wir im t-Test verwendet haben. Für den Regressionskoeffizienten in unserem Beispiel erhält man damit das folgende

[22] Vgl. z. B. Kmenta (1997), S. 416 f.; Ramanathan (1998), S. 169 ff.

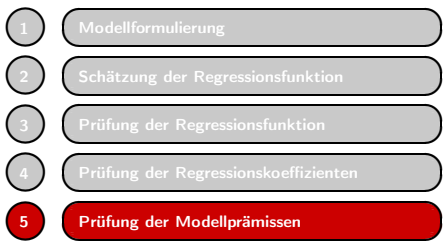
Konfidenzintervall:

$$18,881 - 2,306 \cdot 9,187 \leq \beta_1 \leq 18,881 + 2,306 \cdot 9,187$$
$$-2,304 \leq \beta_1 \leq 40,066$$

Interpretation

Das Ergebnis ist wie folgt zu interpretieren: Mit einer Vertrauenswahrscheinlichkeit von 0,95 liegt der wahre Regressionskoeffizient der Variablen BESUCHE zwischen den Werten -2,304 und 40,066. Je größer das Konfidenzintervall ist, desto unsicherer ist die Schätzung der Steigung der Regressionsgeraden in der Grundgesamtheit, m. a. W. desto unzuverlässiger ist die gefundene Regressionsfunktion bezüglich dieses Parameters. Dieses gilt insbesondere dann, wenn innerhalb des Konfidenzintervalls ein Vorzeichenwechsel liegt, die Richtung des vermuteten Einflusses sich also umkehren kann („Je größer die Zahl der Besuche, desto kleiner die abgesetzte Menge“).

1.2.5 Prüfung der Modellprämissen



Annahmen

Die Güte der Schätzung für die Regressionsparameter, die sich mittels der oben beschriebenen KQ-Methode erzielen lassen, sowie auch die Anwendbarkeit der Tests zur Überprüfung der Güte hängen von gewissen Annahmen ab, die wir bislang stillschweigend unterstellt hatten. Dabei spielt die oben eingeführte Störgröße eine zentrale Rolle.

Die Störgröße wurde eingeführt, um der bestehenden Unsicherheit bei der Modellierung empirischer Sachverhalte Rechnung zu tragen. Da sich die Variation einer empirischen Variablen Y nie vollständig durch eine begrenzte Menge von beobachtbaren Variablen erklären lässt, hatten wir in (1.17) ein stochastisches Modell formuliert, das der Regressionsanalyse zugrunde gelegt wird.

Für die Existenz der Störgröße sind insbesondere folgende Ursachen zu nennen:

- Unberücksichtigte Einflussgrößen
- Fehler in den Daten: Messfehler und Auswahlfehler.

Die Berücksichtigung aller möglichen Einflussgrößen von Y wäre mit einem unvertretbar großen Aufwand verbunden und würde das Modell unhandlich machen. Der Wert eines Modells resultiert daraus, dass es einfacher ist als die Realität und sich auf die Wiedergabe wichtiger struktureller Aspekte begrenzt.

Datenfehler

Fehler in den Daten sind insbesondere Messfehler, bedingt durch begrenzte Messgenauigkeit, und Auswahlfehler, die entstehen, wenn die Daten aufgrund einer Teilauswahl (Stichprobe) gewonnen werden. Ein zufälliger Auswahlfehler ist bei Stichproben unvermeidbar.

Denkt man bei der zu erklärenden Variablen Y an Absatzdaten (Absatzmengen, Marktanteile, Käuferreichweiten, Markenbekanntheit etc.), so handelt es sich dabei meist um Stichprobendaten, die überdies auch nie frei von Messfehlern sind. Als Einflussgrößen wirken neben den Maßnahmen des Anbieters auch die Maßnahmen der Konkurrenten und die des Handels. Hinzu können vielfältige gesamtwirtschaftliche, gesellschaftliche oder sonstige Umwelteinflüsse kommen. Und schließlich resultieren

die einzelnen Käufe aus den Entscheidungen von Menschen, in deren Verhalten immer ein gewisses Maß an Zufälligkeit enthalten ist.

Es ist daher gerechtfertigt, die Störgröße als eine Zufallsgröße aufzufassen und der Regressionsanalyse ein stochastisches Modell zugrunde zu legen. Die beobachteten Daten lassen sich als Realisationen eines Prozesses auffassen, der durch dieses Modell generiert wird. Die Menge der Beobachtungen bildet damit eine *Stichprobe der möglichen Realisationen*.

Bei der Durchführung einer Regressionsanalyse wird eine Reihe von Annahmen gemacht, die das zugrunde gelegte stochastische Modell betreffen. Nachfolgend wollen wir auf die Bedeutung dieser Annahmen und die Konsequenzen ihrer Verletzung eingehen. Da wir uns hier auf die lineare Regressionsanalyse beschränken (mit der sich sehr wohl auch nichtlineare Probleme behandeln lassen), sprechen wir im Folgenden vom klassischen oder *linearen Modell der Regressionsanalyse*.

Lineares Modell

Annahmen des linearen Regressionsmodells

A1. $y_k = \beta_0 + \sum_{j=1}^{J} \beta_j \cdot x_{jk} + u_k \qquad$ mit $k = 1, 2, \ldots, K$ und $K > J + 1$

Das Modell ist *richtig spezifiziert*, d. h.

- es ist linear in den Parametern β_0 und β_j,
- es enthält die relevanten erklärenden Variablen,
- die Zahl der zu schätzenden Parameter (J+1) ist kleiner als die Zahl der vorliegenden Beobachtungen (K).

A2. $\text{Erw}(u_k) = 0$
Die Störgrößen haben den Erwartungswert Null.

A3. $\text{Cov}(u_k, x_{jk}) = 0$
Es besteht keine Korrelation zwischen den erklärenden Variablen und der Störgröße.

A4. $\text{Var}(u_k) = \sigma^2$
Die Störgrößen haben eine konstante Varianz σ^2 (*Homoskedastizität*).

A5. $\text{Cov}(u_k, u_{k+r}) = 0 \qquad$ mit $r \neq 0$
Die Störgrößen sind unkorreliert (*keine Autokorrelation*).

A6. Zwischen den erklärenden Variablen X_j besteht keine lineare Abhängigkeit (*keine perfekte Multikollinearität*).

A7. Die Störgrößen u_k sind *normalverteilt*.

Unter den Annahmen 1 bis 6 liefert die KQ-Methode *lineare Schätzfunktionen* für die Regressionsparameter, die alle wünschenswerten Eigenschaften von Schätzern besitzen, d. h. sie sind *unverzerrt* (erwartungstreu) und *effizient*.[23] Effizienz bedeutet hier, dass sie unter allen linearen und unverzerrten Schätzern eine kleinstmögliche Varianz aufweisen. Im Englischen werden diese Eigenschaften als *BLUE* bezeichnet (Best Linear Unbiased Estimators), wobei mit „Best“ die Effizienz gemeint ist.

BLUE

[23] Dies ist das sog. Gauß-Markow-Theorem. Vgl. dazu Fahrmeir/Kneib/Lang (2009), S. 104; Bleymüller/Weißbach (2015), S. 176; Kmenta (1997), S. 162.

Zur Durchführung von *Signifikanztests* ist außerdem Annahme 7 von Vorteil. Diese Annahme ist auch nicht unplausibel. Da die Störgröße, wie oben dargestellt, die gemeinsame Wirkung sehr vieler und im Einzelnen relativ unbedeutender Einflussfaktoren repräsentiert, die voneinander weitgehend unabhängig sind, lässt sich die Annahme der Normalverteilung durch den „zentralen Grenzwertsatz“ der Statistik stützen.[24]

1.2.5.1 Nichtlinearität

Nichtlinearität kann in vielen verschiedenen Formen auftreten. Abbildung 1.22 zeigt eine lineare und eine nichtlineare Beziehung. Das lineare Regressionsmodell fordert lediglich, dass die Beziehung linear in den Parametern ist. In vielen Fällen ist es daher möglich, eine nichtlineare Beziehung durch Transformation der Variablen in eine lineare Beziehung zu überführen.

Nichtlinearität in den Parametern bzw. Variablen

Derartige nichtlineare Beziehungen zwischen der abhängigen und einer unabhängigen Variablen können durch Wachstums- oder Sättigungsphänomene bedingt sein (z. B. abnehmende Ertragszuwächse der Werbeausgaben). Sie lassen sich oft leicht durch Betrachten des Punktediagramms entdecken. Die Folge von nicht entdeckter Nichtlinearität ist eine Verzerrung der Schätzwerte der Parameter, d. h. die Schätzwerte b_j streben mit wachsendem Stichprobenumfang nicht mehr gegen die wahren Werte β_j.

Generell lässt sich eine Variable X durch eine Variable $X' = f(X)$ ersetzen, wobei f eine beliebige nichtlineare Funktion bezeichnet. Folglich ist das Modell

Linearisierung

$$Y = \beta_0 + \beta_1 X' + u \qquad \text{mit } X' = f(X) \tag{1.24}$$

linear in den Parametern β_0 und β_1 und in X', nicht aber in X. Durch Transformation von X in X' wird die Beziehung linearisiert und lässt sich mittels Regressionsanalyse schätzen.

In allgemeinerer Form lässt sich das lineare Regressionsmodell unter Berücksichtigung nichtlinearer Transformationen der Variablen auch in folgender Form schreiben:

$$f(Y) = \beta_0 + \sum_{j=1}^{J} \beta_j \cdot f_j(X_j) + u \tag{1.25}$$

Abbildung 1.21 zeigt Beispiele für anwendbare nichtlineare Transformationen. Dabei ist jeweils der zulässige Wertebereich angegeben. Der Exponent c in der Potenzfunktion 9 muss vorgegeben werden.

Ein spezielles nichtlineares Modell bildet das *multiplikative Modell* der Form

$$Y = \beta_0 \cdot X_1^{\beta_1} \cdot X_2^{\beta_2} \cdot \ldots \cdot X_J^{\beta_J} \cdot u \tag{1.26}$$

[24] Der zentrale Grenzwertsatz der Statistik besagt, dass die Summenvariable (oder der Mittelwert) von N unabhängigen und identisch verteilten Zufallsvariablen normalverteilt ist und zwar unabhängig von der Verteilung der Zufallsvariablen, wenn N hinreichend groß ist. In der Realität finden sich viele Zufallserscheinungen, die sich aus der Überlagerung zahlreicher zufälliger Effekte ergeben. Der zentrale Grenzwertsatz liefert die Rechtfertigung dafür, in diesen Fällen anzunehmen, dass zumindest angenähert eine Normalverteilung gegeben ist.

Nr.	Bezeichnung	Definition	Bereich
1	Logartihmus	$ln(X)$	$X > 0$
2	Exponential	$exp(X)$	
3	Arkussinus	$sin^{-1}(X)$	$\|X\| \leq 1$
4	Arkustangens	$tan^{-1}(X)$	
5	Logit	$ln(X/(1-X))$	$0 < X < 1$
6	Reziprok	$1/X$	$X \neq 0$
7	Quadrat	X^2	
8	Wurzel	$X^{1/2}$	$X \geq 0$
9	Potenz	X^c	$X > 0$

Abbildung 1.21: Nichtlineare Transformationen

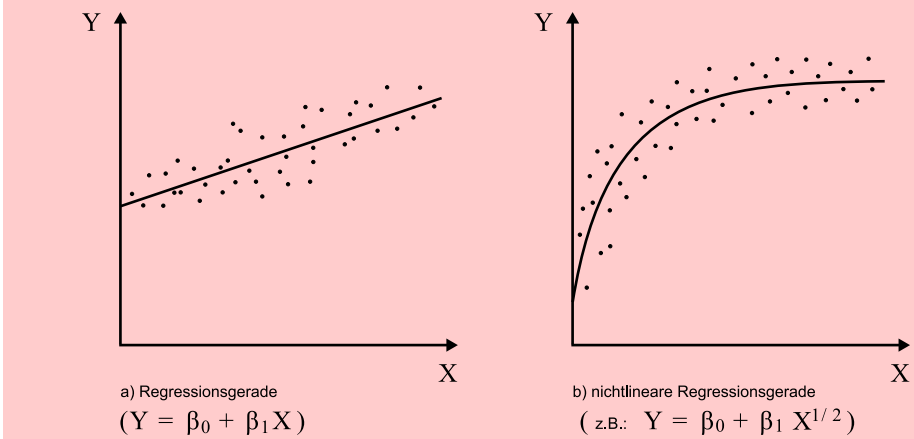


Abbildung 1.22: Lineare und nichtlineare Regressionsbeziehungen

Cobb/Douglas-Produktionsfunktion

Ein Beispiel für ein derartiges Modell ist die bekannte Cobb/Douglas-Produktionsfunktion. Die Exponenten β_j lassen sich als Elastizitäten interpretieren (d. h. die Beziehung zwischen Y und den Variablen X_j ist durch konstante Elastizitäten gekennzeichnet).

Durch Logarithmieren aller Variablen lässt sich das multiplikative Modell in ein lineares Modell überführen und damit mittels Regressionsanalyse schätzen. Man erhält

$$lnY = \beta_0' + \beta_1 \cdot lnX_1 + \beta_2 \cdot lnX_2 + \ldots + \beta_J \cdot lnX_J + u' \tag{1.27}$$

mit $\beta_0' = ln\ \beta_0$ und $u' = ln\ u$

Interaktionseffekte

Eine weitere Form von Nichtlinearität kann im Mehr-Variablen-Fall dadurch auftreten, dass sich die Wirkungen von unabhängigen Variablen nicht-additiv verknüpfen. So kann z. B. eine Preisänderung in Verbindung mit einer Verkaufsförderungsaktion anders wirken als ohne diese. Derartige *Interaktionseffekte* lassen sich wie folgt berücksichtigen:

$$Y = \beta_0 + \beta_1 V + \beta_2 P + \beta_3 V \cdot P + u \tag{1.28}$$

Dabei bezeichnet V die Verkaufsaktion und P den Preis. Das Produkt $V \cdot P$ wird als *Interaktionsterm* bezeichnet, dessen Wirkung der Koeffizient β_3 reflektiert.

Für die Aufdeckung von Nichtlinearität sind statistische *Testmöglichkeiten* vorhanden, auf die hier nur verwiesen werden kann.[25] Hinweise auf das Vorliegen von Nichtlinearität können im übrigen auch die nachfolgend beschriebenen Tests auf Autokorrelation und Heteroskedastizität geben.

1.2.5.2 Erwartungswert der Störgröße ungleich Null

Wenn im Regressionsmodell alle systematischen Einflussgrößen von Y explizit berücksichtigt werden, dann umfasst die Störvariable u nur zufällige Effekte, die positive und negative Abweichungen zwischen beobachteten und geschätzten Werten verursachen. Das Regressionsmodell unterstellt (Annahme 2), dass der Erwartungswert der Störvariable Null ist und sich die Schwankungen somit im Mittel ausgleichen.

Eine Verletzung dieser Annahme ergibt sich z. B., wenn die Werte von Y mit einem konstanten Fehler zu hoch oder zu niedrig gemessen werden. Wir sprechen dann von einem systematischen Messfehler und die Störgröße enthält einen systematischen Effekt. Was ist die Folge? Durch die KQ-Schätzung der Regressionsparameter wird quasi erzwungen, dass der Mittelwert der Residuen Null wird (vgl. Gleichung A5 im mathematischen Anhang). Der systematische Messfehler geht dabei in den Schätzwert des konstanten Gliedes b_0 ein, sodass dieser nicht mehr unverzerrt ist. Werden die Werte von Y konstant überhöht gemessen, so wird auch b_0 zu groß ausfallen. In den meisten Anwendungen ist der Wert von b_0 nur von sekundärem oder gar keinem Interesse und eine Verzerrung wird daher wenig stören.

Systematischer Messfehler

Es ist aber große Vorsicht geboten, wenn man ein Modell ohne konstantes Glied spezifiziert, da sich dann die Verzerrung auf die Regressionskoeffizienten auswirkt. Dies sollte daher nur in wohlbegründeten Ausnahmefällen geschehen.

1.2.5.3 Falsche Auswahl der Regressoren

Das korrekt spezifizierte Regressionsmodell sollte gemäß Annahme A1 alle relevanten Einflussgrößen von Y enthalten. Dies wird sich jedoch oft nicht realisieren lassen, sei es, dass die Erfassung technisch nicht möglich oder zu aufwändig wäre, oder sei es, dass gar nicht alle relevanten Einflussgrößen bekannt sind. Die Modellformulierung bleibt dann unvollständig, d. h. es fehlen erklärende Variablen, und eine mögliche Folge ist die Verzerrung der Schätzwerte.

Glücklicherweise muss dies nicht zwangsläufig die Folge sein, wenn Annahme A3 erfüllt ist, d. h. wenn keine Korrelation zwischen den im Modell berücksichtigten erklärenden Variablen und der Störgröße (die die unberücksichtigten Variablen enthält) besteht. Die Folge ist vielmehr die gleiche wie die eines konstanten Messfehlers. Der Erwartungswert der Störgröße ist nicht mehr Null und es kommt zu einer Verzerrung von b_0.

Anders verhält es sich dagegen, wenn $\mathrm{Cov}(x_{jk}, u_k) > 0$ gilt, also eine positive Korrelation zwischen der Variablen j und der Störgröße besteht. In diesem Fall würde der Schätzwert für b_j zu groß ausfallen. Durch die KQ-Schätzung würde nämlich der Teil der Variation von Y, der von u kommt, fälschlich der Variable X_j zugeordnet werden.

Vernachlässigung relevanter Variablen

[25] Vgl. z. B. Kmenta (1997), S. 517 ff.; v. Auer (2016), S. 299 ff.

Beispiel: Das korrekte Modell lautet:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + v \tag{1.29}$$

und wir spezifizieren fälschlich

$$Y = \beta_0 + \beta_1 X_1 + u \qquad \text{mit} \quad u = \beta_2 X_2 + v \tag{1.30}$$

Wenn X_1 und X_2 korreliert sind, dann sind auch X_1 und u korreliert und es liegt damit eine Verletzung von Annahme A3 vor, die zu einer Verzerrung von b_1 führt.[26] Der Einfluss der vernachlässigten Variablen X_2 schlägt sich in b_1 nieder, und zwar um so mehr, je stärker X_1 und X_2 korreliert sind. Ist dagegen die vernachlässigte Variable X_2 nicht mit X_1 korreliert, so tritt dieser Effekt nicht auf. Es wäre lediglich eine Verzerrung von b_0 möglich. Eine Ausnahme besteht wiederum bei einem Modell ohne konstanten Term: in diesem Fall ist auch eine Verzerrung von b_1 möglich.

Overfitting

Neben der Vernachlässigung relevanter Variablen (underfitting) kann es auch vorkommen, dass ein Modell zu viele erklärende Variable enthält (overfitting). Auch dies kann, wie die Vernachlässigung relevanter Variablen, eine Folge unvollständigen theoretischen Wissens und daraus resultierender Unsicherheit sein. Der Untersucher packt dann aus Sorge davor, relevante Variable zu übersehen, alle verfügbaren Variablen in das Modell, ohne sie einer sachlogischen Prüfung zu unterziehen. Solche Modelle werden auch als „kitchen sink models“ bezeichnet. Diese Vorgehensweise führt zwar nicht zu verzerrten Schätzern für die Regressionskoeffizienten, wohl aber zu ineffizienten Schätzern (d. h. die Varianz der Schätzer ist nicht mehr minimal).[27] Wie in vielen Dingen gilt auch hier: Mehr ist nicht besser.

Je größer die Anzahl von Variablen in der Regressionsgleichung ist, desto eher kann es vorkommen, dass ein tatsächlicher Einflussfaktor nicht signifikant erscheint, weil seine Wirkung nicht mehr hinreichend präzise ermittelt werden kann. Umgekehrt wächst mit steigender Zahl der Regressoren auch die Gefahr, dass eine irrelevante Variable irrtümlich als statistisch signifikant erscheint, obgleich sie nur zufällig mit der abhängigen Variablen korreliert. Es ist also sowohl möglich, dass sich eine irrelevante Variable als statistisch signifikant erweist, als auch, dass ein relevanter Einflussfaktor nicht signifikant erscheint. Letzteres sollte daher auch nicht dazu führen, eine sachlich begründete Hypothese zu verwerfen, solange man kein widersprüchliches Ergebnis erzielt hat. Das wäre z. B. der Fall, wenn ein signifikanter Koeffizient ein anderes Vorzeichen hat, als angenommen. In diesem Fall sollte man seine Hypothese verwerfen oder zumindest überdenken. Dies zeigt die Wichtigkeit theoretischer oder sachlogischer Überlegungen bei der Analyse kausaler Zusammenhänge.[28]

1.2.5.4 Heteroskedastizität

Wenn die Streuung der Residuen in einer Reihe von Werten der prognostizierten abhängigen Variablen nicht konstant ist, dann deutet dies auf Heteroskedastizität hin. Damit ist eine Prämisse des linearen Regressionsmodells verletzt, die verlangt, dass die Varianz der Fehlervariablen u für alle k homogen ist, m. a. W. die Störgröße darf nicht von den unabhängigen Variablen und von der Reihenfolge der Beobachtungen

[26] Eine Alternative zur KQ-Schätzung liefert in diesem Fall die sog. Instrument-Variablen-Schätzung (IV-Schätzung). Siehe hierzu v. Auer (2016), S. 522 ff.

[27] Vgl. z. B. Kmenta (1997), S. 446 ff.

[28] Zu Verfahren, die die richtige Auswahl der Regressoren unterstützen können, vgl. z. B. v. Auer (2016), S. 327 ff. Ein solcher Test ist z. B. der RESET-Test (REgression Specification Error Test) von Ramsay (1969). Vgl. dazu auch Ramanathan (1998), S. 294 ff.

abhängig sein. Ein Beispiel für das Auftreten von Heteroskedastizität wäre eine zunehmende Störgröße in einer Reihe von Beobachtungen etwa aufgrund von Messfehlern, die durch nachlassende Aufmerksamkeit der beobachtenden Person entstehen.

Aufdecken von Heteroskedastizität

Heteroskedastizität führt zu Ineffizienz der Schätzung und verfälscht den Standardfehler des Regressionskoeffizienten. Damit wird auch die Schätzung des Konfidenzintervalls ungenau.

Zur Aufdeckung von Heteroskedastizität empfiehlt sich zunächst eine visuelle Inspektion der Residuen, indem man diese gegen die prognostizierten (geschätzten) Werte von Y plottet. Dabei ergibt sich bei Vorliegen von Heteroskedastizität meist ein Dreiecksmuster, wie in Abbildung 1.23 a oder b dargestellt.

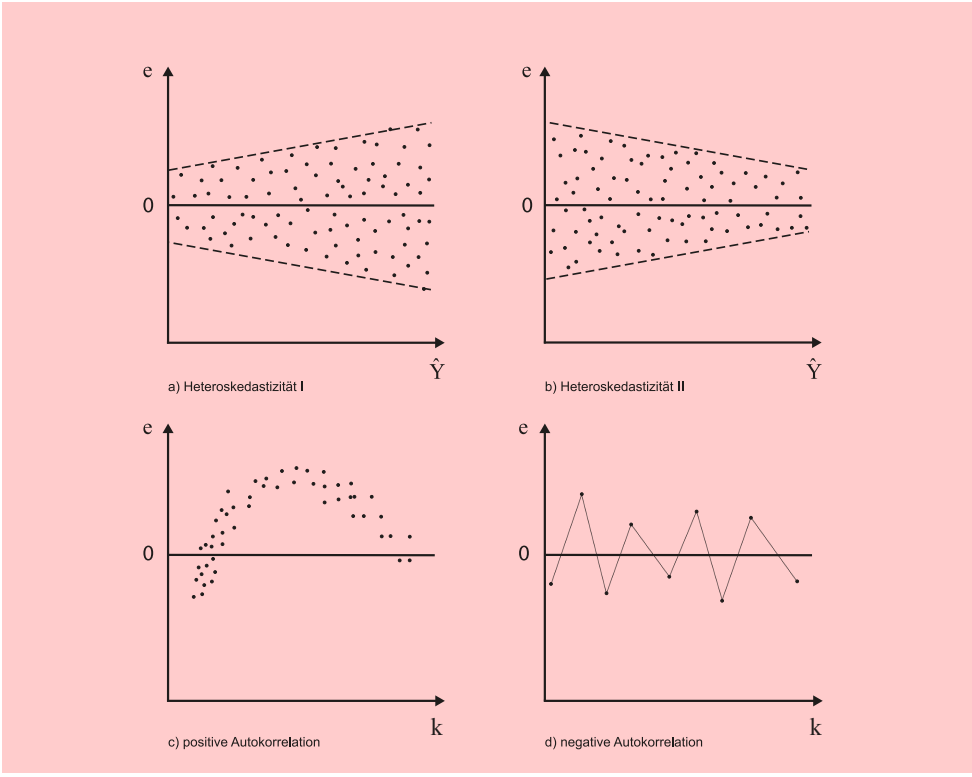


Abbildung 1.23: Heteroskedastizität und Autokorrelation

Goldfeld/Quandt-Test

Der bekannteste Test zur Aufdeckung von Heteroskedastizität bildet der *Goldfeld/Quandt-Test*, bei dem die Stichprobenvarianzen der Residuen in zwei Unterstichproben, z. B. der ersten und zweiten Hälfte einer Zeitreihe, verglichen und ins Verhältnis gesetzt werden.[29] Liegt perfekte Homoskedastizität vor, müssen die Varianzen identisch sein ($s_1^2 = s_2^2$), d. h. das Verhältnis der beiden Varianzen der Teilgruppen entspricht dem Wert Eins. Je weiter das Verhältnis von Eins abweicht, desto unsicherer wird die Annahme gleicher Varianz. Wenn die Störgrößen normalverteilt sind und die Annahme der Homoskedastizität zutrifft, folgt das Verhältnis der Varianzen einer F-Verteilung und kann daher als Teststatistik gegen die Nullhypothese

[29] Zu dieser und anderen Testmöglichkeiten auf Heteroskedastizität vgl. z. B. Kmenta (1997), S. 292 ff.; Greene (2018), S. 304 ff.

gleicher Varianz H_0: $\sigma_1^2 = \sigma_2^2$ getestet werden. Die F-Teststatistik berechnet sich wie folgt:

$$F_{emp} = \frac{s_1^2}{s_2^2} \quad \text{mit} \quad s_1^2 = \frac{\sum_{k=1}^{K_1} e_k^2}{K_1 - J - 1} \quad \text{und} \quad s_2^2 = \frac{\sum_{k=1}^{K_2} e_k^2}{K_2 - J - 1} \tag{1.31}$$

Dabei sind K_1 und K_2 die Fallzahlen in den beiden Teilgruppen und J bezeichnet die Anzahl der unabhängigen Variablen in der Regression. Die Gruppen sind dabei so anzuordnen, dass $s_1^2 \geq s_2^2$ gilt. Der ermittelte F-Wert ist bei vorgegebenem Signifikanzniveau gegen den theoretischen F-Wert für $(K_1 - J - 1, K_2 - J - 1)$ Freiheitsgrade zu testen.

Verfahren von Glesjer

Eine andere Methode zur Aufdeckung von Heteroskedastizität bietet ein *Verfahren von Glesjer*, bei dem eine Regression der absoluten Residuen auf die Regressoren durchgeführt wird:[30]

$$|e_k| = \beta_0 + \sum_{j=1}^{J} \beta_j x_{jk} \tag{1.32}$$

Bei Homoskedastizität gilt die Nullhypothese H_0: $\beta_j = 0$ $(j = 1, 2, \ldots, J)$. Wenn sich signifikant von Null abweichende Koeffizienten ergeben, so muss die Annahme der Homoskedastizität abgelehnt werden.

Test auf Nichtlinearität

Zur Begegnung von Heteroskedastizität kann versucht werden, durch Transformation der abhängigen Variablen oder der gesamten Regressionsbeziehung Homoskedastizität der Störgrößen herzustellen.[31] Dies impliziert meist eine nichtlineare Transformation. Somit ist Heteroskedastizität meist auch ein Problem von Nichtlinearität und der Test auf Heteroskedastizität kann auch als ein Test auf Nichtlinearität aufgefasst werden. Ähnliches gilt auch für das nachfolgend behandelte Problem der Autokorrelation.[32]

1.2.5.5 Autokorrelation

Das lineare Regressionsmodell basiert auf der Annahme, dass die Störgrößen in der Grundgesamtheit unkorreliert sind. Wenn diese Bedingung nicht gegeben ist, sprechen wir von Autokorrelation. Autokorrelation tritt vor allem bei Zeitreihen auf. Die Abweichungen von der Regressions(=Trend)geraden sind dann nicht mehr zufällig, sondern in ihrer Richtung von den Abweichungen, z. B. des vorangegangenen Beobachtungswertes, abhängig.

Autokorrelation führt zu Verzerrungen bei der Ermittlung des Standardfehlers der Regressionskoeffizienten und demzufolge auch bei der Bestimmung der Konfidenzintervalle für die Regressionskoeffizienten.

[30] Vgl. Maddala/Lahiri (2009), S. 214 f. Ein anderer gebräuchlicher Test ist der White-Test von White (1980), der in einigen ökonomischen Computer-Programmen angeboten wird. Vgl. dazu z. B. Kmenta (1997), S. 295 ff.; Greene (2018), S. 304 f.; v. Auer (2016), S. 435.

[31] Vgl. Kockläuner (1988), S. 88 ff.

[32] Zur Erzielung konsistenter (asymptotisch erwartungstreuer) Schätzer bei Vorliegen von Heteroskedastizität werden anstelle der einfachen KQ-Methode, auch Ordinary Least Squares (OLS) genannt, erweiterte Verfahren wie Generalized Least Squares (GLS) oder Weighted Least Squares (WLS) verwendet. Vgl. hierzu Greene (2018), S. 306 ff.; Kmenta (1997), S. 352 ff.; Ramanathan (1998), S. 392 ff.

Zur Aufdeckung von Autokorrelation empfiehlt sich auch hier zunächst eine visuelle Inspektion der Residuen, indem man diese gegen die prognostizierten (geschätzten) Werte von Y plottet. Bei positiver Autokorrelation liegen aufeinander folgende Werte der Residuen nahe beieinander (vgl. Abbildung 1.23 c), bei negativer Autokorrelation dagegen schwanken sie stark (vgl. Abbildung 1.23 d). Visuelle Inspektion

Die rechnerische Methode, eine Reihe von Beobachtungswerten auf Autokorrelation zu prüfen, stellt der *Durbin/Watson-Test* dar. Bei diesem Test wird die Reihenfolge der Residuen der Beobachtungswerte zum Gegenstand der Analyse gemacht. Der Durbin/Watson-Test prüft die Hypothese H_0, dass die Beobachtungswerte nicht autokorreliert sind.[33] Um diese Hypothese zu testen, wird ein empirischer Wert d ermittelt, der die Differenzen zwischen den Residuen von aufeinander folgenden Beobachtungswerten aggregiert.

Durbin/Watson-Statistik

$$d = \frac{\sum_{k=2}^{K} (e_k - e_{k-1})^2}{\sum_{k=1}^{K} e_k^2} \tag{1.33}$$

Zum besseren Verständnis der Formel lässt diese sich wie folgt schreiben:

$$d = \frac{\sum_{k=2}^{K} e_k^2}{\sum_{k=1}^{K} e_k^2} + \frac{\sum_{k=2}^{K} e_{k-1}^2}{\sum_{k=1}^{K} e_k^2} - 2\frac{\sum_{k=2}^{K} e_k \cdot e_{k-1}}{\sum_{k=1}^{K} e_k^2} \approx 1 + 1 - 2\frac{\sum_{k=2}^{K} e_k \cdot e_{k-1}}{\sum_{k=1}^{K} e_k^2}$$

Der rechte Teil der Gleichung gilt für großen Stichprobenumfang K. Entscheidend ist hier die Summe im Zähler des rechten Summanden, die positiv oder negativ werden kann.

Wenn die aufeinander folgenden Residuen nahe bei einander liegen, wie in Abbildung 1.23 c, dann werden jeweils positive Residuen oder negative Residuen miteinander multipliziert. Damit wird die Summe im Zähler des rechten Summanden positiv und der Quotient geht gegen 1 und damit d gegen 0. Niedrige Werte von d deuten daher auf eine positive Autokorrelation hin.

Wechseln dagegen aufeinander folgende Residuen die Vorzeichen, wie in Abbildung 1.23 d, dann geht der rechte Quotient gegen -1 und damit d gegen 4. Hohe Werte von d deuten daher auf eine negative Autokorrelation.

Wechseln die Vorzeichen aufeinander folgender Residuen zufällig, so geht der rechte Quotient gegen 0 und damit d gegen 2. Werte von d nahe 2 bedeuten daher, dass keine Autokorrelation vorliegt.

Abbildung 1.24 veranschaulicht diesen Sachverhalt. Sie zeigt den Annahmebereich und die Ablehnungsbereiche für die Nullhypothese, dass keine Autokorrelation vorliegt. Gleichzeitig deutet sie an, dass bei der Durchführung eines Durbin/Watson-Tests zwei Unschärfebereiche bestehen.

[33] Strenggenommen wird nur die Hypothese geprüft, dass keine lineare Autokorrelation erster Ordnung (zwischen e_k und e_{k-1}) vorliegt. Selbst wenn also die Nullhypothese nicht verworfen wird, heisst das nicht, dass keine nichtlineare Autokorrelation oder dass keine lineare Autokorrelation r-ter Ordnung (also zwischen e_k und e_{k-r}) vorliegt.

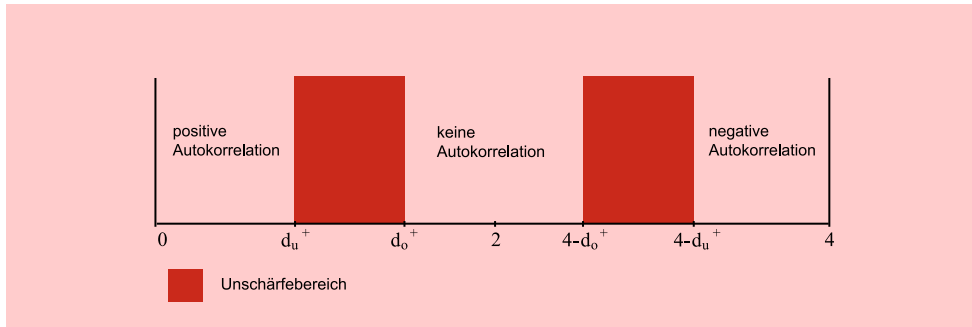


Abbildung 1.24: Ablehnungs- und Unschärfebereich

Fragestellung: Test zum Niveau α von:		
H_0 : keine Autokorrelation	gegen	H_1 : Autokorrelation gegeben
Teststatistik		**Entscheidung**
$d_o^+;\alpha/2 \le d \le 4-d_o^+;\alpha/2$		H_0
$d \le d_u^+;\alpha/2$ oder $d \ge 4-d_u^+;\alpha/2$		H_1
Unschärfebereich		Keine Entscheidung möglich

d = empirischer Wert
$d_u^+;\alpha/2$ = unterer Grenzwert aus der Tabelle zum Niveau $\alpha/2$
$d_o^+;\alpha/2$ = oberer Grenzwert aus der Tabelle zum Niveau $\alpha/2$

Abbildung 1.25: Entscheidungsregeln für den Durbin/Watson-Test

Die unteren und oberen Grenzwerte d_u^+ und für d_o^+ für die Unschärfebereiche lassen sich für unterschiedliche Anzahl von Beobachtungen und Regressoren den Durbin-Watson-Tabellen im Anhang entnehmen.[34] Abbildung 1.25 zeigt die Entscheidungsregeln für die Durchführung des Durbin-Watson-Tests.[35]

1.2.5.6 Multikollinearität und Schätzgenauigkeit

Das lineare Regressionsmodell basiert auf der Prämisse, dass die Regressoren nicht exakt linear abhängig sind, d. h. ein Regressor darf sich nicht als lineare Funktion der übrigen Regressoren darstellen lassen. In diesem Falle würde perfekte Multikollinearität bestehen und die Regressionsanalyse wäre rechnerisch nicht durchführbar.[36] Perfekte Multikollinearität wird selten vorkommen, und wenn, dann meist als Folge von Fehlspezifikationen, z. B. wenn man dieselbe Einflussgröße zweimal als unabhängige Variable in das Regressionsmodell aufnimmt. Die zweite Variable enthält dann keine zusätzliche Information und ist überflüssig.

[34] Testtabellen, die bereits bei sechs Beobachtungswerten beginnen, finden sich bei Savin/White (1977), S. 1989-1996.

[35] Die Durbin/Watson-Tabelle ist indifferent gegenüber der Frage, ob es sich um einen einseitigen oder zweiseitigen Test handelt. Im Falle des zweiseitigen Tests mit der Irrtumswahrscheinlichkeit α sind die Grenzwerte aus der Tabelle mit der Vertrauenswahrscheinlichkeit $1-\alpha/2$ zu bestimmen. Das wäre für $\alpha = 5\%$ die Tabelle A.9 im Anhang.

[36] Vgl. hierzu Formel (B6) im mathematischen Anhang zur Schätzung der Regressionskoeffizienten. Die Matrix $\mathbf{X'X}$ wird dann singulär und die Inverse existiert nicht.

Bei empirischen Daten besteht aber immer ein gewisser Grad an Multikollinearität, der nicht störend sein muss. Auch bei Vorliegen von Multikollinearität liefert die KQ-Methode Schätzer, die wir oben als BLUE bezeichnet haben. Ein hoher Grad an Multikollinearität aber wird zum Problem, denn mit zunehmender Multikollinearität werden die Schätzungen der Regressionsparameter unzuverlässiger. Dies macht sich bemerkbar am Standardfehler der Regressionskoeffizienten, der größer wird.

Multikollinearität

Bei Multikollinearität überschneiden sich die Streuungen der unabhängigen Variablen. Dies bedeutet zum einen Redundanz in den Daten und damit weniger Information. Zum anderen bedeutet es, dass sich die vorhandene Information nicht mehr eindeutig den Variablen zuordnen lässt. Dies kann grafisch mit Hilfe eines Venn-Diagramms veranschaulicht werden.[37] Abbildung 1.26 zeigt dies schematisch für eine Zweifachregression, wobei die Streuungen der abhängigen Variablen Y und der beiden Regressoren jeweils durch Kreise dargestellt sind.[38] Die Multikollinearität kommt in den Überschneidungsflächen C und D zum Ausdruck. Für die Schätzung von b_1 kann nur die Information in Fläche A genutzt werden und für die von b_2 die Information in Fläche B. Die Information in Fläche C dagegen kann den Regressoren nicht individuell zugeordnet werden und deshalb auch nicht für die Schätzung ihrer Koeffizienten genutzt werden. Sie ist deshalb aber nicht völlig verloren, denn sie vermindert den Standardfehler der Regression und erhöht damit das Bestimmtheitsmaß und die Genauigkeit von Prognosen.

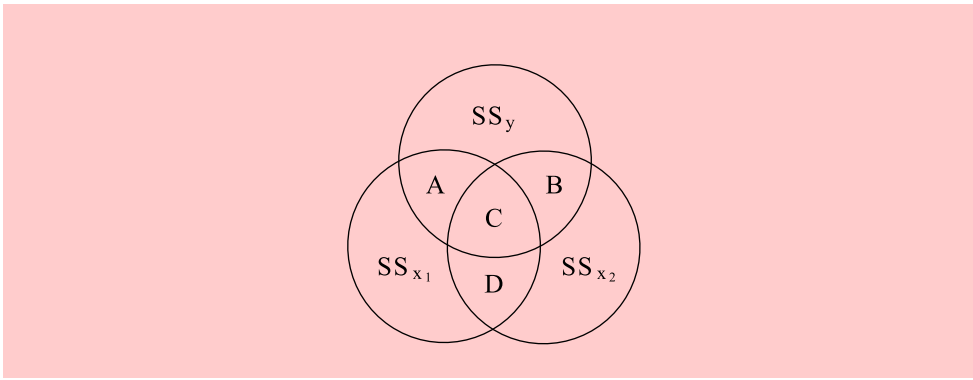


Abbildung 1.26: Venn-Diagramm

Es kann daher infolge von Multikollinearität vorkommen, dass das Bestimmtheitsmaß R^2 der Regressionsfunktion signifikant ist, obgleich keiner der Koeffizienten in der Funktion signifikant ist. Eine andere Folge von Multikollinearität kann darin bestehen, dass sich die Regressionskoeffizienten erheblich verändern, wenn eine weitere Variable in die Funktion einbezogen oder eine enthaltene Variable aus ihr entfernt wird.

Um dem Problem der Multikollinearität zu begegnen, ist zunächst deren Aufdeckung erforderlich, d. h. es muss festgestellt werden, welche Variablen betroffen sind und wie stark das Ausmaß der Multikollinearität ist. Einen ersten Anhaltspunkt kann die Betrachtung der *Korrelationsmatrix* liefern. Hohe Korrelationskoeffizienten ($|r|$ nahe 1) zwischen den unabhängigen Variablen bedeuten ernsthafte Multikollinearität. Die Korrelationskoeffizienten messen allerdings nur *paarweise* Abhängigkeiten. Es

Prüfung auf Multikollinearität

[37] Vgl. hierzu v. Auer (2016), S. 565.
[38] Für die Streuungen (Sum of Squares) gilt $SS_Y = \sum(y_k - \overline{y})^2$ und $SS_{X_j} = \sum(x_{jk} - \overline{x}_j)^2$.

kann deshalb auch hochgradige Multikollinearität trotz durchgängig niedriger Werte für die Korrelationskoeffizienten der unabhängigen Variablen bestehen.

Zur Aufdeckung von Multikollinearität empfiehlt es sich daher, eine Regression jeder unabhängigen Variablen X_j auf die übrigen unabhängigen Variablen durchzuführen und so deren lineare Abhängigkeit zu ermitteln. Ein Maß hierfür bildet das entsprechende Bestimmtheitsmaß R_j^2 (oder der multiple Korrelationskoeffizient R_j). Ein Wert $R_j^2 = 1$ besagt, dass sich die Variable X_j durch eine Linearkombination der anderen unabhängigen Variablen erzeugen lässt und folglich überflüssig ist. Für Werte von R_j^2 nahe 1 gilt das gleiche in abgeschwächter Form. R_j^2 lässt sich damit als *Redundanz* der Variablen X_j interpretieren.

In Statistik-Programmen wie SPSS werden hiermit verwandte Maße zur Prüfung auf Multikollinearität ausgegeben, die sog. *Toleranz* und der *Variance Inflation Factor*.[39]

Toleranz

Toleranz der Variablen X_j

$$T_j = 1 - R_j^2 \tag{1.34}$$

mit

R_j^2 = Bestimmtheitsmaß für Regression der unabhängigen Variablen X_j auf die übrigen Variablen in der Regressionsfunktion

X_j = $f(X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_J)$

VIF

Variance Inflation Factor von Variable X_j

$$VIF_j = \frac{1}{1 - R_j^2} \tag{1.35}$$

Toleranzwerte nahe Null deuten auf starke Multikollinearität der betreffenden Variable mit den übrigen Regressoren hin. Ihre Aufnahme in das Modell ist dann nicht mehr „tolerierbar“. Exakte Grenzen lassen sich allerdings nicht angeben.[40] Für $T_j = 0,2$ erhält man $VIF_j = 5$, oder für $T_j = 0,1$ erhält man $VIF_j = 10$. Derartige Grenzwerte findet man in der Literatur.

Die Bedeutung obiger Maße für Multikollinearität wird aus der folgenden Formel für die Varianz eines geschätzten Regressionskoeffizienten deutlich:[41]

Varianz der Regressionskoeffizienten von Variable X_j

$$Var(b_j) = \frac{\sigma^2}{\sum_{k=1}^{K} (x_{jk} - \overline{x}_j)^2 (1 - R_j^2)} = \sigma^2 \cdot \frac{1}{SS_{X_j}} \cdot \frac{1}{(1 - R_j^2)} \tag{1.36}$$

Es lassen sich damit drei Faktoren für die Schätzgenauigkeit eines Regressionskoeffizienten b_j identifizieren:

a) Die Varianz der Störgrößen: Je größer die Streuung der Störeinflüsse ist, desto ungenauer wird die Schätzung der Koeffizienten.

[39] Vgl. Belsley/Kuh/Welsch (1980), S. 93.

[40] In SPSS werden aus rechentechnischen Gründen Variablen mit einer Toleranz $< 0,0001$ ausgeschlossen.

[41] Vgl. dazu Belsley/Kuh/Welsch (1980), S. 93 ff.; Wooldridge (2016), S. 83 ff.; Fahrmeir/Kneib/Lang (2009), S. 101; Greene (2018), S. 94.

b) Die Streuung der Variablen X_j: Je geringer die Streuung einer Variablen ist, desto ungenauer wird die Schätzung ihres Regressionskoeffizienten.

c) Die Multikollinearität der Variablen X_j: Je größer die lineare Abhängigkeit von den übrigen Regressoren ist, desto ungenauer wird die Schätzung ihres Koeffizienten. Die Varianz des Schätzwertes steigt proportional mit dem VIF.

Die Faktoren a) und b) betreffen das Datenmaterial, der Faktor c) dagegen die Modellbildung.

Den ersten Faktor a) kann man eventuell verkleinern durch Ausschaltung von Störeinflüssen (z. B. Laboruntersuchungen) oder durch Erhöhung der Messgenauigkeit. Das betrifft allerdings nur die Daten zukünftiger Untersuchungen.

Der zweite Faktor b) ergibt sich aus dem Kehrwert der Streuung von X_j:

$$SS_{X_j} = \sum_{k=1}^{K} (x_{jk} - \overline{x}_j)^2$$

Die Größe der Streuung (Sum of Squares) umfasst zwei Aspekte:

- Die Variation der Variablen X_j um ihren Mittelwert
- Den Stichprobenumfang K

Es leuchtet ein, dass die mögliche Wirkung einer Variablen nicht feststellbar ist, wenn sie nicht variiert. Die Streuung ist dann Null. Aber auch bei geringer Variation wird die Schätzung des Regressionskoeffizienten immer ungenau sein. Das ist ein häufiges Problem bei Beobachtungsdaten. Aus diesem Grund werden oft Experimente durchgeführt, in denen die interessierende Variable manipuliert wird, um eine hinreichende Variation zu erhalten. Das ist natürlich nur möglich, wenn es sich um eine kontrollierbare Variable handelt. Die Streuung ist außerdem abhängig vom Stichprobenumfang, denn die Summe der Quadrate steigt mit K.

Der Untersucher kann also zumindest theoretisch die Genauigkeit der Schätzung erhöhen, indem er die unabhängigen Variablen manipuliert und/oder indem er den Stichprobenumfang erhöht. Das hilft natürlich nicht, wenn er sich mit dem vorhandenen Datenmaterial begnügen muss.

Der dritte Faktor c), der die Multikollinearität betrifft, kann vom Untersucher durch Modifikationen der Modellspezifikation verändert werden. Eine einfache Möglichkeit zur Begegnung von hoher Multikollinearität besteht darin, dass man Variablen mit niedriger Toleranz oder großem VIF aus dem Modell entfernt. Durch Verkleinerung des Modells verkleinern sich i.d.R. auch die Bestimmtheitsmaße R_j^2 und damit die Varianzinflationsfaktoren VIF_j. Dies ist unproblematisch, soweit es sich dabei um sachlogisch weniger wichtige Variablen handelt (z. B. der Einfluss des Wetters auf die Absatzmenge von Zahncreme).

Problematisch wird dieser Vorgang, wenn es sich bei der oder den betroffenen Variablen gerade um diejenigen handelt, deren Einfluss den Untersucher primär interessiert. Er steht dann oft vor dem Dilemma, entweder die Variable(n) zu entfernen und damit möglicherweise den Zweck der Untersuchung in Frage zu stellen, oder die Variable(n) in der Gleichung zu belassen und damit die Folgen der Multikollinearität in Kauf zu nehmen. Durch Erhöhung des Stichprobenumfangs kann der Untersucher eventuell die negative Wirkung der Multikollinearität kompensieren.

Andere Maßnahmen zur Beseitigung oder Umgehung von Multikollinearität bilden z. B. Transformationen der Variablen oder Ersetzung der Variablen durch *Faktoren*,

die mittels Faktorenanalyse gewonnen wurden.[42] Um die Wirkung der Multikollinearität besser abschätzen zu können, sollte der Untersucher in jedem Fall auch Alternativrechnungen mit verschiedenen Variablenkombinationen (Modellen) durchführen. Seine Sachkenntnis und sein subjektives Urteil müssen letztlich über die Einschätzung und Behandlung der Multikollinearität entscheiden.

1.2.5.7 Nicht-Normalverteilung der Störgrößen

Die letzte Annahme des linearen Regressionsmodells besagt, dass die Störgrößen normalverteilt sein sollen. Wir hatten darauf hingewiesen, dass diese Annahme für die Kleinstquadrate-Schätzung nicht benötigt wird, d. h. die KQ-Schätzer besitzen auch ohne diese Annahme die BLUE-Eigenschaft.[43]

Die Annahme der Normalverteilung der Störgrößen ist lediglich für die Durchführung statistischer Tests (t-test, F-test) von Bedeutung. Hierbei wird unterstellt, dass die zu testenden Schätzwerte der Regressionsparameter, also b_0 und b_j, normalverteilt sind. Wäre dies nicht der Fall, wären auch die Tests nicht gültig.

Wenn die Störgrößen normalverteilt sind, dann sind auch die Y-Werte, die die Störgrößen als additiven Term enthalten, normalverteilt. Und da die KQ-Schätzer Linearkombinationen der Y-Werte bilden (vgl. mathematischer Anhang), sind folglich auch b_0 und b_j normalverteilt.

Wir hatten oben ausgeführt, dass die Annahme angenähert normalverteilter Störgrößen in vielen Fällen plausibel ist, wenn diese durch Überlagerung zahlreicher und im Einzelnen relativ unbedeutender und voneinander unabhängiger Zufallsgrößen zustande kommt. Eine Rechtfertigung hierfür liefert der zentrale Grenzwertsatz der Statistik. Allerdings kann man nicht davon ausgehen, dass dies generell so ist.

Sind die Störgrößen nicht normalverteilt, so können aber die KQ-Schätzer trotzdem normalverteilt sein. Auch dies folgt wiederum aus dem zentralen Grenzwertsatz und den obigen Ausführungen. Allerdings gilt dies nur asymptotisch mit wachsender Zahl der Beobachtungen K. Ist die Zahl der Beobachtungen groß (etwa $K > 40$), sind damit die Signifikanztests unabhängig von der Verteilung der Störgrößen gültig.[44]

Unempfindlichkeit der Regressionsanalyse

Abbildung 1.27 fasst die wichtigsten Prämissen des linearen Regressionsmodells und die Konsequenzen ihrer Verletzung zusammen. Aufgrund der Vielzahl der Annahmen, die der Regressionsanalyse zugrunde liegen, mag deren Anwendbarkeit sehr eingeschränkt erscheinen. Das aber ist nicht der Fall. Die Regressionsanalyse ist recht unempfindlich gegenüber kleineren Verletzungen der obigen Annahmen und bildet ein äußerst flexibles und vielseitig anwendbares Analyseverfahren.

[42] Vgl. dazu das Kapitel 5 „Faktorenanalyse“ in diesem Buch. Bei einem Ersatz der Regressoren durch Faktoren muss man sich allerdings vergegenwärtigen, dass dadurch womöglich der eigentliche Untersuchungszweck in Frage gestellt wird. Eine andere Methode zur Begegnung von Multikollinearität ist die sog. Ridge Regression, bei der man zugunsten einer starken Verringerung der Varianz eine kleine Verzerrung der Schätzwerte in Kauf nimmt. Vgl. dazu z. B. Kmenta (1997), S. 440 ff.; Belsley/Kuh/Welsch (1980), S. 219 ff.

[43] Vgl. z. B. Kmenta (1997), S. 261.

[44] Zumindest unter sehr allgemeinen Bedingungen, nämlich dass die Störgrößen endliche Varianz besitzen und voneinander unabhängig sind. Vgl. hierzu Greene (2018), S. 60 ff.; Kmenta (1997), S. 262. Zum Testen auf Normalität ist es üblich, die Residuen zu plotten. Da die Normalverteilung symmetrisch ist, sollte dies auch für die Verteilung der Residuen gelten. Zu formalen Tests siehe Kmenta (1997), S. 265 ff.

Prämisse	Prämissenverletzung	Konsequenzen
Linearität in den Parametern	Nichtlinearität	Verzerrung der Schätzwerte
Vollständigkeit des Modells (Berücksichtigung aller relevanten Variablen)	Unvollständigkeit	Verzerrung der Schätzwerte
Homoskedastizität der Störgrößen	Heteroskedastizität	Ineffizienz
Unabhängigkeit der Störgrößen	Autokorrelation	Ineffizienz
Keine lineare Abhängigkeit zwischen den unabhängigen Variablen	Multikollinearität	Verminderte Präzision der Schätzwerte
Normalverteilung der Störgrößen	nicht normalverteilt	Ungültigkeit der Signifikanztests (F-Test und t-Test), wenn K klein ist

Abbildung 1.27: Prämissenverletzungen des linearen Regressionsmodells

1.3 Fallbeispiel

In einer Untersuchung über potenzielle Ursachen von Veränderungen im Margarineabsatz erhebt der Manager eines Margarineherstellers Daten über potenzielle, von ihm vermutete Einflussgrößen der Absatzveränderungen. Aufgrund seiner Erfahrung vermutet der Manager, dass die von ihm kontrollierten Größen Preis, Ausgaben für Verkaufsförderung sowie Zahl der Vertreterbesuche einen ursächlichen Einfluss auf den Margarineabsatz in seinen Verkaufsgebieten haben. Aus diesem Grunde erhebt er Daten über die Ausprägungen dieser Einflussgrößen in 37 Verkaufsgebieten, die zufällig ausgesucht werden. Er hofft, aufgrund dieser Stichprobe ein zuverlässiges Bild über die Wirkungsweise dieser Einflussgrößen auf den Margarineabsatz in allen Verkaufsgebieten zu gewinnen.

1.3.1 Blockweise Regressionsanalyse

Mit einer blockweisen Regressionsanalyse, in SPSS als Methode „Einschluss“ (Enter) bezeichnet, kann der Benutzer eine einzelne Variable oder Blöcke von Variablen in eine Regressionsgleichung einbeziehen. Um mittels des Programms SPSS ein Regressionsmodell unter Verwendung dieser Methode zu berechnen und zu überprüfen, ist zunächst die Prozedur „Regression“ aus dem Menüpunkt „Analysieren“ auszuwählen und sodann die Option „Linear“ (vgl. Abbildung 1.28).

Dialogfenster

Im nunmehr geöffneten Dialogfenster „Lineare Regression“ (vgl. Abbildung 1.29) werden zunächst die abhängige Variable (hier: MENGE) und eine oder mehrere unabhängige Variablen (hier: PREIS, AUSGABEN, BESUCHE) aus der Variablenliste ausgewählt und mittels der Option „Einschluss“ in die Regressionsfunktion einbezogen. Nach Anklicken von „OK“ erhält man das Ergebnis der Analyse, das in Abbildung 1.30 wiedergegeben ist.

Das erste wichtige Ergebnis sind die Regressionskoeffizienten b_j für die drei unabhängigen Variablen BESUCHE, PREIS, AUSGABEN sowie das konstante Glied. Diese finden sich im unteren Bereich der Abbildung 1.30 in der Tabelle „Koeffizienten“ in der ersten mit „B“ bezeichneten Spalte.

Die geschätzte Regressionsfunktion lautet damit:

$$\text{MENGE} = 722 + 34,26 \cdot \text{PREIS} + 0,635 \cdot \text{Ausgaben} + 8,31 \cdot \text{BESUCHE}$$

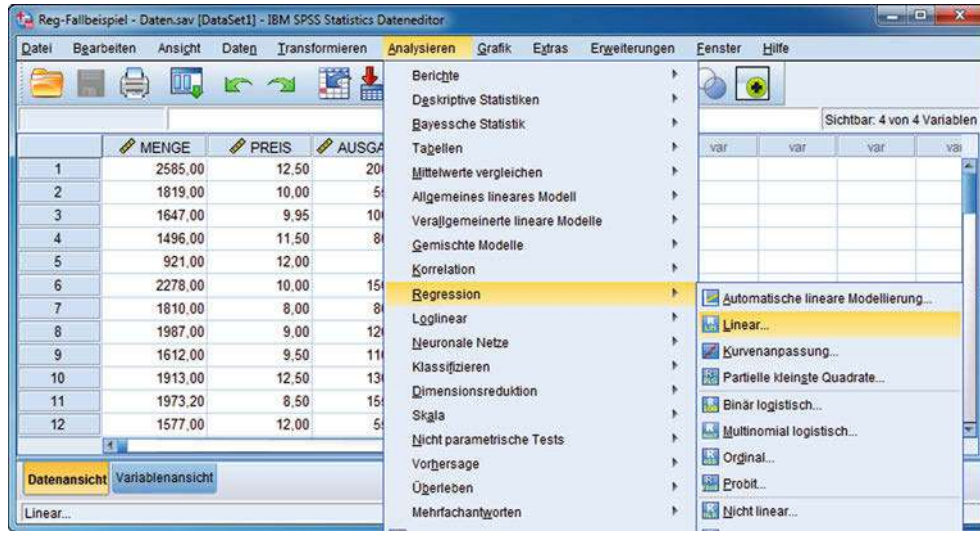


Abbildung 1.28: Daten-Editor mit Auswahl des Analyseverfahrens „Regression (Linear)“

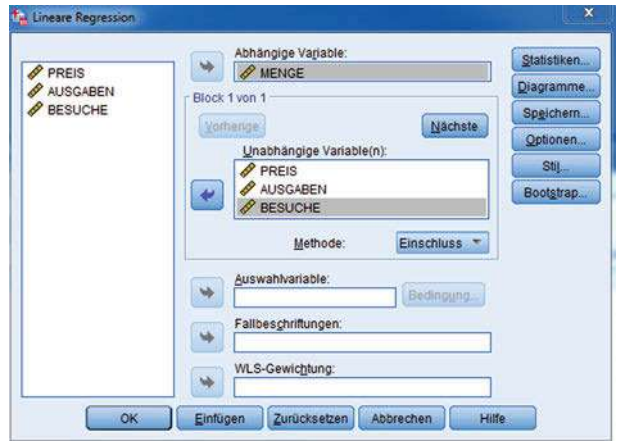


Abbildung 1.29: Dialogfenster „Lineare Regression“

Globale Gütemaße

Modell-zusammenfassung

In dem mit „Modellzusammenfassung“ überschriebenen Abschnitt finden sich die globalen Gütemaße. Das Bestimmtheitsmaß (R-Quadrat) beträgt hier $R^2 = 0,85$ (Formel 1.14). Die Größe $R = 0,92$ ist der multiple Korrelationskoeffizient (Wurzel aus R^2). Das korrigierte Bestimmtheitsmaß gemäß Formel (1.16) beträgt 0,83. Mit „Standardfehler des Schätzers“ ist die Standardabweichung der Residuen (Formel 1.20) gemeint, die hier 167,4 beträgt.

Der Wert für R^2, der besagt, dass rund 85 % der Streuung der Absatzmenge durch die drei Regressoren erklärt wird, ist für eine Marktuntersuchung dieser Art ein relativ hoher Wert. Allgemein gültige Aussagen, ab welcher Höhe ein R^2 als gut einzustufen ist, lassen sich jedoch nicht machen, da dies von der jeweiligen Problemstellung abhängig ist. Bei stark zufallsbehafteten Prozessen (z. B. Wetter, Börse) und großer Fallzahl kann auch ein R^2 von 0,1 akzeptabel sein.

Modellzusammenfassung[b]

Modell	R	R-Quadrat	Korrigiertes R-Quadrat	Standardfehler des Schätzers	Durbin-Watson-Statistik
1	,920[a]	,846	,832	167,39254	2,779

a. Einflußvariablen : (Konstante), BESUCHE, PREIS, AUSGABEN

b. Abhängige Variable: MENGE

ANOVA[a]

Modell		Quadratsumme	df	Mittel der Quadrate	F	Sig.
1	Regression	5064564,293	3	1688188,098	60,249	,000[b]
	Nicht standardisierte Residuen	924668,657	33	28020,262		
	Gesamt	5989232,950	36			

a. Abhängige Variable: MENGE

b. Einflußvariablen : (Konstante), BESUCHE, PREIS, AUSGABEN

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten: Regressionskoeffizient B	Nicht standardisierte Koeffizienten: Std.-Fehler	Standardisierte Koeffizienten: Beta	T	Sig.
1	(Konstante)	722,227	241,353		2,992	,005
	PREIS	-34,261	17,354	-,135	-1,974	,057
	AUSGABEN	,635	,054	,808	11,686	,000
	BESUCHE	8,306	1,787	,322	4,648	,000

a. Abhängige Variable: MENGE

Abbildung 1.30: SPSS-Output für die Regressionsanalyse

F-Test

Der F-Test wird in dem mit „ANOVA“ (Analysis of Variance) überschriebenen Abschnitt wiedergegeben (vgl. Formel 1.18). In der mit „Regression“ bezeichneten Zeile wird zunächst die durch das Regressionsmodell erklärte Streuung (Quadratsumme) ausgewiesen, daneben die Anzahl der Freiheitsgrade (df) und die erklärte Varianz (Mittel der Quadrate), die sich aus dem Quotient von Streuung und Freiheitsgraden ergibt. Analog kann man in der Zeile „Nichtstandardisierte Residuen“ die nicht erklärte Streuung, die zugehörigen Freiheitsgrade und die nicht erklärte Varianz ablesen.

Die Anzahl der Freiheitsgrade (df) ergibt sich durch:

$$df = J = 3 \qquad \text{für die erklärte Streuung}$$
$$df = K - J - 1 = 33 \qquad \text{für die nicht erklärte Streuung}$$

(Die Zahl der Freiheitgrade für die gesamte Streuung ergibt sich durch $3 + 33 = 36$) Für die F-Statistik erhält man damit gemäß Formel (1.18) oder (1.19) den Wert

$$F_{emp} = 60,25$$

Zum Testen der Nullhypothese, dass kein systematischer Zusammenhang besteht, ist dieser Wert mit einem theoretischen F-Wert für eine geforderte Irrtumswahrscheinlichkeit zu vergleichen. Nachfolgend sind die theoretischen F-Werte für verschiedene Irrtumswahrscheinlichkeiten, die man für die obigen Freiheitsgrade einem Tabellenwerk für die F-Verteilung entnehmen kann (siehe Anhang A.2 des Buches), wiederge-

geben:

$F = 2,9$ für Irrtumswahrscheinlichkeit 0,05
$F = 4,5$ für Irrtumswahrscheinlichkeit 0,01
$F = 5,2$ für Irrtumswahrscheinlichkeit 0,005

Der hier erzielte F-Wert ist weit größer und damit hoch signifikant. Folglich kann die Nullhypothese, dass das Modell nichts erklärt, abgelehnt werden.

Führt man den Test mittels p-Wert durch (siehe oben), kann man sich das Nachschlagen in Tabellen ersparen. In SPSS wird der p-Wert (empirische Signifikanz) des errechneten F-Wertes rechts in der Spalte „Sig." ausgewiesen. Der Wert beträgt hier 0,000 und es kann somit kein Zweifel an der Signifikanz des Modells bestehen.

Prüfung der Regressionskoeffizienten

Koeffizienten

In der Tabelle „Koeffizienten", der wir schon die Regressionskoeffizienten entnommen hatten, finden sich in der zweiten Spalte die Standardfehler s_{b_j} der Regressionskoeffizienten (vgl. Formel B4 im mathematischen Anhang). Diese werden für die Ermittlung der t-Werte sowie der Konfidenzintervalle der Koeffizienten (vgl. Abbildung 1.32) benötigt.

Die folgende Spalte enthält die standardisierten Regressionskoeffizienten $\hat{b}_j$ (Beta-Werte). Wir erkennen, dass die Ausgaben den höchsten Beta-Wert annehmen. Daraus können wir schließen, dass diese den stärksten Einfluss auf die Absatzmenge haben.

Entsprechend ist auch der t-Wert für die Ausgaben am höchsten. Auch hier sind, analog zum F-Test, die p-Werte (Signifikanzniveaus) der Regressionsparameter angegeben. Diese sind für die Variablen AUSGABEN und BESUCHE jeweils Null, womit deren Einfluss sich als hoch signifikant erweist.

Für den Preis dagegen erhält man mit p = 0,057 einen Wert, der leicht über dem üblicherweise geforderten Signifikanzniveau von 0,05 bzw. 5% liegt. Es wäre aber vorschnell, daraus zu folgern, dass der Preis hier keinen Einfluss hat. Eine Betrachtung des Vorzeichens des Preiskoeffizienten zeigt, dass dieses negativ und somit logisch richtig ist. Ein höherer Preis vermindert i.d.R. die Nachfrage. Mit diesem Vorwissen ließe sich hier ein einseitiger t-Test durchführen, für den sich der p-Wert halbieren lässt (siehe oben). Der Wert $p/2 = 0,028$ liegt deutlich unter 0,05 und somit sind hier alle drei Regressoren als signifikant anzusehen.

Als Faustregel lässt sich merken, dass ein Koeffizient (bei zweiseitigem Test) signifikant ist mit 5 % Irrtumswahrscheinlichkeit, wenn $t \geq 2$ gilt, also der zugehörige t-Wert größer Zwei ist. Damit lässt sich die Signifikanz eines t-Wertes auch ohne Tabelle und p-Wert beurteilen. Dies gilt allerdings nur für eine größere Anzahl von Beobachtungen, genaugenommen für $K \geq 60$, wie ein Blick in die t-Tabelle im Anhang zeigt. Der t-Wert des Preises liegt hier leicht unter Zwei.

Weitere Statistiken

Statistiken

Neben den durch das Programm SPSS standardmäßig ausgegebenen Statistiken (Schätzer, Anpassungsgüte des Modells) können im Dialogfenster „Statistiken" (vgl. Abbildung 1.31) weitere Statistiken ausgewählt werden. Hierzu gehören die Konfidenzintervalle der Regressionskoeffizienten sowie Statistiken, die dazu dienen, die Einhaltung der Prämissen des linearen Regressionsmodells zu überprüfen.

Abbildung 1.31: Dialogfenster „Statistiken“

Abbildung 1.32 enthält die Konfidenzintervalle der drei Regressionskoeffizienten sowie des konstanten Gliedes (95 %-Konfidenzintervall für B). Man sieht, dass der Koeffizient der Variablen PREIS das größte Konfidenzintervall unter den drei Regressionskoeffizienten besitzt und folglich dessen Schätzung am ungenauesten ist. Noch ungenauer ist allerdings der Schätzwert des konstanten Gliedes.

Koeffizienten[a]

Modell		95,0% Konfidenzintervalle für B		Kollinearitätsstatistik	
		Untergrenze	Obergrenze	Toleranz	VIF
1	(Konstante)	231,190	1213,264		
	PREIS	-69,568	1,045	,998	1,002
	AUSGABEN	,525	,746	,978	1,023
	BESUCHE	4,671	11,941	,976	1,024

a. Abhängige Variable: MENGE

Abbildung 1.32: Konfidenzintervalle und Kollinearitätsstatistik

Prüfung auf Multikollinearität

Zwecks Aufdeckung von Multikollinearität soll hier in einem ersten Schritt die Korrelationsmatrix auf erkennbare Abhängigkeiten unter den Regressoren geprüft werden (vgl. Abbildung 1.33). Starke Korrelationen unter den Regressoren liegen hier nicht vor, was jedoch noch keine Gewähr für das Fehlen von Multikollinearität bietet. Diese kann auch vorliegen, wenn alle paarweisen Korrelationen niedrig sind.

	MENGE	PREIS	AUSGABEN	BESUCHE
MENGE	1	-0,110	0,854	0,436
PREIS	-0,110	1	0,014	0,043
AUSGABEN	0,854	0,014	1	0,148
BESUCHE	0,436	0,043	0,148	1

Abbildung 1.33: Korrelationsmatrix

Kollinearitäts-statistik

In Abbildung 1.32 sind neben den Konfidenzintervallen der Regressionsparameter auch deren Toleranzen und Variance Inflation Factors (VIF) angegeben (vgl. Formel (1.34) und (1.35)). Die vorliegenden Werte lassen keine nennenswerte Multikollinearität erkennen.

Schwellenwert

Im Programm SPSS wird die Toleranz jeder unabhängigen Variable vor Aufnahme in die Regressionsgleichung geprüft. Die Aufnahme unterbleibt, wenn der Toleranzwert unter einem Schwellenwert von 0,0001 liegt. Dieser Schwellenwert, der sich vom Benutzer auch ändern lässt, bietet allerdings keinen Schutz gegen Multikollinearität, sondern gewährleistet nur die rechnerische Durchführbarkeit der Regressionsanalyse. Eine exakte Grenze für „ernsthafte Multikollinearität“ lässt sich nicht angeben (vgl. Abschnitt 1.2.5.6).

Analyse der Residuen

Prämissenverletzung

Zwecks Prüfung der Prämissen des linearen Regressionsmodells, die die Verteilung der Störgrößen betreffen, muss man auf die Residuen zurückgreifen, da die Störgrößen nicht beobachtbar sind. Hierbei geht es z. B. um Prüfung auf Autokorrelation und Heteroskedastizität oder die Prüfung auf Normalverteilung der Residuen.

In Abbildung 1.34 sind neben den beobachteten und geschätzten Werten der abhängigen Variablen MENGE, y_k und $\hat{y}_k$, auch die Residuen $e_k = \hat{y}_k - y_k$ in der rechten Spalte aufgelistet. In der ersten Spalte sind außerdem die standardisierten Residuen angegeben, die man durch Division der Residuen durch ihre Standardabweichung (den Standardfehler s = 167,4) erhält. Durch die Standardisierung ergibt sich jeweils ein Mittelwert von 0 und eine Standardabweichung von 1. Man bekommt diese Tabelle, indem man im Dialogfenster „Statistiken“ (Abbildung 1.31) „Fallweise Diagnose“ und „Alle Fälle“ wählt. Die Daten lassen sich auch zwecks weiterer Verarbeitung in der Datendatei abspeichern. Dazu ist die Option „Speichern“ im Dialogfenster „Lineare Regression“ aufzurufen (vgl. Abbildung 1.29).

Anstatt alle Fälle aufzulisten, kann man auch die Option „Ausreißer außerhalb“ wählen, wobei ein Schwellenwert angegeben werden kann. Voreingestellt ist der Wert 3 für ± 3 Standardabweichungen um den Nullpunkt. Es werden dann nur diejenigen Fälle aufgelistet, bei denen der Absolutwert des Residuums diesen Schwellenwert übersteigt und die damit als Ausreißer identifiziert werden.

Die Identifikation von Ausreißern ist wichtig, da diese Beobachtungen möglicherweise falsch oder untypisch sind und das Ergebnis verfälschen können. Die KQ-Schätzung reagiert empfindlich auf Ausreißer, da sie die Quadrate der Residuen minimiert. Die Wirkung eines Ausreißers hängt dabei nicht nur von der Größe des Residuums ab, sondern auch von der Abweichung der X-Werte von ihrem jeweiligen Mittelwert (leverage effect).[45]

[45] Siehe hierzu z. B. Belsley/Kuh/Welsch (1980), S. 16 ff.; Fox (2008), S. 241 ff.; Greene (2018), S. 104 ff.

Fallweise Diagnose[a]

Fallnummer	Standardisierte Residuen	MENGE	Nicht standardisierter vorhergesagter Wert	Nicht standardisierte Residuen
1	,687	2585,00	2469,9338	115,06625
2	1,202	1819,00	1617,7660	201,23398
3	-1,147	1647,00	1838,9247	-191,92468
4	,467	1496,00	1417,8854	78,11459
5	-,376	921,00	983,8658	-62,86578
6	,587	2278,00	2179,7878	98,21225
7	-,359	1810,00	1870,0359	-60,03585
8	,278	1987,00	1940,3953	46,60470
9	-1,232	1612,00	1818,2035	-206,20345
10	,818	1913,00	1776,0354	136,96462
11	-,391	1973,20	2038,6870	-65,48698
12	,215	1577,00	1540,9372	36,06281
13	-,462	2125,60	2202,8589	-77,25888
14	-1,366	1866,20	2094,8515	-228,65151
15	,436	1620,60	1547,5846	73,01543
16	,096	2628,80	2612,6907	16,10933
17	,374	2298,60	2236,0036	62,59636
18	-,090	2225,90	2240,9231	-15,02306
19	-,187	1777,70	1808,9384	-31,23840
20	-1,527	1912,40	2167,9279	-255,52785
21	,247	2034,40	1993,1050	41,29496
22	-,075	2445,40	2457,9927	-12,59272
23	1,167	2667,00	2471,6104	195,38965
24	-,882	1172,10	1319,7169	-147,61689
25	-2,257	1275,30	1653,0450	-377,74497
26	2,545	2421,60	1995,5254	426,07464
27	-1,381	1761,40	1992,5073	-231,10729
28	,722	1569,50	1448,6338	120,86623
29	,489	1399,90	1318,1156	81,78441
30	-,320	1812,50	1866,0826	-53,58255
31	,824	2092,00	1953,9925	138,00753
32	1,251	1811,10	1601,7520	209,34801
33	-,287	1563,80	1611,8156	-48,01561
34	-,330	2212,80	2268,0386	-55,23864
35	,039	2432,60	2426,0284	6,57159
36	1,266	1855,70	1643,7717	211,92827
37	-1,046	1791,50	1966,6305	-175,13048

a. Abhängige Variable: MENGE

Abbildung 1.34: Y-Werte und Residuen (fallweise Diagnose)

Residuenstatistik[a]

	Minimum	Maximum	Mittelwert	Std.-Abweichung	N
Nicht standardisierter vorhergesagter Wert	983,8658	2612,6907	1902,5027	375,07645	37
Nicht standardisierte Residuen	-377,74496	426,07465	,00000	160,26616	37
Standardisierter vorhergesagter Wert	-2,449	1,893	,000	1,000	37
Standardisierte Residuen	-2,257	2,545	,000	,957	37

a. Abhängige Variable: MENGE

Abbildung 1.35: Statistik der Schätzwerte und der Residuen

Abbildung 1.35 zeigt eine Zusammenstellung von Minima und Maxima sowie Mittelwert und Standardabweichung der fallweisen Werte. Alle standardisierten Residuen liegen innerhalb eines Intervalls von ±3 Standardabweichungen um den Nullpunkt, d.h. es sind keine ernsthaften Ausreißer vorhanden. Zwei Werte liegen außerhalb eines Intervalls von ±2 Standardabweichungen. Es sind dies die Fälle 25 und 26.

Betrachtet man diese beiden Fälle als Ausreißer und eliminiert sie, so verbessert sich die Schätzung. Das Bestimmtheitsmaß steigt von 0,85 auf 0,89 und die Wirkung des Preises wird deutlicher. Der Regressionskoeffizient ändert sich von -34,3 auf -43,1 und der p-Wert sinkt von 0,057 auf 0,006. Die Wirkung des Preises ist damit jetzt hoch signifikant. Bezüglich der übrigen Koeffizienten ergeben sich keine wesentlichen Änderungen.

Diagramme zur Prüfung der Residuen

Für die Prüfung der Residuen ist deren Visualisierung äußerst hilfreich. Die Prozedur „Regression“ von SPSS bietet hierfür diverse Diagramme an. Diese können über das Dialogfenster „Diagramme“ aufgerufen werden (vgl. Abbildung 1.36).

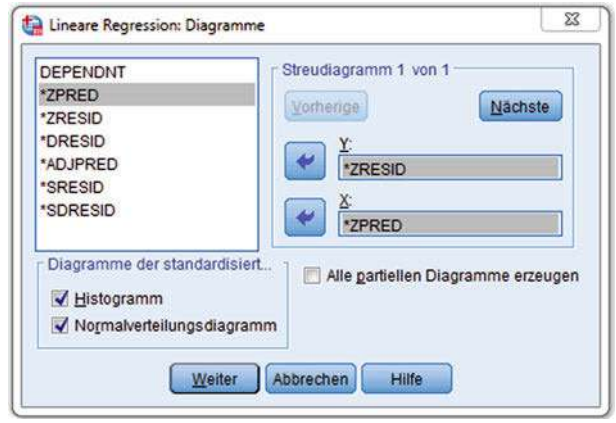


Abbildung 1.36: Dialogfenster „Diagramme“

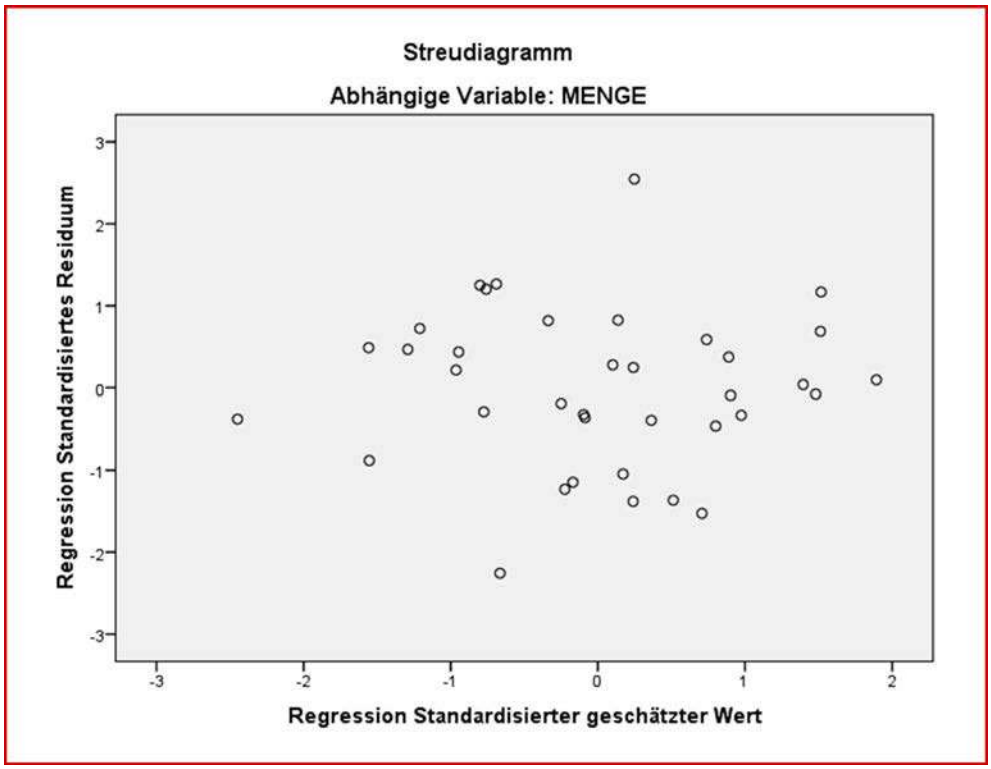


Abbildung 1.37: Streudiagramm der Residuen über der Menge

Abbildung 1.37 zeigt ein Streudiagramm der standardisierten Residuen. Es kann wie folgt erstellt werden. Mit *ZRESID werden in SPSS die Daten der standardisierten Residuen bezeichnet, die wir plotten wollen. Sie sollen auf der Y-Achse des Streudiagramms abgetragen werden. Für die horizontale X-Achse wählen wir die standardisierten $\hat{y}$-Werte (standardisierte geschätzte Mengen), die mit *ZPRED bezeichnet werden. Stattdessen hätten wir auch die nichtstandardisierten $\hat{y}$-Werte, die mit DEPENDT bezeichnet werden, wählen können. Alle diese Werte lassen sich, wie schon erwähnt, zwecks weiterer Verwendung in der Daten-Datei abspeichern.

Die Residuen sollten nicht in einer systematischen Beziehung zu den unabhängigen Variablen stehen (Annahme A3 des linearen Regressionsmodells). Da die geschätzte abhängige Variable eine Linearkombination der unabhängigen Variablen bildet, kann sie ersatzweise benutzt werden. Detailliertere Aufschlüsse erlangt man, wenn man die Residuen über jede der unabhängigen Variablen plotted. Dies kann nach Abspeichern der Werte mit den Grafik-Funktionen von SPSS erfolgen.

Die Residuen scheinen hier zufällig verteilt zu sein ohne erkennbares Muster. Sie streuen gleichmäßig um eine horizontale Linie durch den Nullpunkt der Y-Achse. Es ist somit keine systematische Beziehung zu den $\hat{y}$-Werten (Schätzwerten der abhängigen Variable) erkennbar. Die beiden „Ausreißer“ (Fall 25 und 26) sind deutlich erkennbar. Zwecks Prüfung auf Normalverteilung kann ein Histogramm der Residuen erstellt werden. Die Eingabe in das Dialogfenster zeigt Abbildung 1.36 und das Ergebnis zeigt Abbildung 1.38. Die Verteilung erscheint annähernd symmetrisch und nach beiden Seiten abfallend. Sie widerspricht damit nicht der Annahme einer Normalverteilung,

der sie sich mit steigender Anzahl der Beobachtungen wahrscheinlich annähern würde. Kleinere Abweichungen von der Normalverteilung beeinträchtigen nicht die Gültigkeit der statistischen Tests.

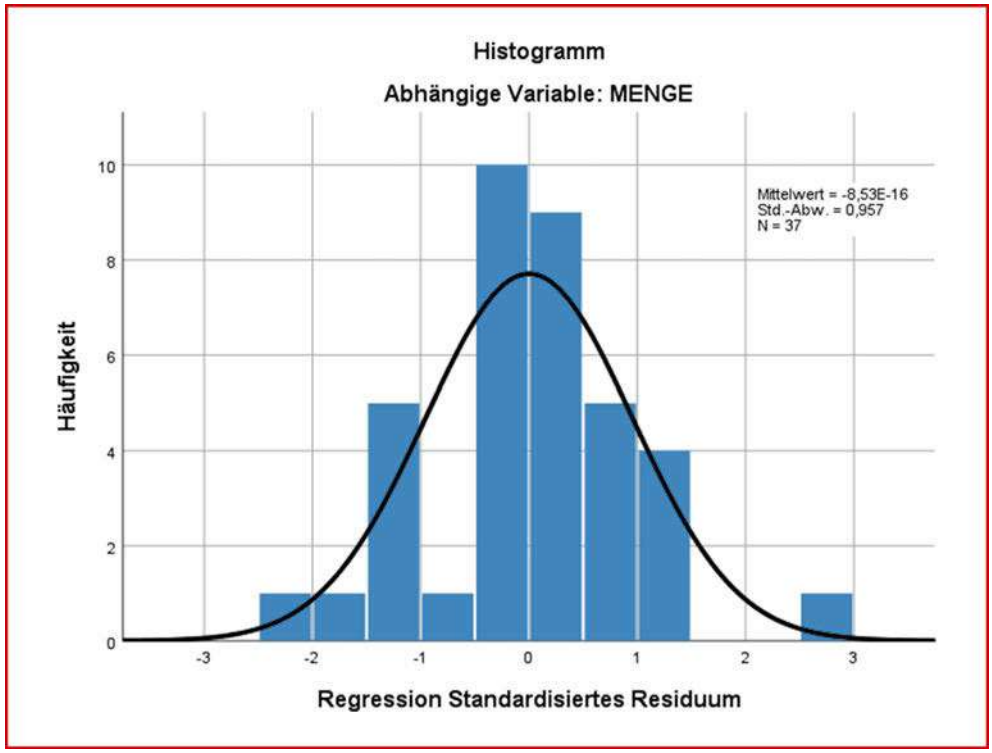


Abbildung 1.38: Histogramm der standardisierten Residuen mit Normalverteilungskurve

Zusätzlich kann zur Prüfung auf Normalverteilung ein P-P-Diagramm (Probability-Probability-Plot) erstellt werden.[46] Der Aufruf erfolgt durch Wahl der Option „Normalverteilungsdiagramm“ im Dialogfenster „Diagramme“ (Abbildung 1.36). Man erhält hier das P-P-Diagramm in Abbildung 1.39. Es werden die bei Normalverteilung erwarteten Wahrscheinlichkeiten der Residuen gegen die beobachteten Wahrscheinlichkeiten geplottet. Bei perfekter Normalverteilung der Residuen würden alle Punkte auf der Diagonalen liegen. Es ergeben sich hier leichte Abweichungen, die auch aus dem Histogramm ersichtlich sind.

Prüfung auf Heteroskedastizität

Die Annahme A4 des linearen Regressionsmodells besagt, dass die Störgrößen eine konstante Varianz haben sollen (Homoskedastizität). Die Streuung der Residuen über der geschätzten abhängigen Variable in Abbildung 1.37 deutet nicht auf eine Verletzung dieser Annahme (Vorliegen von Heteroskedastizität) hin. Auch eine Regression der absoluten Residuen auf die unabhängigen Variablen gemäß (1.32) zeigt hier keine signifikanten Koeffizienten, die auf Heteroskedastizität hindeuten würden.

[46] Das P-P-Diagramm ist ähnlich dem bekannteren Q-Q-Diagramm (Quantil-Quantil-Plot) und liefert gleichartige Information.

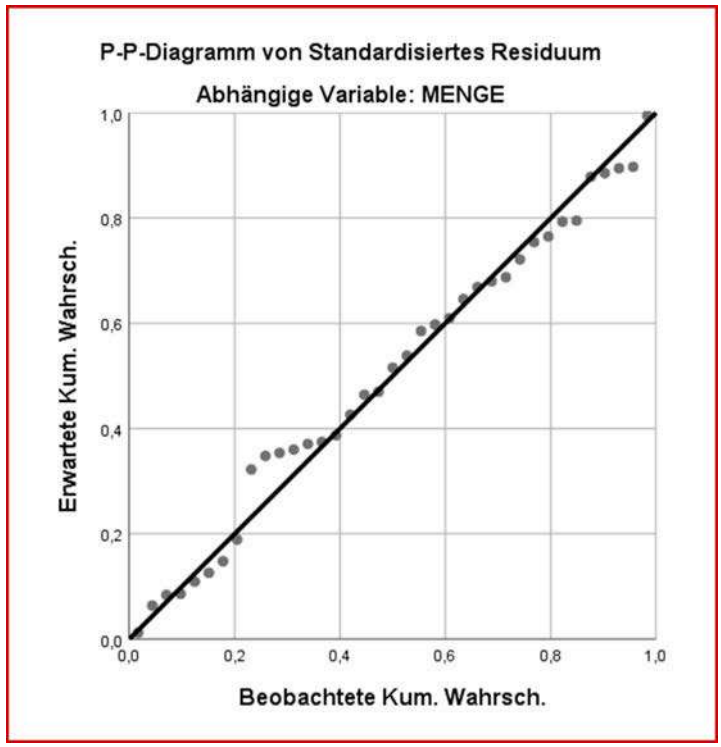


Abbildung 1.39: P-P-Diagramm für die Residuen

Durbin/Watson-Test

Da es sich bei den Beobachtungen hier nicht um Zeitreihendaten handelt, sondern um Querschnittsdaten, deren Reihenfolge sich beliebig verändern lässt, macht eine Prüfung auf Vorliegen von Autokorrelation eigentlich keinen Sinn. Wir wollen aber trotzdem kurz die Durchführung des Durbin/Watson-Testes an den vorliegenden Daten demonstrieren.

Der Wert der Durbin/Watson-Statistik, $d = 2{,}779$, wurde in Abbildung 1.30 ausgewiesen. Als Grenzwerte ergeben sich aufgrund der Durbin-Watson-Tabelle im Anhang für $\alpha = 5\%$ und bei 37 Fällen und drei Regressoren (Tabelle A.10 für den *zweiseitigen* Test): $d_u^+ = 1{,}21$ und $d_o^+ = 1{,}56$. Man erhält damit $4-d+u = 2{,}79$ und $4-d+o = 2{,}44$. Hieraus ergibt sich, dass der empirische Wert von d im Unschärfebereich liegt und somit keine Aussage über das Vorliegen von Autokorrelation getroffen werden kann (vgl. die Entscheidungsregeln für den Durbin/Watson-Test in Abbildung 1.25).

1.3.2 Schrittweise Regressionsanalyse

Stepwise Regression

Das Programm SPSS bietet eine Reihe von Möglichkeiten, um aus einer Menge von unabhängigen Variablen unterschiedliche Kombinationen auszuwählen und somit unterschiedliche Regressionsmodelle zu formulieren. Mit den drei unabhängigen Variablen „PREIS“, „AUSGABEN“ und „BESUCHE“ lassen sich insgesamt sieben verschiedene Modelle (Regressionsgleichungen) bilden: drei mit einer unabhängigen Variable, drei mit zwei unabhängigen Variablen und eines mit drei unabhängigen Variablen. Die An-

zahl der möglichen Kombinationen erreicht mit wachsender Anzahl der unabhängigen Variablen sehr schnell beträchtliche Größen. Es ist zwar möglich, alle Kombinationen durchrechnen zu lassen. Für den Untersucher verbleibt das Problem, die alternativen Modelle zu vergleichen und unter diesen auszuwählen. Weniger aufwändig sind die beiden folgenden Vorgehensweisen:

- Der Untersucher formuliert ein oder einige Modelle, die ihm aufgrund von theoretischen oder sachlogischen Überlegungen sinnvoll erscheinen und überprüft diese empirisch durch Anwendung der Regressionsanalyse (zur Auswahl der unabhängigen Variablen wird hierzu in SPSS die Methode „Einschluss“ verwendet).
- Der Untersucher lässt sich vom Computer eine Auswahl von Modellen, die sein Datenmaterial gut abbilden (dies ist in SPSS mittels der Methode „Schrittweise“ möglich), zeigen und versucht sodann, diese sinnvoll zu interpretieren.

Gefahren

Die zweite Alternative ist besonders verlockend und findet in der empirischen Forschung durch die Verfügbarkeit leistungsfähiger Computer-Programme zunehmende Verbreitung. Es besteht hierbei jedoch die Gefahr, dass sachlogische Überlegungen in den Hintergrund treten können, d. h. dass der Untersucher mehr dem Computer als seinem gesunden Menschenverstand vertraut. Der Computer kann nur nach statistischen Kriterien wählen, nicht aber erkennen, ob ein Modell auch inhaltlich sinnvoll ist.

Statistisch signifikante Zusammenhänge sollten vom Untersucher nur dann akzeptiert werden, wenn sie seinen sachlogischen Erwartungen entsprechen. Andererseits sollte der Untersucher bei Nichtsignifikanz eines Zusammenhanges nicht folgern, dass kein Zusammenhang besteht, wenn ansonsten das Ergebnis sachlich korrekt ist. Andernfalls sollte man bei widersprüchlichen Ergebnissen oder sachlogisch unbegründeten Einflussfaktoren nicht zögern, diese aus dem Regressionsmodell zu entfernen, auch wenn der Erklärungsanteil dadurch sinkt.

Algorithmusgesteuerter Auswahlprozess

Nachdem wir gezeigt haben, wie in SPSS mit der Methode „Einschluss“ die unabhängigen Variablen ausgewählt und blockweise in die Regressionsgleichung einbezogen werden, zeigen wir nun die schrittweise Regression, bei der die Auswahl der Variablen automatisch (durch einen Algorithmus gesteuert) erfolgt. In SPSS lässt sie sich durch die Anweisung „Schrittweise“ (Stepwise) aufrufen (vgl. Abbildung 1.40). Bei der schrittweisen Regression werden die unabhängigen Variablen einzeln nacheinander in die Regressionsgleichung einbezogen, wobei jeweils diejenige Variable ausgewählt wird, die ein bestimmtes Gütekriterium maximiert.

Im ersten Schritt wird eine einfache Regression mit derjenigen Variablen durchgeführt, die die höchste (positive oder negative) Korrelation mit der abhängigen Variablen aufweist. In den folgenden Schritten wird dann jeweils die Variable mit der höchsten partiellen Korrelation ausgewählt. Aus der Rangfolge der Aufnahme lässt sich die statistische Wichtigkeit der Variablen erkennen.

Die Anzahl der durchgeführten Analysen bei der schrittweisen Regression ist bedeutend geringer als die Anzahl der kombinatorisch möglichen Regressionsgleichungen. Bei 10 unabhängigen Variablen sind i. d. R. auch nur 10 Analysen gegenüber 1.023 möglichen Analysen durchzuführen. Die Zahl der durchgeführten Analysen kann allerdings schwanken. Einerseits kann sie sich verringern, wenn Variablen ein bestimmtes Aufnahmekriterium nicht erfüllen. Andererseits kann es vorkommen, dass eine bereits ausgewählte Variable wieder aus der Regressionsgleichung entfernt wird, weil sie durch die Aufnahme anderer Variablen an Bedeutung verloren hat und das Aufnahmekri-

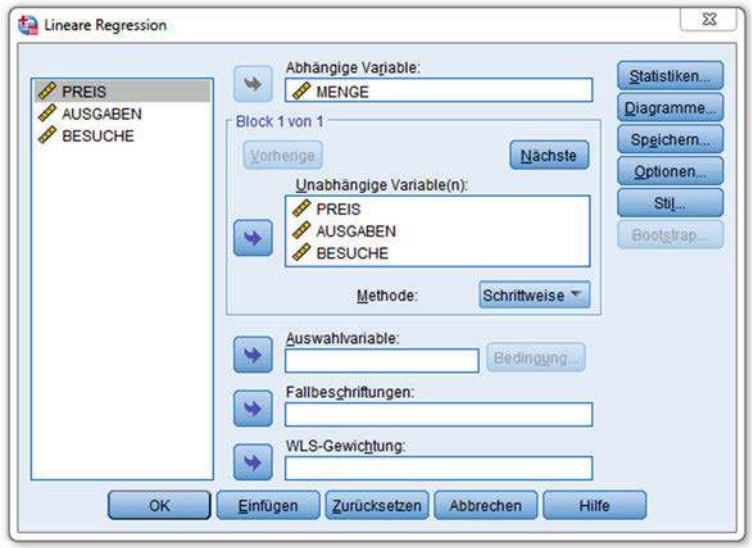


Abbildung 1.40: Dialogfenster „Lineare Regression“: Auswahl „Schrittweise“

terium nicht mehr erfüllt. Es besteht allerdings keine Gewähr, dass die schrittweise Regression immer zu einer optimalen Lösung führt.

Abbildung 1.41 und Abbildung 1.42 zeigen das Ergebnis der schrittweisen Regressionsanalyse für das Fallbeispiel. Das Zielkriterium für die sukzessive Auswahl der unabhängigen Variablen bildet die Erhöhung von R-Quadrat. Im ersten Schritt wird diejenige Variable ausgewählt, die mit der abhängigen Variablen den höchsten Korrelationskoeffizienten und damit das höchste R-Quadrat hat. Das ist hier die Variable AUSGABEN und man erhält Modell 1 in Abbildung 1.42. Im zweiten Schritt wird dann die Variable BESUCHE ausgewählt, die R-Quadrat am stärksten erhöht. und man erhält Modell 2. Hiermit endet der *Auswahlalgorithmus*, d.h. die Variable PREIS wird nicht aufgenommen.

Die Variable PREIS wird nicht aufgenommen, da sie das Aufnahmekriterium des Auswahlalgorithmus hier nicht erfüllt. Das ist der p-Wert, den der Preis bei Aufnahme in das Modell erhalten würde. Wie wir schon wissen, gilt hier der Wert $p = 0{,}057$ (vgl. Abbildung 1.30), der auch unten in Abbildung 1.42 angegeben ist. Eine Variable wird nur dann aufgenommen, wenn ihr p-Wert unter einem Schwellenwert PIN liegt. Voreingestellt ist $PIN = 0{,}05$. Im Dialogfenster „Optionen“ (vgl. Abbildung 1.43) kann dieser Wert variiert werden. Erhöhen wir PIN auf 0,06, so wird auch der Preis aufgenommen und wir erhalten als Modell 3 das Modell, welches wir bereits mit der blockweisen Regressionsanalyse geschätzt haben.

Im Dialogfenster „Optionen“ ist das Aufnahmekriterium mit „F-Wahrscheinlichkeit“ bezeichnet. Gemeint ist der p-Wert für den F-Wert des partiellen Korrelationskoeffizienten zwischen abhängiger und unabhängiger Variable. Er ist gleich dem p-Wert des t-Wertes der Variable nach Aufnahme in das Modell. Die partiellen Korrelationskoeffizienten sind in Abbildung 1.42 unten neben den p-Werten angegeben.

Für die schrittweise Regression sind jeweils zwei *Schwellenwerte* anzugeben, ein Wert für die Aufnahme einer Variablen in das Modell und ein Wert für die

Aufgenommene/Entfernte Variablen[a]

Modell	Aufgenommene Variablen	Entfernte Variablen	Methode
1	AUSGABEN	.	Schrittweise Selektion (Kriterien: Wahrscheinlichkeit von F-Wert für Aufnahme <= ,050, Wahrscheinlichkeit von F-Wert für Ausschluß >= ,100).
2	BESUCHE	.	Schrittweise Selektion (Kriterien: Wahrscheinlichkeit von F-Wert für Aufnahme <= ,050, Wahrscheinlichkeit von F-Wert für Ausschluß >= ,100).

a. Abhängige Variable: MENGE

Modellzusammenfassung[c]

Modell	R	R-Quadrat	Korrigiertes R-Quadrat	Standardfehler des Schätzers
1	,854[a]	,730	,722	215,09415
2	,910[b]	,827	,817	174,38043

a. Einflußvariablen : (Konstante), AUSGABEN

b. Einflußvariablen : (Konstante), AUSGABEN, BESUCHE

c. Abhängige Variable: MENGE

Abbildung 1.41: SPSS-Output für die schrittweise Regressionsanalyse

Elimination einer Variablen aus dem Modell:

PIN: p-Wert des F-Wertes für Aufnahme einer Variablen. Voreingestellt ist der Wert PIN = 0,05.

POUT: p-Wert des F-Wertes für Elimination einer Variablen. Voreingestellt ist der Wert POUT = 0,1.

Es muss immer POUT > PIN gelten, da der Algorithmus ansonsten möglicherweise kein Ende findet. Je kleiner PIN und POUT, desto mehr werden die Anforderungen für die Aufnahme einer Variablen oder deren Verbleib im Modell verschärft. Alternativ können anstelle der p-Werte des F-Wertes auch die F-Werte selbst als Kriterium verwendet und im Dialogfenster „Optionen“ eingestellt werden. Die beiden Kriterien sind nicht völlig identisch, da der p-Wert für einen bestimmten F-Wert sich mit der Anzahl der Variablen im Modell ändern kann. Für große Stichproben gleichen sich die Ergebnisse an. Den p-Werten ist der Vorzug zu geben, auch weil die Einstellung der F-Werte schwieriger ist.

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten: Regressionskoeffizient B	Nicht standardisierte Koeffizienten: Std.-Fehler	Standardisierte Koeffizienten: Beta	T	Sig.
1	(Konstante)	1069,992	92,672		11,546	,000
	AUSGABEN	,671	,069	,854	9,719	,000
2	(Konstante)	379,108	174,467		2,173	,037
	AUSGABEN	,634	,057	,807	11,204	,000
	BESUCHE	8,160	1,860	,316	4,388	,000

a. Abhängige Variable: MENGE

Ausgeschlossene Variablen[a]

Modell		Beta In	T	Sig.	Partielle Korrelation	Kollinearitäts-statistik: Toleranz
1	PREIS	-,122[b]	-1,408	,168	-,235	1,000
	BESUCHE	,316[b]	4,388	,000	,601	,978
2	PREIS	-,135[c]	-1,974	,057	-,325	,998

a. Abhängige Variable: MENGE

b. Einflußvariablen im Modell: (Konstante), AUSGABEN

c. Einflußvariablen im Modell: (Konstante), AUSGABEN, BESUCHE

Abbildung 1.42: (Fortsetzung von Abbildung 1.41)

Abbildung 1.43: Dialogfenster „Optionen“

1.3.3 SPSS-Kommandos

In Abbildung 1.44 ist abschließend die Syntaxdatei mit den SPSS-Kommandos für das Fallbeispiel wiedergegeben. Vergleiche hierzu die Ausführungen im einleitenden Kapitel dieses Buches. Wie oben beschrieben, wird zunächst eine blockweise Regression und anschließend eine schrittweise Regression durchgeführt.

Wenn der Datensatz fehlende Werte enthalten würde, was in der Praxis häufig vorkommt, dann kann dies durch das MISSING-VALUES-Kommando berücksichtigt

```
* MVA: Fallbeispiel Regressionsanalyse.
* DATENDEFINITION.
DATA LIST FREE / MENGE PREIS AUSGABEN BESUCHE.
MISSING VALUES ALL (9999).

BEGIN DATA.
2585,0  12,50   2000    109
1819,0  10,00   550     107
1647,0  9,95    1000    99
……..
1791,5  12,50   1600    79
END DATA.

* Regressionsanalyse für den Margarinemarkt nach der Methode "Enter".
REGRESSION
   VARIABLES MENGE PREIS AUSGABEN BESUCHE
  /MISSING LISTWISE
  /STATISTICS R ANOVA COEFF CI TOL
  /DESCRIPTIVES CORR
  /DEPENDENT MENGE
  /ENTER PREIS AUSGABEN BESUCHE
  /CASEWISE DEPENDENT PRED RESID OUTLIERS (0)
  /SCATTERPLOT (*RESID,*PRED)
  /RESIDUALS DURBIN HISTOGRAM(ZRESID) NORMPROB(ZRESID)
  /CASEWISE PLOT(ZRESID) ALL
  /SAVE PRED RESID ZRESID.

* Regressionsanalyse für den Margarinemarkt nach der Methode "Stepwise".
REGRESSION
   VARIABLES MENGE PREIS AUSGABEN BESUCHE
  /MISSING LISTWISE
  /CRITERIA PIN (0.05) POUT (0.1)
  /DEPENDENT MENGE
  /METHOD STEPWISE PREIS AUSGABEN BESUCHE.
```

Abbildung 1.44: SPSS-Job zur Regressionsanalyse

werden. Wäre z. B. die „0“ für die Ausgaben in Beobachtung 5 (vgl. Abbildung 1.6) ein fehlender Wert (user missing value), dann müsste dieser Wert anstelle der 9999 im MISSING-VALUES-Kommando eingesetzt werden.

1.4 Anwendungsempfehlungen

Für die praktische Anwendung der Regressionsanalyse sollen abschließend einige Empfehlungen gegeben werden, die rezeptartig formuliert sind und den schnellen Zugang zur Anwendung der Methode erleichtern sollen.

1. Das Problem, das es zu untersuchen gilt, muss genau definiert werden: Welche Größe soll erklärt werden? Die zu erklärende Variable und die erklärenden Variablen sollten metrisches Skalenniveau haben. Binäre Variablen lassen sich wie metrische Variablen behandeln. Nominale Variablen mit mehr als zwei Ausprägungen lassen sich einbeziehen, wenn man sie in Dummy-Variablen transformiert.

2. Es ist viel Sachkenntnis und Überlegung einzubringen, um mögliche Einflussgrößen, die auf die zu erklärende Variable einwirken, zu erkennen und zu definieren. Die wichtigen Einflussgrößen sollten im Modell enthalten sein, aber mehr muss nicht besser sein. Eine Variable sollte nur dann berücksichtigt werden, wenn sachlogische Gründe hierfür bestehen.

3. Die Zahl der Beobachtungen muss genügend groß sein. Sie sollte möglichst doppelt so groß sein wie die Anzahl der Variablen in der Regressionsgleichung.

4. Vor Beginn der Rechnung sollten aufgrund der vorhandenen Sachkenntnis zunächst hypothetische Regressionsmodelle mit den vorhandenen Variablen formuliert werden. Dabei sollten auch die Art und Stärke der Wirkungen von berücksichtigten Variablen überlegt werden.
5. Nach Schätzung einer Regressionsfunktion ist zunächst das Bestimmtheitsmaß auf Signifikanz zu prüfen. Wenn kein signifikantes Testergebnis erreichbar ist, muss der ganze Regressionsansatz verworfen werden.
6. Anschließend sind die einzelnen Regressionskoeffizienten sachlogisch (auf Vorzeichen) und statistisch (auf Signifikanz) zu prüfen.
7. Die gefundene Regressionsgleichung ist auf Einhaltung der Prämissen des linearen Regressionsmodells zu prüfen.
8. Eventuell sind Variablen aus der Gleichung zu entfernen oder neue Variablen aufzunehmen. Die Modellbildung ist oft ein iterativer Prozess, bei dem der Untersucher auf Basis von empirischen Ergebnissen neue Hypothesen formuliert und diese anschließend wieder überprüft.
9. Wenn die gefundene Regressionsgleichung alle Prämissen-Prüfungen überstanden hat, erfolgt die Überprüfung an der Realität.

1.5 Mathematischer Anhang

Ergänzend zum Text sind nachfolgend die Formeln zur Durchführung einer einfachen und einer multiplen Regressionsanalyse in knapper Form zusammengestellt.

A. Einfache Regression

Die zu schätzende Regressionsfunktion lautet:

$$\hat{Y} = b_0 + b_1 X \tag{A1}$$

Durch Einbeziehung der Residuen e erhält man:

$$Y = b_0 + b_1 X + e \tag{A2}$$

Für die einzelnen Beobachtungen schreiben wir:

$$y_k = b_0 + b_1 x_k + e_k \tag{A3}$$

Schätzung der Parameter

Gemäß dem KQ-Kriterium sind die Parameter b_0 und b_1 gesucht, für die die Summe der quadrierten Residuen minimal wird:

$$S = \sum_{k=1}^{K} e_k^2 \qquad \rightarrow \quad \text{min!} \tag{A4}$$

Mit (A3) erhält man für das Zielkriterium den Ausdruck

$$S = \sum_{k=1}^{K} (y_k - b_0 - b_1 x_k)^2 \qquad \rightarrow \quad \text{min!} \tag{A5}$$

Durch partielle Differentiation nach b_0 und b_1 erhält man unter Weglassung von Index k an den Summenzeichen die folgenden Bedingungen erster Ordnung für das gesuchte Minimum:

$$\frac{\delta S}{\delta b_0} = 2\sum(y_k - b_0 - b_1 x_k)(-1) = 0 \tag{A6}$$

$$\frac{\delta S}{\delta b_1} = 2\sum(y_k - b_0 - b_1 x_k)(-x_k) = 0 \tag{A7}$$

Daraus folgt:

$$\sum(y_k - b_0 - b_1 x_k) = 0 \qquad \Rightarrow \sum e_k = 0 \tag{A8}$$

$$\sum(y_k - b_0 - b_1 x_k)x_k = 0 \qquad \Rightarrow \sum e_k x_k = 0 \tag{A9}$$

Durch Umformung erhält man hieraus die sog. *Normalgleichungen*:

$$\sum y_k = Kb_0 + b_1 \sum x_k \tag{A10}$$

$$\sum y_k x_k = b_0 \sum x_k + b_1 \sum x_k^2 \tag{A11}$$

Durch Auflösen von (A10) nach b_0 erhält man

$$b_0 = \frac{1}{K}\sum y_k - b_1 \frac{1}{K}\sum x_k = \overline{y} - b_1 \overline{x} \tag{A12}$$

Dies entspricht Formel (1.9) zur Berechnung des konstanten Gliedes der Regressionsfunktion. Durch Einsetzen in (A11) erhält man:

$$\sum y_k x_k = \frac{1}{K}\sum y_k \sum x_k - b_1 \frac{1}{K}(\sum x_k)(\sum x_k) + b_1 \sum x_k^2$$

Durch Auflösen nach b_1 erhält man hieraus Formel (1.8)

$$b_1 = \frac{K\sum x_k y_k - (\sum x_k)(\sum y_k)}{K\sum x_k^2 - (\sum x_k)^2} \tag{A13}$$

Mit Hilfe der Mittelwerte der x- und y-Werte lässt sich diese Gleichung wie folgt vereinfachen:

$$b_1 = \frac{\sum(x_k - \overline{x})\sum(y_k - \overline{y})}{\sum(x_k - \overline{x})^2} \tag{A14}$$

Schätzfehler der Parameter:

Der Standardfehler des Regressionskoeffizienten b_1 errechnet sich durch

$$s_{b_1} = s\sqrt{\frac{1}{\sum(x_k - \overline{x})^2}} \tag{A15}$$

wobei s den Standardfehler der Regression bezeichnet:

$$s = \sqrt{\frac{\sum e_k^2}{K - J - 1}} \tag{A16}$$

Für den Standardfehler des konstanten Gliedes b_0 gilt:

$$s_{b_0} = s\sqrt{\frac{1}{K} + \frac{\overline{x}^2}{\sum(x_k - \overline{x})^2}} \tag{A17}$$

B. Multiple Regression

Die zu schätzende *multiple Regressionsfunktion* lautet

$$\hat{Y} = b_0 + b_1 X_1 + \ldots + b_j X_j \tag{B1}$$

und durch Einbeziehung der Residuen e erhält man:

$$Y = b_0 + b_1 X_1 + \ldots + b_j X_j + e \tag{B2}$$

Für die K Beobachtungen lässt sich dies in Matrix-Form wie folgt darstellen:

$$\begin{pmatrix} y_1 \\ \vdots \\ y_K \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \ldots & x_{J1} \\ \vdots & \vdots & & \vdots \\ 1 & x_{1K} & & x_{JK} \end{pmatrix} \cdot \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_J \end{pmatrix} + \begin{pmatrix} e_1 \\ \vdots \\ e_K \end{pmatrix} \tag{B3}$$

In Matrizenschreibweise erhalten wir hierfür den folgenden Ausdruck:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{e} \tag{B4}$$

mit

y	:	K-Vektor der Beobachtungswerte der abhängigen Variablen
X	:	(K x (J+1))-Matrix mit den Beobachtungswerte der Regressoren
b	:	(J+1)-Vektor der Regressionskoeffizienten plus konstantes Glied
e	:	K-Vektor der Residualgrößen

In dieser Darstellung betrachten wir das konstante Glied wie einen Koeffizienten und zwar für eine künstliche Variable X_0, die nur aus Einsen besteht. Diese haben wir als erste Spalte in die Datenmatrix **X** eingefügt.

Schätzung der Parameter

Das KQ-Kriterium lautet damit analog zu (A4) und (A5):

$$S = \mathbf{e}'\mathbf{e} = (\mathbf{y} - \mathbf{Xb})'(\mathbf{y} - \mathbf{Xb}) \qquad \rightarrow \quad \text{min!} \tag{B5}$$

Durch partielle Differentiation nach $b = [b_0, b_1, b_2, \ldots, b_J]'$ erhält man die *Schätzwerte der Regressionskoeffizienten* durch:

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{B6}$$

Die Matrix $\mathbf{X}'\mathbf{X}$ ist eine quadratische Matrix mit J+1 Zeilen bzw. Spalten. Ihre Größe ist also unabhängig von der Anzahl der Beobachtungen. Sie hat den gleichen Rang wie **X** und sie lässt sich nur invertieren, wenn sie vollständigen Rang J+1 besitzt. Daraus folgt, dass

- die Anzahl der Beobachtungen größer als die Anzahl der erklärenden Variablen sein muss (K > J),
- die erklärenden Variablen voneinander unabhängig sein müssen bzw. dass zwischen den Spalten von **X** keine linearen Abhängigkeiten bestehen dürfen.

Um bei Vorliegen von linearen Abhängigkeiten trotzdem eine Lösung erzielen zu können, wird z. B. in SPSS oder anderen Programmen zur Regressionsanalyse intern meist eine schrittweise Regressionsanalyse durchgeführt, bei welcher dann eventuell linear abhängige Variablen einfach übergangen werden.

Die Schätzwerte der abhängigen Variablen erhält man als Funktion unabhängigen Variablen durch:

$$\hat{\mathbf{y}} = \mathbf{Xb} \tag{B7}$$

Unter Verwendung von (B6) lassen sie sich die geschätzten y-Werte auch als Funktion der beobachteten y-Werte ausdrücken:

$$\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{Hy} \tag{B8}$$

Die Matrix **H** wird als „Hut-Matrix“ (hat matrix) bezeichnet, da sie **y** in $\hat{\mathbf{y}}$ transformiert und damit **y** den Hut aufsetzt. **H** ist eine quadratische KxK-Matrix. Sie besitzt wichtige diagnostische Bedeutung, da ihre Diagonalelemente zur Identifikation von Ausreißern (outliers) dienen können.[47]

Schätzfehler der Parameter

Die *Varianz-Kovarianz-Matrix* der Regressionskoeffizienten lautet:

$$V(\mathbf{b}) = s^2(\mathbf{X}'\mathbf{X})^{-1} \tag{B9}$$

wobei sich der Standardfehler s wie bei der einfachen Regression gemäß (A16) aus den quadrierten Residuen errechnet.

Bezeichnet man mit a_{jj} die Diagonalelemente der Inversen von $\mathbf{X}'\mathbf{X}$

$$a_{jj} = [(\mathbf{X}'\mathbf{X})^{-1}]_{jj} \qquad (\text{j=0, 1, ..., J}) \tag{B10}$$

dann erhält man die *Standardfehler der Regressionskoeffizienten* b_j durch

$$s_{b_j} = s\sqrt{a_{jj}} \qquad (\text{j=0, 1, ..., J}) \tag{B11}$$

Besser interpretierbar ist die folgende Formel für den Standardfehler der Regressionskoeffizienten:[48]

$$s_{b_j} = s\sqrt{\frac{1}{\sum(x_{jk} - \overline{x}_j)^2(1 - R_j^2)}} \tag{B12}$$

Ein Vergleich mit Formel (A17) für die einfache Regression macht deutlich, dass der Standardfehler eines Regressionskoeffizienten bei der multiplen Regression auch von der linearen Abhängigkeit zwischen X_j und den übrigen Regressoren abhängt. Für $R_j^2 \to 1$ geht der Standardfehler gegen ∞.

[47] Vgl. Belsley/Kuh/Welsch (1980), S. 16 ff.
[48] Vgl. z. B. Wooldridge (2016), S. 83 ff.; Greene (2018), S. 94; Fox (2008), S. 107

Die Formeln für die multiple Regression unter Anwendung von Matrizenrechnung lassen sich natürlich auch für eine einfache Regressionsanalyse anwenden. Die Matrix $\mathbf{X'X}$ reduziert sich in diesem Fall auf eine 2x2-Matrix.

Mit Hilfe obiger Formeln lässt sich eine Regressionsanalyse auch für größere Probleme z. B. mit MS Excel oder einem vergleichbaren Kalkulationsprogramm, welches Matrixfunktionen enthält, durchführen.

Literaturhinweise

A. Basisliteratur zur Regressionsanalyse

Auer, L. v. (2016), Ökonometrie: Eine Einführung, 7. Auflage, Berlin u. a.

Bleymüller, J./Weißbach, R. (2015), Statistik für Wirtschaftswissenschaftler, 17. Auflage, München.

Fahrmeir, L./ Kneib, T. / Lang, S. (2009), Regression – Modelle, Methoden und Anwendungen, Berlin/Heidelberg.

Fahrmeir, L./Heumann, C./Künstler, R./Pigeot, I./Tutz, G. (2016), Statistik – Der Weg zur Datenanalyse, 8. Auflage, Berlin u. a.

Greene, W. (2018), Econometric Analysis, 8. Auflage, Essex.

Hair, J./Black, W./Babin, B./Anderson, R. (2010), Multivariate Data Analysis, 7. Auflage, Englewood Cliffs (N.J.), Kapitel 4.

B. Zitierte Literatur

Auer, L. v. (2016), Ökonometrie: Eine Einführung, 7. Auflage, Berlin u. a.

Backhaus, K./Erichson, B./Weiber, R. (2015), Fortgeschrittene Multivariate Analyseverfahren, 3. Auflage, Berlin/Heidelberg.

Belsley, D./Kuh, E./Welsch, R. (1980), Regression Diagnostics, New York u. a.

Bleymüller, J./Weißbach, R. (2015), Statistik für Wirtschaftswissenschaftler, 17. Auflage, München.

Bortz, J./Schuster, C. (2010), Statistik für Human- und Sozialwissenschaftler, 7. Auflage, Berlin u. a.

Fahrmeir, L./Heumann, C./Künstler, R./Pigeot, I./Tutz, G. (2016), Statistik – der Weg zur Datenanalyse, 8. Auflage, Berlin/Heidelberg.

Fahrmeir, L./Kneib, T./Lang, S. (2009), Regression – Modelle, Methoden und Anwendungen, Berlin u. a.

Fox, J. (2008), Applied Regression Analysis and Generalized Linear Models, 3. Auflage, Los Angeles u. a.

Greene, W. (2018), Econometric Analysis, 8. Auflage, Essex.

Hammann, P./Erichson, B. (2000), Marktforschung, 4. Auflage, Stuttgart.

IBM (2004), SPSS 13.0 Command Syntax Reference, Chicago.

Kmenta, J. (1997), Elements of Econometrics, 2. Auflage, New York.

Kockläuner, G. (1988), Angewandte Regressionsanalyse mit SPSS, Braunschweig u. a.

Little, J. D. C. (1970), Models and Managers: The Concept of a Decision Calculus, in: *Management Science*, Vol. 16, Nr. 8, S. 466–485.

Maddala, G./Lahiri, K. (2009), Introduction to Econometrics, 4. Auflage, New York.

Ramanathan, R. (1998), Introductory Econometrics with Applications, 4. Auflage, Fort Worth.

Savin, N./White, K. (1977), The Durbin-Watson Test for Serial Correlation with Extreme Sample Size or many Regressors, in: *Econometrica*, Vol. 45, Nr. 8, S. 1989–1996.

Schneeweiß, H. (1990), Ökonometrie, 4. Auflage, Heidelberg.

Studenmund, A. (2001), Using Econometrics: A practical Guide, 4. Auflage, Boston (MA).

Weiber, R./Mühlhaus, D. (2014), Strukturgleichungsmodellierung, 2. Auflage, Berlin/Heidelberg.

Wonnacott, R./Wonnacott, T. (1979), Econometrics, 2. Auflage, New York.

Wonnacott, T./Wonnacott, R. (1981), Regression: A Second Course in Statistics, Malabar.

Wooldridge, J. (2016), Introductory Econometrics: A modern Approach, 6. Auflage, Cincinnati (OH).

2 Zeitreihenanalyse

K. Backhaus, B. Erichson, W. Plinke, R. Weiber, *Multivariate Analysemethoden*,
https://doi.org/10.1007/978-3-662-56655-8_3

2.1 Problemstellung

Bei der Datenanalyse wird generell zwischen der Analyse von Querschnittsdaten, die zu einem Zeitpunkt bei verschiedenen Untersuchungsobjekten erhoben wurden, und Zeitreihendaten (Längsschnittdaten), die zu verschiedenen Zeitpunkten erhoben wurden, unterschieden, wobei sich diese auch kombinieren lassen. Während die in diesem Buch behandelten Methoden primär der Querschnittsanalyse dienen, soll in diesem Kapitel gesondert auf die Zeitreihenanalyse eingegangen werden. Sie bildet ein wichtiges Anwendungsgebiet der im vorstehenden Kapitel behandelten Regressionsanalyse.

Die Zeitreihenanalyse dient neben der Beschreibung und Erklärung der zeitlichen Entwicklung einer Variablen Y insbesondere auch deren Prognose, d. h. der Schätzung von Werten dieser Variablen für zukünftige Zeitpunkte oder Perioden. Jede weitreichende Entscheidung basiert auf Prognosen. Die Zeitreihenanalyse ist daher für die Stützung von Entscheidungsproblemen jeglicher Art von großer Wichtigkeit. Abb. 2.1 zeigt beispielhaft einige typische Fragestellungen, die primär wirtschaftliche Probleme betreffen. Nicht alle diese Fragestellungen können hier behandelt werden. Vielmehr soll nur ein Einblick in Grundlagen der Zeitreihenanalyse mit Hilfe des Instrumentariums der Regressionsanalyse gegeben werden.

Prognose

Eine *Zeitreihe* ist formal eine Menge von Werten einer Variablen Y (z. B. Käuferzahl, Absatzmenge, Volkseinkommen, Bevölkerung, Wetter), die im Zeitablauf gemessen wurden und gemäß der Zeit geordnet sind. Sie lässt sich ausdrücken durch

Zeitreihe

$$\langle y_1, y_2, y_3, ... y_t, ... y_T \rangle \tag{2.1}$$

wobei der Index t (z. B. $t = 1, 2, 3, ..., T$) eine Abbildung der Zeit bildet. Nach dem Wertevorrat von t lässt sich zwischen stetigen und diskreten Zeitreihen unterscheiden. Hier sollen nur diskrete Zeitreihen betrachtet werden.[1] Der Index t bezeichnet bzw. zählt in diesem Fall äquidistante Zeitpunkte oder Perioden, in die die Zeit eingeteilt wird (z. B. Tage, Wochen, Monate). Durch $(1 : T)$ sei der *Beobachtungszeitraum* der Zeitreihe bezeichnet.

Zeitvariable

Die *Zeit* kann als eine Variable aufgefasst werden, die sich im Gegensatz zu anderen Variablen völlig gleichförmig und unabhängig von allem anderen Geschehen entwickelt.[2] Sie hat eine ordnende Funktion auf Ereignisse, indem sie diese in eine feste und unabänderliche Reihenfolge bringt. Bei Querschnittsdaten dagegen spielt die Reihenfolge der Daten keine Rolle und kann beliebig verändert werden.

Nach dem zeitlichen Bezug der Variablen Y unterscheidet man

- *zeitpunkt-bezogene Variablen*: Zustands- oder Bestandsgrößen, die punktuell gemessen werden, z. B. an einem Stichtag (Lagerbestand am Monatsende, Finanzvermögen am Jahresende),
- *zeitraum-bezogene Variablen*: Strömungsgrößen, die über einen Zeitraum gemessen werden (Verbrauch pro Tag, Umsatz pro Monat, Einkommen pro Jahr).

[1] Stetige Zeitreihen werden z. B. durch automatische Aufzeichnung physikalischer oder meteorologischer Prozesse (z. B. Temperatur, Luftdruck) erzeugt.

[2] Der klassische Zeitbegriff wurde durch Isaak Newton (1643–1727) geprägt: „Absolute, true, and mathematical time, from its own nature, passes equably without relation to anything external“. Seit Albert Einstein (1879–1955) wissen wir zwar, dass auch die Zeit nicht unabhängig und konstant verläuft, sondern sich mit zunehmender Geschwindigkeit verlangsamt und bei Lichtgeschwindigkeit sogar zum Stillstand kommt. Für die uns hier interessierenden Phänomene aber können wir dies vernachlässigen.

- Ist das Volumen des Margarinemarktes über die Zeit konstant (abgesehen von zufälligen Schwankungen) oder besteht ein Trend (Wachstum oder Schrumpfung)?[3]
- Gibt es im Konsum von Margarine saisonale Schwankungen?[4]
- Hat das Wetter (z. B. die Temperatur) einen Einfluss auf den Margarinekonsum?[5]
- Wurde durch die Änderung der Verpackung eines Produktes ein Strukturbruch in der Entwicklung des Abatzvolumens bewirkt?[6]
- Lässt sich aus den Absatzzahlen eines neuen Produktes über die ersten 12 Monate nach der Markteinführung erkennen, welchen Marktanteil das Produkt im zweiten oder dritten Jahr erreichen wird und ob es am Markt erfolgreich sein wird?[7]
- Ein Automobilhersteller möchte herausfinden, ob der GfK-Konsumklima-Index oder der Ifo-Geschäftsklima-Index geeignete Frühindikatoren zur Prognose seines Absatzvolumens sind.[8]
- Ein Hersteller der Kosmetik-Branche möchte herausfinden, ob seine Produktwerbung zeitlich verzögert wirkt, und wenn ja, mit welchem Time-Lag.[9]
- Anfang der 90er Jahre begann die Ausbreitung von zwei Innovationen, die seitdem unser Leben erheblich verändert haben: Das Internet (Einführung des WWW) und das Mobiltelefon (Handy). Wie hat sich die Nutzung seitdem in Deutschland entwickelt und welche zukünftige Entwicklung ist zu erwarten?[10]

Abbildung 2.1: Typische Fragestellungen der Zeitreihenanalyse

Ein Problem bei Zeitreihenanalysen resultiert daraus, dass viele der üblicherweise verwendeten Perioden, wie Monate, Quartale oder Jahre, nicht wirklich äquidistant sind und daher eine Bereinigung der Zeitreihe notwendig machen können.[11] Die Länge eines Monats variiert zwischen 28 und 31 Tagen, d. h. die Länge des Februars in einem Normaljahr beträgt nur 90,32 % der Länge eines Monats mit 31 Tagen (z. B. Juli oder August). Entsprechend variieren die Werte von monatlich oder auch quartalsmäßig gemessenen Strömungsgrößen bei konstanter Strömung (z. B. Energieerzeugung, Fördermenge von Öl). Für ökonomische Analysen ist neben der Zahl der Kalender-

Äquidistanz

[3] Siehe dazu das Fallbeispiel in diesem Kapitel.
[4] Siehe dazu das Fallbeispiel in diesem Kapitel.
[5] Siehe dazu das Fallbeispiel in diesem Kapitel.
[6] z. B. Schneeberger (1994).
[7] z. B. Mahajan/Muller/Bass (1990), P. Mertens and S. Rässler (2012).
[8] z. B. Backhaus/Simon (1981), Niederhübner (1994).
[9] z. B. Lambin (1969), Palda (1984).
[10] Siehe dazu Kapitel 10.
[11] Siehe hierzu Rinne/Specht (2002), S. 33 ff.

tage auch die Zahl der Werk- oder Arbeitstage relevant, die wiederum durch gesetzliche Feiertage eingeschränkt werden. Hierdurch können sich auch Schwankungen bei wöchentlich gemessenen Werten ergeben. Von geringerer Bedeutung ist dagegen die unterschiedliche Länge zwischen Normaljahren und Schaltjahren.

Quantitative Prognoseverfahren

Qualitative Prognoseverfahren

Prognoseverfahren, die sich auf Zeitreihenanalysen stützen, werden als *quantitative Prognoseverfahren* bezeichnet.[12] Sie beinhalten im Kern ein mathematisches Modell des Prozesses, der die Zeitreihe erzeugt hat. Im Unterschied dazu bezeichnet man Prognoseverfahren, die sich nicht auf ein mathematisches Modell stützen, als *qualitative Prognoseverfahren*. Sie basieren primär auf den subjektiven Einschätzungen von Experten (z. B. Delphi-Methode, Szenario-Technik).[13]

Nach der Anzahl der Zeitreihen, auf die sich das Prognoseverfahren stützt, lassen sich die quantitativen Prognoseverfahren in zwei Gruppen einteilen.

Quantitative Prognoseverfahren

a) Zeitreihenextrapolation

Entwicklungsprognose

Die Zeitreihenextrapolation basiert auf der Analyse einer einzelnen Zeitreihe. Es wird nach einem Muster in der Entwicklung der Zeitreihendaten gesucht und durch Extrapolation dieses Musters in die Zukunft werden Prognosen abgeleitet (sog. *Entwicklungsprognosen*). Dabei wird die Annahme gemacht, dass das Muster stabil bleibt. Die hierbei verwendeten Prognoseverfahren werden auch als *Extrapolationsverfahren* bezeichnet. Sie lassen sich grob in drei Gruppen einteilen:

- Zeitregression
- Glättungsmethoden (z. B. gleitende Durchschnitte, Exponentielle Glättung)
- Autoregressive Methoden (z. B. ARIMA-Prozesse nach Box/Jenkins, adaptive Filter)[14]

Hier soll nur die Zeitregression behandelt werden. Im Kern beinhaltet sie ein deterministisches Modell, welches von der Zeit abhängig ist und das auf Basis der Zeitreihendaten zu schätzen ist. Zwecks Erstellung von Prognosen sind entsprechende Zeitwerte in das geschätzte Modell einzusetzen. Die Modelle lassen sich daher relativ einfach handhaben und sind auch dem Verständnis des Benutzers besser zugänglich, als viele andere Verfahren, bei denen die Modellierung der stochastischen Komponente einer Zeitreihe im Vordergrund steht. Weitere Vorteile sind, dass die Zeitregression auch angewendet werden kann, wenn nur verhältnismäßig wenige Zeitreihendaten vorliegen, und dass sie sowohl für kurzfristige wie auch mittel- und langfristige Prognosen verwendet werden kann.

[12] Zu Zeitreihenanalysen und Prognoseverfahren vgl. z. B. Hamilton (1994), Hanke/Reitsch (2009), Makridakis/Wheelwright/Hyndman (1998), P. Mertens and S. Rässler (2012), Rinne/Specht (2002).

[13] Einen Überblick über quantitative wie qualitative Prognoseverfahren geben z. B Armstrong, J. (2002), Hanke/Reitsch (2009), Makridakis/Wheelwright/Hyndman (1998).

[14] SPSS bietet unter dem Menüpunkt „Analysieren/Zeitreihen“ Prozeduren zur Schätzung von Modellen der exponentiellen Glättung und von ARIMA-Modellen (Autoregressive Integrated Moving Average Models) an, auf die hier nicht eingegangen wird.

b) Kausale Prognoseverfahren

Kausale Prognoseverfahren basieren auf der Analyse der Beziehungen zwischen mehreren Zeitreihen. Es wird dazu ein Modell spezifiziert, welches die kausalen Beziehungen zwischen den Zeitreihenvariablen abbildet und welches sodann auf Basis der Zeitreihendaten zu schätzen ist.

Hier wird von *Strukturmodellen* oder *ökonometrischen Modellen* gesprochen.[15] Man unterscheidet:

- Eingleichungsmodelle
- Mehrgleichungsmodelle

In Eingleichungsmodellen wird die Abhängigkeit der Prognosevariablen Y von einer oder mehreren Prädiktorvariablen modelliert (z. B. die Abhängigkeit der Pkw-Nachfrage vom Preis, den Werbeausgaben, dem verfügbaren Einkommen der Bevölkerung und dem Konsumklima). Im Gegensatz zu Entwicklungsprognosen spricht man hier auch von *Wirkungsprognosen*. Soweit es sich bei den Prädiktorvariablen um nicht kontrollierte Variablen handelt (wie hier dem verfügbaren Einkommen und dem Konsumklima), erfordert die Prognose von Y, dass zunächst Prognosen der Prädiktorvariablen zu erstellen sind, was die Anwendung erschwert. Dies kann u.U. erleichtert werden, wenn die Entwicklung der Prädiktorvariablen gegenüber der abhängigen Variable zeitlich voraus läuft. So ist z. B. der GfK-Konsumklima-Index ein *vorauseilender Indikator* für die Pkw-Nachfrage oder die Zahl der Neuzulassungen von Pkws für die Nachfrage nach Autozubehör.[16]

Wirkungsprognosen

Mehrgleichungsmodelle werden erforderlich, wenn wechselseitige Abhängigkeiten zwischen den Variablen bestehen und modelliert werden sollen. So bestehen z. B. Wechselwirkungen zwischen dem Volkseinkommen und privatem Konsum oder zwischen der Absatzmenge eines Produktes und seinem Bekanntheitsgrad.

Wir wollen uns hier auf Eingleichungsmodelle beschränken. In Eingleichungsmodellen lassen sich Entwicklungs- und Wirkungsprognosen auch kombinieren. Zur Schätzung der Modelle kann die im vorstehenden Kapitel behandelte Regressionsanalyse verwendet werden.

Anwendungsbeispiel

Anwendungsbeispiel

Die Analyse einer einzelnen Zeitreihe soll zunächst an einem kleinen Beispiel demonstriert werden. Der Manager des hier betrachteten Margarineherstellers ist etwas überrascht, aber auch erfreut über den hohen Absatz im Verkaufsgebiet 1 (vgl. Abbildung 1.6 in Kapitel 1). Falls das hohe Niveau anhalten oder sich gar weiter erhöhen sollte, müsste er allerdings eine Änderung hinsichtlich der Belieferung dieses Gebietes vornehmen. Für seine Entscheidungsfindung benötigt er eine Analyse und Prognose der Absatzentwicklung in diesem Verkaufsgebiet. Zu diesem Zweck besorgt er sich die Absatzdaten der letzten 10 Jahre, die in Abbildung 2.2 wiedergegeben sind.

[15] Vgl. hierzu die ökonometrische Literatur, z. B. Greene (2018), Hanssens/Parsons/Schultz (2003), Kmenta (1997), Ramanathan (2002), Schneeweiß (1990), Studenmund (2017), Wooldridge (2016).
[16] Vgl. z. B. Backhaus/Simon (1981), Niederhübner (1994).

Zeit (Periode t)	Menge (Kartons pro Periode)
1	1.657
2	1.864
3	1.950
4	2.204
5	2.288
6	2.410
7	2.414
8	2.534
9	2.739
10	2.785

Abbildung 2.2: Zeitreihe der Absatzmenge eines Verkaufsgebietes

2.2 Vorgehensweise

Die allgemeine Vorgehensweise bei der Analyse und Prognose einer Zeitreihe lässt sich in fünf Schritte gliedern (vgl. Abbildung 2.3).

Abbildung 2.3: Ablaufschritte der Analyse und Prognose einer Zeitreihe

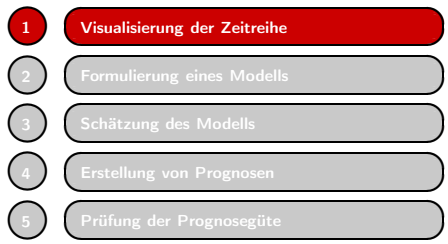
Am Beginn der Analyse sollte immer die Visualisierung der Zeitreihendaten stehen, die unerlässlich für die Formulierung eines geeigneten Modells ist. Es ist wichtig, dass sich der Untersucher vom Verlauf der Zeitreihe ein „Bild“ macht, was bei Betrachtung einer langen Zahlenkolonne kaum möglich ist.

Eine Visualisierung der zu analysierenden Daten ist zwar für jede Art der Datenanalyse von Vorteil, doch stößt man gerade bei der multivariaten Datenanalyse schnell an Grenzen, da sich maximal drei Dimensionen darstellen lassen. Daher ist es ein Ziel vieler multivariater Analysemethoden, durch Analyse und Komprimierung der Daten eine Visualisierung zu ermöglichen. Bei Zeitreihendaten dagegen bietet sich immer eine Visualisierung über die Zeit an.

Die Formulierung des Modells sollte sich auf das „Bild“ stützen, welches sich der Untersucher von der Zeitreihe gemacht hat, aber auch auf weitere sachlogische oder theoretische Überlegungen. Es existiert ein umfangreiches Arsenal von Modelltypen, auf welches man bei der Modellformulierung zurückgreifen kann. Oft ist es zweckmäßig, alternative Modelle zu erproben oder miteinander zu kombinieren.

Im Zentrum des Ablaufschemas steht die Durchführung einer Regressionsanalyse, bezüglich derer wir uns hier auf das vorstehende Kapitel stützen können. Hieran schließt sich die Verwendung des geschätzten Modells zur Erstellung von Prognosen an. Durch Gegenüberstellung dieser Prognosen mit der Realität erfolgt schließlich die Prüfung der Prognosegüte des Modells bzw. die Modellvalidierung.

2.2.1 Visualisierung der Zeitreihe



Visualisierung von Zeitreihendaten

Für die Visualisierung von Zeitreihendaten bieten Programme wie Excel oder SPSS vielfältige Funktionen an (Streudiagramme, Liniendiagramme, Balkendiagramme). Abbildung 2.4 zeigt die Absatzdaten aus Abbildung 2.2 als Streudiagramm. Häufig werden die Streupunkte durch Linien verbunden, sodass sich ein Polygonzug wie in Abbildung 2.18 ergibt. Die Darstellung einer Zeitreihe als Balkendiagramm zeigt z. B. Abbildung 2.19.

Das Streudiagramm zeigt einen deutlich ansteigenden Verlauf der Absatzmenge über die Zeit. Es lässt aber auch erkennen, dass der Verlauf nicht ganz linear, sondern leicht konkav gebogen ist. Das Wachstum der Absatzmenge scheint sich im Verlauf der Zeit etwas abzuschwächen.

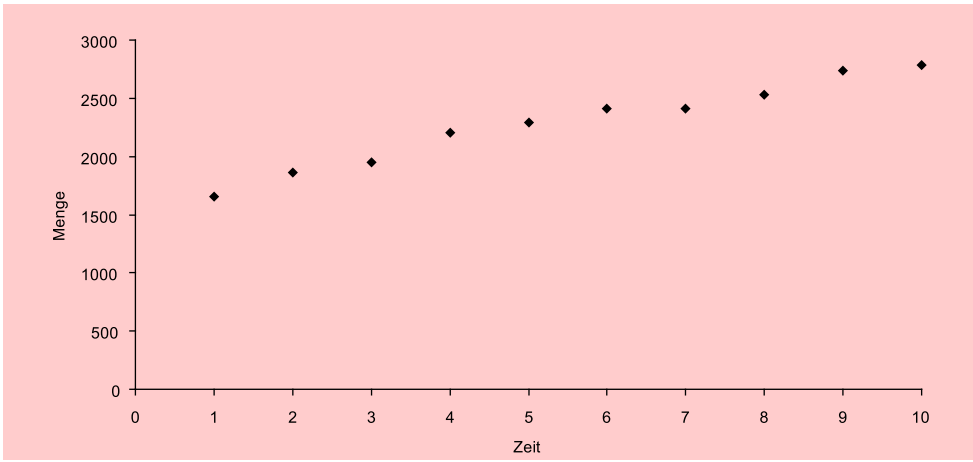
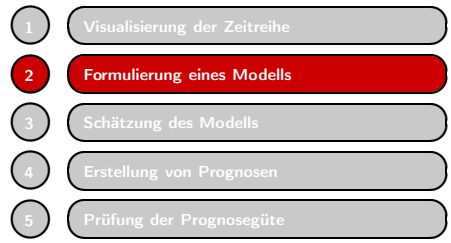


Abbildung 2.4: Streudiagramm der Zeitreihendaten

2.2.2 Formulierung eines Modells



Zeitreihenverläufe können höchst unterschiedliche Formen annehmen und es existiert eine zahllose Menge von Modelltypen, die sich an den Verlauf einer Zeitreihe mehr oder weniger gut anpassen lassen. Die Wahl des „richtigen“ Modells bildet sicherlich die wichtigste, aber auch schwierigste Entscheidung bei der Anwendung quantitativer Prognoseverfahren.

Zeitreihenzerlegung

Ein Grundprinzip der Modellierung von Zeitreihen bildet die *Zeitreihenzerlegung*, d. h. die Zerlegung einer Zeitreihe in unterschiedliche Komponenten. Man unterscheidet:

$$\text{Additive Zeitreihenzerlegung:} \qquad Y = A + K + S + u \tag{2.2}$$

$$\text{Multiplikative Zeitreihenzerlegung:} \qquad Y = A \cdot K \cdot S \cdot u \tag{2.3}$$

mit

Y = Prognosevariable mit den Werten $\langle y_1, y_2, y_3, ...y_t, ...y_T \rangle$
A = Trendkomponente
K = Konjunkturkomponente
S = Saisonkomponente
u = zufällige Komponente (Störgröße)

Die Trendkomponente A repräsentiert die langfristige Entwicklung der Größe Y. Sie kann positiv (Wachstum) oder negativ (Schrumpfung), linear oder nichtlinear sein. K und S sind zyklische Schwankungen, die im Zeitraum von mehreren Jahren (Konjunktur) oder einem Jahr (Saison) periodisch wiederkehren. Sie verlaufen zwangsläufig nichtlinear. A, K und S werden gemeinsam auch als *systematische Komponente* bezeichnet, im Gegensatz zur zufälligen Komponente u.

Systematische Komponente

Die vorliegende Zeitreihe lässt keine zyklische Komponente erkennen. Modell (2.2) reduziert sich damit auf

$$Y = A + u \tag{2.4}$$

Die Trendkomponente A muss nun spezifiziert werden. Wir wollen mit einem einfachen Modell beginnen und nehmen einen linearen Trendverlauf an, obgleich wir einen leicht nichtlinearen Verlauf der Zeitreihe erkannt haben. Wir erhalten damit das *lineare Trendmodell*:

Lineares Trendmodell

$$Y = \alpha + \beta \cdot t + u \tag{2.5}$$

α und β sind unbekannte Parameter, die auf Basis der vorliegenden Daten zu schätzen sind. Der *Trendparameter* β gibt den Zuwachs von Y pro Periode an. Im linearen Trendmodell wird dieser Zuwachs als konstant angenommen.

2.2.3 Schätzung des Modells

Regressionsanalyse

Zur Schätzung der unbekannten Parameter α und β auf Basis der vorliegenden Daten verwenden wir die Regressionsanalyse. Dazu ist der Zeitindex mit den Werten $t = 1, 2, 3, ...$ als unabhängige Variable zu verwenden. Anstelle der Indexzahlen 1, 2, 3, ... könnten hier auch Jahreszahlen (z. B. 1991, 1992, 1993, ...) stehen, was auf den Schätzwert für den Trendparameter β keinen Einfluss hat (es würde sich nur der Schätzwert für das konstante Glied α verändern). Die Durchführung der Regressionsanalyse liefert die folgende Regressionsfunktion:

$$\hat{Y} = a + b \cdot t = 1.619,5 + 120,9 \cdot t \qquad (R^2 = 0,972)$$

Der Schätzwert $b = 120,9$ besagt hier, dass die Absatzmenge im Verkaufsgebiet pro Periode um rund 121 Kartons anwächst. Die geschätzte Funktion ist in Abbildung 2.5 zusammen mit den beobachteten Werten dargestellt. Die beobachteten Werte weichen nur geringfügig von der geschätzten Funktion ab.

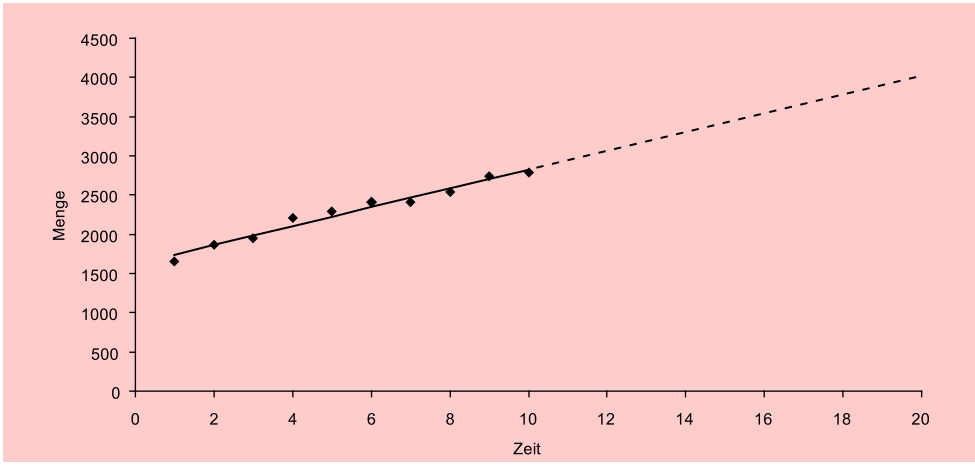


Abbildung 2.5: Lineares Trendmodell: Geschätzte Funktion und prognostizierter Verlauf

Güte der Schätzung

Zur Beurteilung der Güte der geschätzten Regressionsfunktion liefert die Regressionsanalyse u.a. die folgenden Statistiken (vgl. dazu die Ausführungen in Kapitel 1):

Bestimmtheitsmaß (R-Quadrat):	$R^2 =$	$0,972$
F-Statistik:	$F_{emp} =$	$276,8$
Standardfehler der Regression:	$s =$	$66,0$
Durbin/Watson-Statistik:	$d =$	$1,551$

Das Bestimmtheitsmaß sagt uns, dass die Zeitvariable t die Streuung in den Absatzdaten zu rund 97 %, also fast vollständig, erklärt. Die Änderung über die Zeit ist konstant, d.h. z.B. Wachstum der Marketingaktivität aufgrund von Inflation etc. Der Wert der F-Statistik weist auf hohe Signifikanz des Modells hin (der tabellarische F-Wert beträgt hier lediglich 5,32 bei einer Vertrauenswahrscheinlichkeit von 95 %). Der Standardfehler besagt, dass die mittlere Abweichung (Standardabeichung) zwischen der geschätzten Regressionsfunktion und den beobachteten Absatzmengen 66 Kartons beträgt. Der Wert der Durbin/Watson-Statistik, der im Idealfall d=2 beträgt, weist hier nicht auf das Vorliegen von Autokorrelation hin.

2.2.4 Erstellung von Prognosen

Prognose

Mit Hilfe der geschätzten Funktion (2.6) ist es möglich, ohne weitere Informationen Prognosen zu erstellen. Der Prognosewert für die in der Zukunft liegende Periode T+k ergibt sich durch Funktion (2.7).

Punktprognose

Man spricht in diesem Fall von einer *Punktprognose*. Im Unterschied dazu gibt eine *Intervallprognose* einen Bereich um den Prognosewert $\hat{y}_{T+k}$ an, in dem der zukünftige wahre Wert mit einer bestimmten Vertrauenswahrscheinlichkeit bzw. Konfidenz (z. B. 95 %) liegen wird.

$$\hat{Y} = a + b \cdot t \tag{2.6}$$

$$\hat{y}_{T+k} = a + b \cdot (T + k) \tag{2.7}$$

2.2.4.1 Punktprognose

Mittels (2.7) erhält man für die folgende Periode T+1 = 11 den Prognosewert

$$\hat{y}_{11} = 1.619,5 + 120,9 \cdot 11 = 2.949$$

Analog lassen sich Prognosen für beliebige Perioden erstellen. Für Periode 20 erhält man z. B.

$$\hat{y}_{20} = 1.619,5 + 120,9 \cdot 20 = 4.038$$

Stützbereich

In Abbildung 2.5 ist zusätzlich zur geschätzten Funktion im Beobachtungsintervall (1 : 10), der auch als *Stützbereich* des Prognosemodells bezeichnet wird, der prognostizierte Verlauf der Absatzmenge für die Perioden 11 bis 20 als gestrichelte Linie dargestellt.

2.2.4.2 Prognosefehler

Prognosen sind leider immer mit Fehlern verbunden („besonders wenn sie auf die Zukunft gerichtet sind“[17]). Die Basis für die Berechnung des Prognosefehlers bildet der Standardfehler der Regressionsschätzung (vgl. Abschnitt 1.2.3.3 in Kapitel 1), der sich aus den Residuen $e_t = y_t - \hat{y}_t$ ergibt. Der *Standardfehler* errechnet sich durch: Standardfehler

$$s = \sqrt{\frac{1}{T-2}\sum_{t=1}^{T} e_t^2} = 66,0 \tag{2.8}$$

Es ist einsichtig, dass der Prognosefehler um so größer sein wird, je weiter die Prognose in die Zukunft reicht. Um dies zu berücksichtigen, ist der Standardfehler der Schätzung mit einem Faktor > 1 zu multiplizieren, dessen Größe vom Prognosehorizont T+k abhängig ist. Für den Fehler einer Prognose für Periode T+k gilt:[18]

Prognosefehler Prognosefehler

$$s_p(T+k) = s \cdot \sqrt{1 + \frac{1}{T} + \frac{(T+k-\overline{t})^2}{\sum_t (t-\overline{t})^2}} = s \cdot \sqrt{1 + \frac{1}{T} + \frac{(T+k-\overline{t})^2}{(T-1)\cdot s_t^2}} \tag{2.9}$$

mit

s = Standardfehler der Regression
$\overline{t}$ = Mittelwert der Zeitvariablen t
s_t= Standardabweichung der Zeitvariablen t

Mit den Werten $s = 66$, $\overline{t} = 5,5$ und $s_t = 3,028$ ergibt sich hier für den Prognosewert $\hat{y}_{11} = 2.949,4$ der Prognosefehler[19]

$$s_p(11) = 66 \cdot \sqrt{1 + \frac{1}{10} + \frac{(10+1-5,5)^2}{9 \cdot 3,028^2}} = 66 \cdot \sqrt{1 + 0,1 + 0,37} = 66 \cdot 1,21 = 80$$

Interpretation

Interpretation: Gemäß dem hier zugrunde gelegten Modell (2.5) bildet die Absatzmenge Y eine Zufallsgröße, deren Verteilung durch die Störgröße u bedingt ist. Die Abweichungen zwischen dem prognostizierten Wert $\hat{y}_{11} = 2.949,4$ und dem realisierten Wert y_{11} würde, falls wiederholte Realisierungen möglich wären, im Durchschnitt 80 Kartons betragen.

Für Periode 20 hatten wir den Wert $\hat{y}_{20} = 4.037,5$ prognostiziert. Analog ergibt sich hier der Prognosefehler

$$s_p(20) = 66 \cdot \sqrt{1 + \frac{1}{10} + \frac{(10+10-5,5)^2}{9 \cdot 3,028^2}} = 66 \cdot \sqrt{1 + 0,1 + 2,55} = 66 \cdot 1,91 = 126$$

Da die Prognose viel weiter in die Zukunft reicht, ist auch der zugehörige Prognosefehler bedeutend größer. Aus (2.9) ist ersichtlich, dass der Prognosefehler mit dem Prognosehorizont $T + k$ anwächst. Am geringsten ist der Prognosefehler im Mittelpunkt der Zeitreihe bzw. des Stützbereiches, hier im Punkt $\overline{t} = 5,5$. Prognosehorizont

[17] Die Bemerkung stammt vermutlich von Hermann Josef Abs (1901–1994), dem legendären Bankier und langjährigen Vorstandsvorsitzenden der Deutschen Bank.
[18] Vgl. z. B. Kmenta (1997), S. 251.
[19] Die Werte von $\overline{t}$ und s_t lassen sich in SPSS mittels „Analysieren / Deskriptive Statistiken“ berechnen.

2.2.4.3 Prognoseintervall

Prognoseintervall

Um anzugeben, in welchem Bereich der noch unbekannte zukünftige Wert y_{11} mit einer bestimmten Wahrscheinlichkeit liegen wird, ist ein Prognoseintervall (Konfidenzintervall) zu erstellen:

$$y_{T+k} = \hat{y}_{T+k} \pm t_{\alpha/2} \cdot s_p(T+k) \tag{2.10}$$

bzw.

$$\hat{y}_{T+k} - t_{\alpha/2} \cdot s_p(T+k) \leq y_{T+k} \leq \hat{y}_{T+k} + t_{\alpha/2} \cdot s_p(T+k)$$

Dabei bezeichnet $t_{\alpha/2}$ das Quantil der t-Verteilung (Student-Verteilung) für die Vertrauenswahrscheinlichkeit $1-\alpha$ bei zweiseitigem Test und T-2 Freiheitsgraden.

Für die Vertrauenswahrscheinlichkeit $1-\alpha = 0,95$ und T$-2 = 8$ Freiheitsgrade lässt sich aus der t-Tabelle (Anhang A.1) der Wert $t_{\alpha/2} = 2,306$ entnehmen. Für die unbekannte Absatzmenge y_{11} ergibt sich damit das *Prognoseintervall*:

$$\begin{aligned} \hat{y}_{11} - t_{\alpha/2} \cdot s_p(11) &\leq y_{11} \leq \hat{y}_{11} + t_{\alpha/2} \cdot s_p(11) \\ 2.949 - 2,3 \cdot 80 &\leq y_{11} \leq 2.949 + 2,3 \cdot 80 \\ 2.765 &\leq y_{11} \leq 3.133 \end{aligned}$$

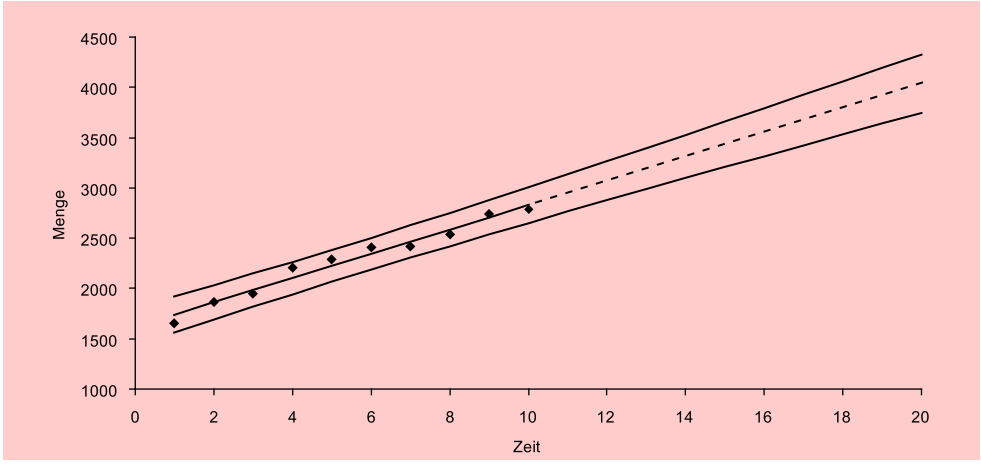


Abbildung 2.6: Lineares Trendmodell: Geschätzte Funktion, prognostizierter Verlauf und 95 %-Prognoseintervall

Interpretation

Interpretation: Mit 95 %iger Sicherheit ist zu erwarten, dass die Absatzmenge in der kommenden Periode zwischen 2.765 und 3.133 Kartons liegen wird. Ließe sich die Absatzmenge 100 Mal realisieren, so würde sie in 95 Fällen in diesen Bereich fallen.

Für Periode 20 hatten wir den Wert $\hat{y}_{20} = 4.037,5$ prognostiziert mit dem Prognosefehler $s_p(20) = 126$. Für das Prognoseintervall erhält man:

$$3.727 \leq y_{20} \leq 4.328$$

Abbildung 2.6 zeigt den Verlauf des Prognoseintervalls über die Zeit. Proportional zum Prognosefehler wächst auch das Prognoseintervall mit k an. Am geringsten ist es, ebenso wie der Prognosefehler, im Mittelpunkt der Zeitreihe bzw. des Stützbereiches.

2.2.5 Prüfung der Prognosegüte

Die üblichen Gütemaße der Regressionsanalyse (Bestimmtheitsmaß, F-Statistik, Standardfehler) sagen nur etwas darüber aus, wie gut die Anpassung des geschätzten Modells an die beobachteten Werte im Stützzeitraum (1:T) ist. Hierauf basieren auch die Berechnung von Prognosefehler und Prognoseintervall. Zur Prüfung der Prognosegüte und damit der Validierung des Modells ist es dagegen erforderlich, die prognostizierten Werte $\hat{y}_{T+k}$ mit den realisierten Werten außerhalb des Stützbereiches im *Prognosebereich* (T+1:T+K) zu vergleichen.

Validierung

Da die realisierten Werte im Zeitpunkt der Prognose nicht bekannt sind (sonst müsste man sie ja nicht prognostizieren), ist daher eine sofortige Prüfung der Prognosegüte nicht möglich. Vielmehr muss man warten, bis diese anfallen, was z. B. bei Jahresdaten recht lange dauern kann.

Eine Alternative besteht darin, die Prognosegüte durch *Ex-Post-Prognosen* zu prüfen. Dazu wird der Stützbereich für die Schätzung des Modells gegenüber dem Beobachtungszeitraum verkürzt, d. h. man verzichtet bei der Schätzung des Modells auf die letzten Zeitreihenwerte und vergleicht diese sodann mit den ex post prognostizierten Werten.

Ex-Post-Prognose

Verkürzen wir hier den Stützbereich von zehn auf acht Perioden, so erhalten wir anstelle der Schätzung

$$\hat{Y} = 1.619,5 + 120,9 \cdot t \qquad (R^2 = 0,972)$$

die folgende Funktion:

$$\hat{Y} = 1.610,5 + 123,3 \cdot t \qquad (R^2 = 0,953)$$

Damit können wir ex post die Absatzmengen für die Perioden 9 und 10 prognostizieren. Die so prognostizierten Werte sind nachfolgend den realisierten Werten gegenübergestellt:

t	$\hat{y}_t$	y_t	$e_t = \hat{y}_t - y_t$
9	2.720	2.739	-19
10	2.843	2.785	58

Zur Beurteilung der *empirischen Prognosegüte* können die folgenden Fehlermaße berechnet werden:[20]

Mittlerer absoluter Fehler (Mean Absolute Deviation)

MAD

$$MAD = \frac{1}{K}\sum_{k=1}^{K}|e_{T+k}| \qquad = \frac{19+58}{2} = 38,5 \qquad (2.11)$$

[20] Vgl. z. B. Hüttner (1994), S. 350 f., Rinne/Specht (2002), S. 134 f.

MAPE

Mittlerer absoluter prozentualer Fehler (Mean Absolute Percentage Error)

$$MAPE = \frac{1}{K}\sum_{k=1}^{K}\left|\frac{e_{T+k}}{y_{T+k}}\right| \cdot 100 \qquad = \frac{0,0069 + 0,0208}{2} \cdot 100 = 1,4\,\% \tag{2.12}$$

mit

e_t = Prognoseabweichung (einfacher Prognosefehler), $e_t = \hat{y}_t - y_t$
T = Ende des Stützbereichs, hier $T = 8$
K = Prognosehorizont (Prognosedistanz), hier: $K = 2$

Der MAPE besitzt gegenüber dem MAD sowie weiteren gebräuchlichen Fehlermaßen den Vorteil, dass sich damit auch die Prognosegüte zwischen unterschiedlichen Zeitreihen vergleichen lässt, da er unabhängig von der Maßeinheit der Zeitreihen ist.

Theil's U

Ein weiteres interessantes Fehlermaß bildet *Theil's U-Statistik*:[21]

$$U = \sqrt{\frac{\sum_{k=1}^{K}(\hat{y}_{T+k} - y_{T+k})^2}{\sum_{k=1}^{K}(y_{T+k} - y_{T+k-1})^2}} \tag{2.13}$$

mit

$$U = \begin{cases} < 1 & : \text{besser als naive Prognose} \\ = 1 & : \text{naive Prognose} \\ > 1 & : \text{schlechter als naive Prognose} \end{cases}$$

Die U-Statistik vergleicht die Prognosegüte eines geschätzten Modells mit der naiven Prognose, die darin besteht, den jeweils letzten Beobachtungswert als Prognose der folgenden Periode zu verwenden. Im Zähler von (2.13) steht der quadrierte Fehler der Prognose für Periode T+k und im Nenner der entsprechende quadrierte Fehler der naiven Prognose. Man erhält hier:

$$U = \sqrt{\frac{(\hat{y}_9 - y_9)^2 + (\hat{y}_{10} - y_{10})^2}{(y_9 - y_8)^2 + (y_{10} - y_9)^2}} = \sqrt{\frac{(2.720 - 2.739)^2 + (2.843 - 2.785)^2}{(2.739 - 2.534)^2 + (2.785 - 2.739)^2}}$$
$$= \sqrt{\frac{(-19)^2 + (58)^2}{(205)^2 + (46)^2}} = \sqrt{\frac{3.725}{44.141}} = 0,29$$

Der erhaltene Wert besagt, dass das lineare Trendmodell die Absatzmenge deutlich besser prognostiziert als die naive Vorgehensweise.

2.3 Nichtlineare Zeitreihenmodelle

2.3.1 Nichtlineare Trendmodelle

Reale Trendverläufe sind nur selten linear. Das lineare Trendmodell bildet daher nur eine Approximation des wahren Verlaufs, die aber meist ausreichend ist, wenn es um kurzfristige Prognosen geht. Für langfristige Prognosen wird dagegen meist ein nichtlineares Trendmodell geeigneter sein.

[21] Vgl. z. B. Rinne/Specht (2002), S. 137.

2.3.1.1 Das Quadratwurzel-Modell

Wie oben schon bemerkt wurde, lässt das Streudiagramm der Zeitreihe erkennen, dass sich das Wachstum der Absatzmenge im Verlauf der Zeit etwas abschwächt, d. h. der Verlauf ist nicht ganz linear, sondern leicht konkav gebogen. Dies gilt für zahlreiche ökonomische Zeitreihen oder andere Prozesse (auch Bäume wachsen bekanntlich nicht in den Himmel). Ein einfaches und häufig verwendetes nichtlineares Trendmodell, das ein sich abflachendes Wachstum bzw. einen degressiv ansteigenden Verlauf aufweist, ist das Quadratwurzel-Modell, dessen Verlaufsform in Abbildung 2.7 dargestellt ist.

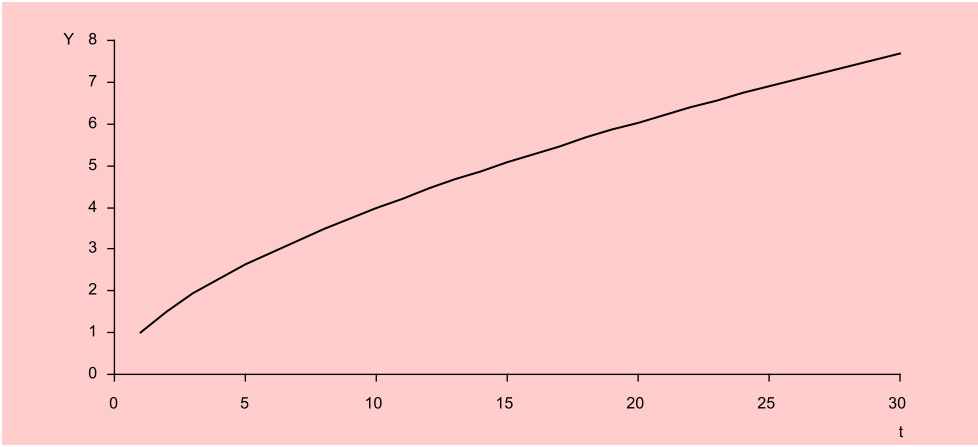


Abbildung 2.7: Quadratwurzel-Funktion $Y = \sqrt{t}$

Das *Quadratwurzel-Modell* lautet:

$$Y = \alpha + \beta \cdot \sqrt{t} + u \tag{2.14}$$

Das Modell ist zwar nichtlinear in Bezug auf die Zeitvariable t, aber linear in Bezug auf die zu schätzenden Parameter. Es handelt sich um ein sog. *intrinsisch lineares Modell*, das sich mit Hilfe der linearen Regressionsanalyse schätzen lässt. Zur Schätzung des Quadratwurzel-Modells kreieren wir eine neue Variable $X = \sqrt{t}$, deren Werte in Abbildung 2.8 aufgelistet sind.

Intrinsisch linear

Die Regression der Menge auf die transformierte Zeitvariable $X = \sqrt{t}$ liefert:

$$\hat{Y} = 1.117,1 + 519,57X = 1.117,1 + 519,57\sqrt{t} \qquad (R^2 = 0,981)$$

Das Bestimmtheitsmaß lässt erkennen, dass die Anpassungsgüte des Quadratwurzel-Modells an die beobachteten Daten besser ist als die des linearen Modells. Der gute Fit ist auch in Abbildung 2.9 zu erkennen.
Die Punktprognose für Periode T+1 = 11 ergibt:

$$\hat{y}_{11} = 1.117,1 + 519,57\sqrt{11} = 2.840$$

und liegt um mehr als 100 Einheiten niedriger als beim linearen Modell. Noch erheblich größer ist der Unterschied bei der Prognose für Periode 20. Man erhält hier den Wert

$$\hat{y}_{20} = 1.117,1 + 519,57\sqrt{20} = 3.441$$

der fast 600 Einheiten niedriger liegt als der des linearen Modells.

Zeit (Periode t)	Menge (Kartons pro Periode)	$X = \sqrt{t}$
1	1.657	1,000
2	1.864	1,414
3	1.950	1,732
4	2.204	2,000
5	2.288	2,236
6	2.410	2,449
7	2.414	2,646
8	2.534	2,828
9	2.739	3,000
10	2.785	3,162

Abbildung 2.8: Zeitreihe der Absatzmenge und transformierte Zeitvariable

Prognosefehler

Analog zu (2.9) lässt sich der Prognosefehler nach nichtlinearer Transformation der Zeitvariablen t wie folgt berechnen:

$$s_p(T+k) = s \cdot \sqrt{1 + \frac{1}{T} + \frac{(x_{T+k} - \overline{x})^2}{\sum_t (x_t - \overline{x})^2}} = s \cdot \sqrt{1 + \frac{1}{T} + \frac{(x_{T+k} - \overline{x})^2}{(T-1) \cdot s_x^2}} \qquad (2.15)$$

mit

x_{T+k} = Wert der transformierten Zeitvariablen für Periode $T+k$
s = Standardfehler der Regression
$\overline{x}$ = Mittelwert der transformierten Zeitvariablen
s_x = Standardabweichung der transformierten Zeitvariablen

Es gilt hier $x_{T+k} = \sqrt{T+k}$ und damit $x_{11} = \sqrt{11} = 3,317$. Mit den Werten $s = 51,55$, $\overline{x} = 2,247$ und $s_x = 0,708$ erhält man hier für den Prognosewert $\hat{y}_{11} = 2.840$ den Prognosefehler

$$s_p(11) = 51,55 \cdot \sqrt{1 + \frac{1}{10} + \frac{(3,317 - 2,247)^2}{9 \cdot 0,708^2}}$$

$$= 51,55 \cdot \sqrt{1 + 0,1 + 0,254} = 51,55 \cdot 1,163 = 60$$

Mit $x_{20} = \sqrt{20} = 4,472$ erhält man für den Prognosewert $\hat{y}_{20} = 3.441$ den Prognosefehler

$$s_p(20) = 51,55 \cdot \sqrt{1 + \frac{1}{10} + \frac{(4,472 - 2,247)^2}{9 \cdot 0,708^2}}$$

$$= 51,55 \cdot \sqrt{1 + 0,1 + 1,097} = 51,55 \cdot 1,482 = 76$$

Die Prognosefehler des Quadratwurzel-Modells sind deutlich kleiner als die des linearen Modells.

Prognoseintervall

Prognoseintervall

Mit Hilfe der obigen Prognosefehler und $t_{\alpha/2} = 2{,}306$ (für die Vertrauenswahrscheinlichkeit $1 - \alpha = 0,95$ und T−2=8 Freiheitsgrade) erhält man die folgenden Prognoseintervalle für die Perioden 11 und 20:

$$2.702 \leq y_{11} \leq 2.979$$
$$3.265 \leq y_{20} \leq 3.617$$

Abbildung 2.9 zeigt für das Quadratwurzel-Modell den Verlauf der Prognose und des Prognoseintervalls über die Zeit.

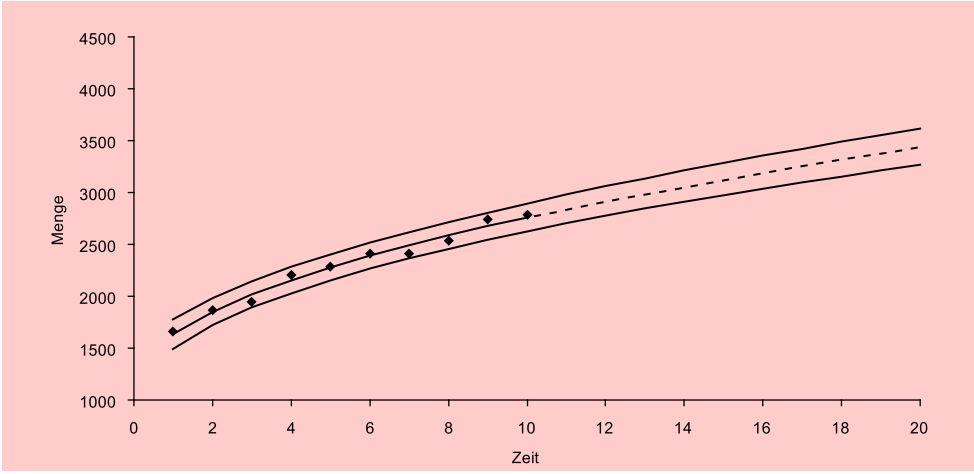


Abbildung 2.9: Quadratwurzel-Modell: Geschätzte Funktion, prognostizierter Verlauf und 95 %-Prognoseintervall

2.3.1.2 Das Logarithmische Modell

Es existiert eine Vielzahl weiterer nichtlinearer Trendmodelle. Zu den Modellen, die wie das Quadratwurzel-Modell ein sich abflachendes Wachstum (einen degressiv steigenden Verlauf) aufweisen, gehört das logarithmische Modell

Degressiv steigend

$$Y = \alpha + \beta \cdot \ln(t) + u \tag{2.16}$$

dessen Verlauf in Abbildung 2.10 dem Verlauf des Quadratwurzel-Modells gegenübergestellt ist. Das logarithmische Modell verläuft erheblich flacher als das Quadratwurzel-Modell. Die Schätzung des logarithmischen Modells liefert hier:

$$\hat{Y} = 1.540,6 + 492,56 \cdot \ln(t) \qquad (R^2 = 0,945)$$

Die Anpassungsgüte an die beobachteten Daten ist in diesem Fall schlechter als die des linearen Modells.

2.3.1.3 Das Multiplikative Modell

Ein sehr flexibles und ebenfalls häufig verwendetes nichtlineares Modell bildet das *multiplikative Modell* (vgl. Formel 1.26):

$$Y = \alpha \cdot t^{\beta} \cdot u \tag{2.17}$$

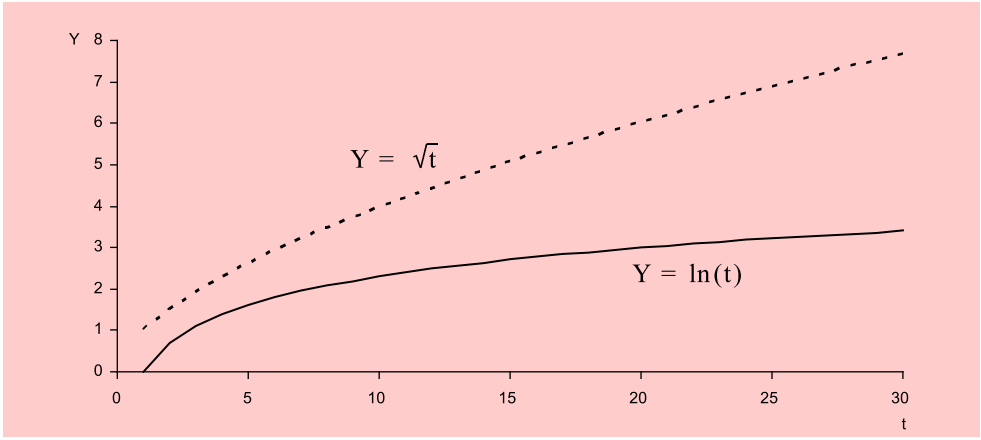


Abbildung 2.10: Logarithmische Funktion $Y = \ln(t)$ und Quadratwurzel-Funktion (gestrichelte Linie)

Das multiplikative Modell kann, je nach der Größe von β, unterschiedliche Verlaufsformen annehmen (vgl. Abbildung 2.11):

(a) $\beta > 1$: progressiv steigender Verlauf (konvexer Verlauf)

(b) $0 < \beta < 1$: degressiv steigender Verlauf (konkaver Verlauf)

(c) $\beta < 0$: degressiv fallender Verlauf

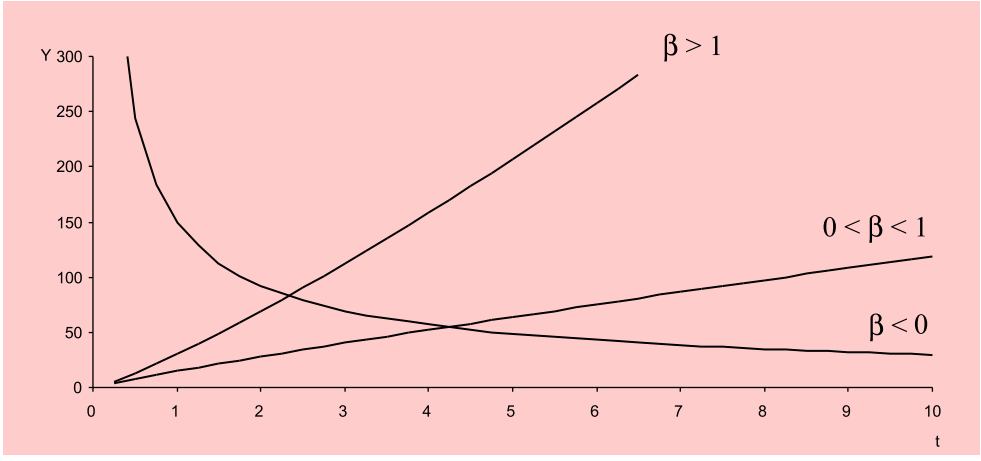


Abbildung 2.11: Multiplikatives Modell: Funktionsverläufe für unterschiedliche Werte von β

Linearisierte Form

Durch Logarithmierung des multiplikativen Modells erhält man die linearisierte Form:

$$\ln(Y) = \ln(\alpha) + \beta \cdot \ln(t) + \ln(u)$$

Die Schätzung mittels linearer Regression für die vorliegenden Absatzmengen liefert:

$$\ln(\hat{Y}) = 7,378 + 0,228 \cdot \ln(t) \qquad (R^2 = 0,971)$$

Nach Delogarithmieren erhält man damit die folgende Funktion:

$$\hat{Y} = 1.600 \cdot t^{0,228} \qquad (R^2 = 0,971)$$

Der Verlauf des multiplikativen Modells liegt zwischen dem des Quadratwurzel-Modells und dem des logarithmischen Modells.

2.3.1.4 Das Potenz-Modell

Ähnlich flexibel, wie das multiplikative Modell, ist das Potenz-Modell, das allerdings drei unbekannte Parameter enthält:

$$Y = \alpha + \beta \cdot t^{\gamma} + u \tag{2.18}$$

Wie das multiplikative Modell nimmt es für unterschiedliche Werte von γ verschiedene Verlaufsformen an:

(a) $\quad \gamma > 1:$ progressiv steigender Verlauf (konvexer Verlauf)
(b) $\quad 0 < \gamma < 1:$ degressiv steigender Verlauf (konkaver Verlauf)

Für $\gamma = 0,5$ ergibt sich das Quadratwurzel-Modell, das einen Spezialfall des Potenz-Modells bildet.

Das Potenz-Modell ist ein echtes bzw. *intrinsisch nichtlineares Modell*. Es ist nichtlinear in Bezug auf den unbekannten Parameter γ und lässt sich daher nicht mittels linearer Regressionsanalyse schätzen. Für die Lösung des Kleinstquadrateprinzips existiert keine analytische Lösung. Vielmehr bedarf es hierfür der Anwendung iterativer Methoden (siehe dazu Kapitel 10: Nichtlineare Regression). Man erhält hier:

Intrinsisch nichtlinear

$$\hat{Y} = 1.298 + 362,3 \cdot t^{0,611} \qquad (R^2 = 0,984)$$

Der Verlauf des Potenz-Modells ist in diesem Fall nur geringfügig steiler als der des Quadratwurzel-Modells und die Anpassungsgüte nur marginal besser.

2.3.1.5 Vergleich der Modelle

In der Abbildung 2.12 sind die fünf obigen Trend-Modelle mit ihren Gütemaßen sowie Prognosewerten und Prognosefehlern für Periode 11 zusammengestellt. Den besten Fit weisen hier das Potenz-Modell und das Quadratwurzel-Modell auf. Während das Bestimmtheitsmaß des Potenz-Modells noch etwas größer ist als das des Quadratwurzel-Modells, ist der F-Wert des Potenz-Modells bedeutend niedriger, da ein Parameter mehr zu schätzen ist als beim Quadratwurzel-Modell und den anderen Modellen.

Fit

Betrachtet man die prognostizierten Werte für Periode 11, so fällt auf, dass diese beim logarithmischen Modell und beim multiplikativen Modell unter dem letzten Beobachtungswert (2.785 für Periode 10) liegen. Im Beobachtungszeitraum kommt dies nicht vor, sondern die Absatzmenge hat sich kontinuierlich jeweils gegenüber der Vorperiode erhöht. Damit erscheint es naheliegend, dass der Manager des Margarineherstellers diese Modelle nicht auswählen wird. Da auch das lineare Modell nicht adäquat erscheint, verbleiben zur Erstellung seiner Absatzprognose hier nur das Quadratwurzel-Modell oder das Potenz-Modell. Diese beiden Modelle weisen auch den geringsten Prognosefehler auf.

Modell	geschätzte Funktion	R^2	F	$\hat{y}_{11}$	$s_p(11)$
Linear	$\hat{Y} = 1.620 + 120,9t$	0,972	277	2.949	80
Quadratwurzel	$\hat{Y} = 1.117 + 519,6\sqrt{t}$	0,983	459	2.840	60
Logarithmisch	$\hat{Y} = 1.542 + 492,6 \cdot \ln(t)$	0,945	139	2.722	104
Multiplikativ	$\hat{Y} = 1.600 \cdot t^{0,228}$	0,971	265	2.761	96
Potenz	$\hat{Y} = 1.298 + 362,3 \cdot t^{0,611}$	0,984	216	2.866	59

Abbildung 2.12: Vergleich der Modelle

2.3.2 Berücksichtigung von Strukturbrüchen

Die in Abbildung 2.13 dargestellten Trendverläufe weisen jeweils einen Strukturbruch auf. Derartige Strukturbrüche findet man häufig bei Zeitreihenanalysen, z. B. wenn durch Änderung der wirtschaftlichen Rahmenbedingungen eine Änderung in der zeitlichen Entwicklung einer betrachteten Variablen Y bewirkt wird. Strukturbrüche lassen sich durch eine Dummy-Variable berücksichtigen, deren Werte vor dem Strukturbruch in Periode t' Null sind und danach Werte größer als Null annehmen.

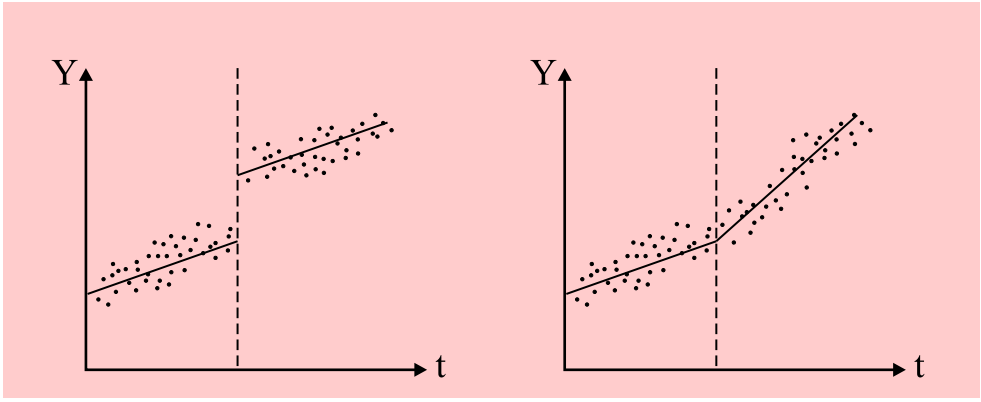


Abbildung 2.13: Strukturbrüche

Niveauänderung

a) Niveauänderung in Periode t = t':

$$Y = \alpha + \beta_1 \cdot t + \beta_2 \cdot q + u \qquad \text{mit } q = \begin{cases} 0 & \textit{für} \quad t < \text{t'} \\ 1 & \textit{für} \quad t \geq \text{t'} \end{cases} \tag{2.19}$$

Die Werte der Dummy-Variablen q über die Zeit t sind nachfolgend dargestellt:

t	1	2	3	...	t'−1	t'	t'+1	t'+2	...	t'+k
q	0	0	0	...	0	1	1	1	...	1

Vor dem Strukturbruch gilt damit:

$$t < \text{t'}: \qquad Y = \alpha + \beta_1 \cdot t + u \tag{2.20}$$

Nach dem Strukturbruch erhöht sich das Niveau um β_2:

$$t \geq \text{t'}: \qquad Y = \alpha + \beta_1 \cdot t + \beta_2 \cdot q + u = (\alpha + \beta_2) + \beta_1 \cdot t + u \tag{2.21}$$

Durch lineare Regression von Y auf die Variablen t und q lassen sich die drei Parameter α, β_1 und β_2 in (2.19) schätzen.

b) Trendänderung in Periode t = t': Trendänderung

$$Y = \alpha + \beta_1 \cdot t + \beta_2 \cdot q + u \qquad \text{mit } q = \begin{cases} 0 & \textit{für} \quad t < \text{t'} \\ t - \text{t'} + 1 & \textit{für} \quad t \geq \text{t'} \end{cases} \tag{2.22}$$

Die Werte der Dummy-Variablen q über die Zeit t sind nachfolgend dargestellt:

t	1	2	3	...	t'−1	t'	t'+1	t'+2	...	t'+k
q	0	0	0	...	0	1	2	3	...	k+1

Vor dem Strukturbruch gilt damit:

$$t < \text{t'}: \qquad Y = \alpha + \beta_1 \cdot t + u \tag{2.23}$$

Ab dem Strukturbruch erhöht sich die Steigung um β_2:

$$\begin{aligned} t \geq \text{t'}: \qquad Y =& \alpha + \beta_1 \cdot t + \beta_2 \cdot (t - \text{t'} + 1) + u \\ =& (\alpha - \beta_2 \cdot \text{t'} + \beta_2) + (\beta_1 + \beta_2) \cdot t + u \\ =& \alpha\text{'} + (\beta_1 + \beta_2) \cdot t + u \end{aligned} \tag{2.24}$$

Für den rechten Ast der Trendfunktion, falls man diesen nach links verlängern würde, ergibt sich der fiktive Schnittpunkt α' mit der Y-Achse, der auch negativ sein kann.

Mittels linearer Regression von Y auf die Variablen t und q lassen sich wiederum die drei unbekannten Parameter α, β_1 und β_2 in (2.22) schätzen.

2.3.3 Berücksichtigung von zyklischen Schwankungen

Bei diskreten Zeitreihen lassen sich auch zyklische Schwankungen, wie Saisoneffekte oder Konjunkturschwankungen, mit Hilfe von Dummy-Variablen berücksichtigen. Als Beispiel diene das folgende Modell mit Trend und Saisoneffekten: Saisoneffekte und Konjunkturschwankungen

$$Y = \alpha + \beta_1 \cdot t + \gamma_1 \cdot q_1 + \gamma_2 \cdot q_2 + \gamma_3 \cdot q_3 + u \tag{2.25}$$

$$\text{mit } q_1 = \begin{cases} 1 & \text{für Frühjahr} \\ 0 & \text{sonst} \end{cases}, q_2 = \begin{cases} 1 & \text{für Sommer} \\ 0 & \text{sonst} \end{cases}, q_3 = \begin{cases} 1 & \text{für Herbst} \\ 0 & \text{sonst} \end{cases}$$

Für den Fall, dass Monatsdaten vorliegen, ergibt sich für die Dummy-Variablen die in Abbildung 2.14 dargestellte Datenstruktur.

Die vierte Dummy-Variable für Winter wird in obiger Modellformulierung nicht benötigt. Das vierte Quartal bildet quasi die Basis und die Koeffizienten der drei Dummy-Variablen geben die Abweichungen von dieser Basis an. Der Effekt des Basisquartals ist im konstanten Glied α enthalten. Würde man auch die vierte Dummy-Variable in das Modell aufnehmen, ließe es sich wegen perfekter Multikollinearität nicht mehr schätzen, da die Addition der vier Dummy-Variablen immer den konstanten Wert Eins ergibt. Multikollinearität

Trend			Dummy-Variablen			
		t (Monat)	q_1	q_2	q_3	q_4
1. Zyklus (Jahr)	1. Saison (Quartal)	1	1	0	0	0
		2	1	0	0	0
		3	1	0	0	0
	2. Saison (Quartal)	4	0	1	0	0
		5	0	1	0	0
		6	0	1	0	0
	3. Saison (Quartal)	7	0	0	1	0
		8	0	0	1	0
		9	0	0	1	0
	4. Saison (Quartal)	10	0	0	0	1
		11	0	0	0	1
		12	0	0	0	1
2. Zyklus (Jahr)	1. Saison (Quartal)	13	1	0	0	0
		14	1	0	0	0
		15	1	0	0	0

Abbildung 2.14: Datenstruktur der Dummy-Variablen für saisonale Schwankungen

Entfernt man das konstante Glied aus der Regressionsgleichung, so lassen sich alle vier Dummy-Variablen einbeziehen und man erhält das folgende Modell:

$$Y = \beta_1 \cdot t + \gamma_1 \cdot q_1 + \gamma_2 \cdot q_2 + \gamma_3 \cdot q_3 + \gamma_4 \cdot q_4 + u \tag{2.26}$$

Durch Wegfall des konstanten Gliedes wird hier die Multikollinearität vermieden.

2.3.4 Umsetzung mit SPSS

Zur Durchführung von Zeitreihenanalysen auf Basis regressionsanalytischer Modelle lassen sich in SPSS die Prozeduren „Lineare Regression“, „Kurvenanpassung“ und „Nichtlineare Regression“ verwenden.

2.3.4.1 Lineare Regression

Die Schätzung der obigen Zeitreihenmodelle lässt sich mit Hilfe der Prozedur „Lineare Regression“ durchführen. Eine Ausnahme bildet das Potenz-Modell (2.18), bei dem es sich um ein intrinsisch nichtlineares Modell handelt, für dessen Schätzung die Prozedur „Nichtlineare Regression“ erforderlich ist.[22] Alle anderen Modelle lassen sich

Transformation

mit Hilfe nichtlinearer Transformationen oder der Erzeugung von Dummy-Variablen auf eine lineare Form bringen. Die erforderlichen nichtlinearen Transformationen können mit Hilfe des Menüpunktes „Transformieren / Variable berechnen“ vorgenommen werden. Damit lassen sich neue Variablen berechnen, mittels derer sodann die lineare Regression durchgeführt werden kann. Die Prozedur „Lineare Regression“ enthält allerdings keine Option zur Durchführung von Prognoserechnungen. Die Erstellung von

Prognose

Prognosen muss der Untersucher auf Basis der geschätzten Modelle selber vornehmen.

[22] Zur Nichtlinearen Regressionsanalyse siehe Backhaus/Erichson/Weiber (2015).

Um den Verlauf der geschätzten Funktion in die Zukunft zu extrapolieren, kann wiederum der Menüpunkt „Transformieren / Variable berechnen“ oder ein Taschenrechner verwendet werden. Zur Berechnung des Prognosefehlers gemäß (2.15) liefert die Prozedur „Lineare Regression“ den Standardfehler der Regression s. Den Mittelwert $\overline{x}$ und die Standardabweichung s_x der transformierten Prädiktorvariablen kann man mit der Prozedur „Deskriptive Statistik“ berechnen. Somit sind für die Berechnung des Prognosefehlers keine Berechnungen auf Basis der individuellen Zeitreihendaten erforderlich, was bei langen Zeitreihen sehr aufwändig sein könnte.

2.3.4.2 Kurvenanpassung

Die Arbeit wird wesentlich erleichtert, wenn man die Prozedur „Kurvenanpassung“ (Curve Fit) nutzen kann. In dieser Prozedur werden dem Benutzer das lineare Trendmodell sowie zehn weitere nichtlineare Trendmodelle zur Verfügung gestellt (siehe Abbildung 2.15). Es werden globale Gütemaße (Bestimmtheitsmaß und F-Wert) sowie die Standardfehler und t-Werte der Regressionsparameter ermittelt, und die beobachteten und geschätzten Werte der Zeitreihenvariablen Y werden grafisch dargestellt. Außerdem lassen sich durch Extrapolation des geschätzten Modells auch Punkt- und Intervallprognosen durchführen.[23]

Trendmodell

Für die Erstellung von Prognosen ist es erforderlich, dass die unabhängige Variable „Zeit“ nicht aus der Datendatei ausgewählt wird, sondern durch die Prozedur „Kurvenanpassung“ erzeugt wird. Weiterhin ist in der Dialogbox die Option „Speichern“ (Save) aufzurufen. Dort lassen sich eine Vertrauenswahrscheinlichkeit (90 %, 95 % oder 99 %) sowie die Periode, bis zu der Prognosen erstellt werden sollen (Prognosehorizont), wählen. Die prognostizierten Werte sowie die zugehörigen Prognoseintervalle werden anschließend in der Datendatei gespeichert. Hierzu werden drei neue Variablen erzeugt: „FIT“ (geschätzte bzw. prognostizierte Werte), „LCL“ (Lower Confidence Limit) und „UCL“ (Upper Confidence Limit).

Prognose

In Abbildung 2.15 sind die elf Modelle zusammengestellt. Alle Modelle sind intrinsisch linear, d. h. sie lassen sich, soweit erforderlich, mittels nichtlinearer Transformationen auf eine lineare Form bringen und somit mittels linearer Regression schätzen. Im rechten Teil von Abbildung 2.15 sind die linearisierten Formen dargestellt. Auf die Einbeziehung der Störgröße wurde hier verzichtet.

Die Modelle 1, 2 und 6 wurden oben behandelt. Das Modell 6 (Exponent) entspricht dem multiplikativen Modell (2.17).

Die Modelle 4 (Quadratisch) und 5 (Kubisch) sind Polynome zweiten und dritten Grades. Sie erzielen zwar für die Zeitreihendaten der Absatzmenge einen guten Fit, sind aber für mittel- oder langfristige Prognosen ungeeignet, da sie sich wegen der großen Exponenten sehr schnell vom Wertebereich der Beobachtungsdaten entfernen und auch negative Werte annehmen können.

Die übrigen Modelle sind zur Modellierung der obigen Zeitreihe, die einen recht flach gekrümmten, degressiv steigenden (konkaven) Verlauf besitzt, nicht geeignet. Das Inverse Modell 3 besitzt einen sehr stark gekrümmten Verlauf und würde viel zu niedrige Prognosen erbringen. Die Schätzung der Modelle 7 bis 11 führt hier zu progressiv steigenden (konvexen) Verläufen.

Konkav

Konvex

Die Modelle 5 (Kubisch) und 9 (Logistisch) können auch s-förmige Verläufe annehmen.[24] Das logistische Modell ist ein sog. Wachstumsmodell (Growth Model), das für

S-förmig

[23] Vgl. IBM Cooperation (o.J.), S. 79 ff.

Nr.	Name	Formel	Linearisierung
1	Linear	$Y = \alpha + \beta \cdot t$	
2	Logarithmisch	$Y = \alpha + \beta \cdot \ln(t)$	
3	Invers	$Y = \alpha + \beta / t$	$Y = \alpha + \beta \cdot 1/t$
4	Quadratisch (Quadratic)	$Y = \alpha + \beta_1 \cdot t + \beta_2 \cdot t^2$	
5	Kubisch (Cubic)	$Y = \alpha + \beta_1 \cdot t + \beta_2 \cdot t^2 + \beta_3 \cdot t^3$	
6	Potenz (Power)	$Y = \alpha \cdot t^\beta$	$\ln(Y) = \ln(\alpha) + \beta \cdot \ln(t)$
7	Zusammengesetzt (Compound)	$Y = \alpha \cdot \beta^t$	$\ln(Y) = \ln(\alpha) + \ln(\beta) \cdot t$
8	S-Kurve (S-curve)	$Y = e^{\alpha + \beta / t}$	$\ln(Y) = \alpha + \beta \cdot 1/t$
9	Logistisch	$Y = \frac{1}{1/M + \alpha \cdot \beta^t}$	$\ln(1/Y - 1/M) = \ln(\alpha) + \ln(\beta) \cdot t$
10	Wachstum (Growth)	$Y = e^{\alpha + \beta \cdot t}$	$\ln(Y) = \alpha + \beta \cdot t$
11	Exponentiell (Exponential)	$Y = \alpha \cdot e^{\beta \cdot t}$	$\ln(Y) = \ln(\alpha) + \beta \cdot t$

Abbildung 2.15: Modelle der SPSS-Prozedur „Kurvenanpassung“

wachsendes t gegen eine Wachstumsgrenze, den Maximalwert M, konvergiert.[25] Sein Verlauf ist symmetrisch um den Wendepunkt (vgl. Abbildung 2.16), der bei $y_w = M/2$ liegt. Es findet in vielen Bereichen Verwendung, so z. B. in der Innovations- und Diffusionsforschung[26], in der Epidemiologie sowie auch als Approximation der Verteilungsfunktion der Gauss'schen Normalverteilung, für die kein analytischer Ausdruck existiert (vgl. dazu auch Kapitel 5: Logistische Regression). In der Marktforschung wird das logistische Modell u.a. zur Modellierung der Absatzmenge bei der Einführung von neuen Produkten verwendet.

Das logistische Modell ist eigentlich ein intrinsisch nichtlineares Modell. Es lässt sich nur linearisieren, wenn die Wachstumsgrenze M bekannt ist. In der Prozedur „Kurvenanpassung“ muss daher M durch den Benutzer spezifiziert werden. Der anzugebende Wert muss größer sein als der größte Wert der vorliegenden Zeitreihe. Zur Schätzung des logistischen Modells siehe auch Kapitel 1 im Buch *Fortgeschrittene Multivariate Analysemethoden*: Nichtlineare Regression.

Die Formulierung des logistischen Modells in SPSS ist etwas ungewöhnlich. Üblicher ist die folgende Formulierung:

$$Y = \frac{M}{1 + e^{\alpha - \beta \cdot t}} \tag{2.27}$$

die sich für bekanntes M wie folgt linearisieren lässt:

$$\ln\left(\frac{M}{Y} - 1\right) = \alpha - \beta \cdot t \tag{2.28}$$

[24] Das als „S-Kurve“ bezeichnete Exponentialmodell 8 kann, ähnlich dem Inversen Modell 3, nur konvexe oder konkave Verläufe annehmen, je nachdem, ob $\beta > 0$ oder $\beta < 0$ gilt. Die Bezeichnung als „S-Kurve“ ist daher irreführend.

[25] Bei dem Modell 10 („Wachstumsfaktor“ bzw. „Growth“) handelt es sich nicht um ein Wachstumsmodell im engeren Sinne, da es nicht gegen einen oberen Grenzwert konvergiert. Einen Überblick über Wachstumsmodelle geben z. B. Mertens, P. (1994), S. 157 ff., Hammann/Erichson (2000), S. 446 ff.

[26] Vgl. z. B. Meffert/Steffenhagen (1977), S. 70 ff., Weiber (1992).

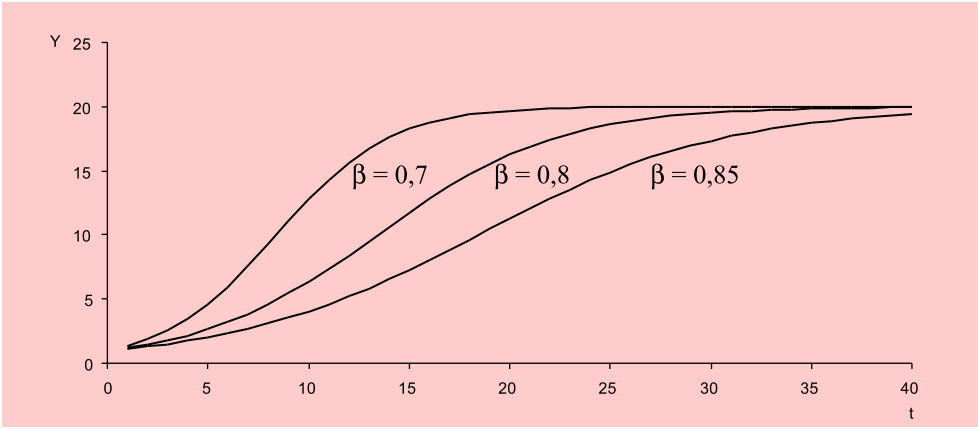


Abbildung 2.16: Logistisches Modell für verschiedene Werte von β

Abbildung 2.16 zeigt den Verlauf für $M = 20$ und verschiedene Werte von β ($0 < \beta < 1$). Für größere Werte von β wird der Verlauf flacher. Der Parameter α dagegen verändert die Form des Modells nicht, sondern bewirkt mit zunehmender Größe eine Verschiebung der Funktion nach rechts. In Modell 9 muss $\alpha > 0$ gelten, während in der Formulierung gemäß (2.27) der Parameter α beliebige Werte annehmen kann.

2.4 Fallbeispiel

2.4.1 Problemstellung

Während in dem oben betrachteten Absatzgebiet der Margarineabsatz über die letzten Jahre kontinuierlich angestiegen ist, mussten dagegen in den meisten anderen Absatzgebieten z. T. erhebliche Einbußen hingenommen werden. Der Manager unseres Margarineherstellers fragt sich daher, ob dies an unzureichenden Marketingbemühungen des Unternehmens liegt oder ob sich die Marktbedingungen geändert haben. Überdies kam es im Jahresablauf des öfteren zu Lieferengpässen, da die Absatzmengen stark schwankten. Für die Produktionsplanung wäre daher sowohl eine kurzfristige Prognose des Absatzvolumens für die jeweils nächsten Monate wie auch eine mittelfristige Prognose für die kommenden Jahre von großer Wichtigkeit.

Prognose des Absatzvolumens

Um dem Problem auf den Grund zu gehen, beabsichtigt der Manager unseres Margarineherstellers, zunächst den Gesamtmarkt für Margarine in Deutschland zu analysieren und zu prognostizieren. Hierzu besorgt er sich von der Gesellschaft für Konsumforschung (GfK AG) in Nürnberg die monatlichen Absatzdaten der letzten vier Jahre, die in Abbildung 2.17 als Streudiagramm dargestellt sind.[27] Auch die Absatzmengen des Gesamtmarktes weisen eine erhebliche Streuung auf. Es stellt sich die Frage, ob diese Streuung eher zufällig bedingt ist, oder ob sich dahinter eine systematische Entwicklung verbirgt, die sich modellieren und für Prognosen nutzen lässt.

[27] Der GfK AG sei an dieser Stelle für die Überlassung der Daten gedankt.

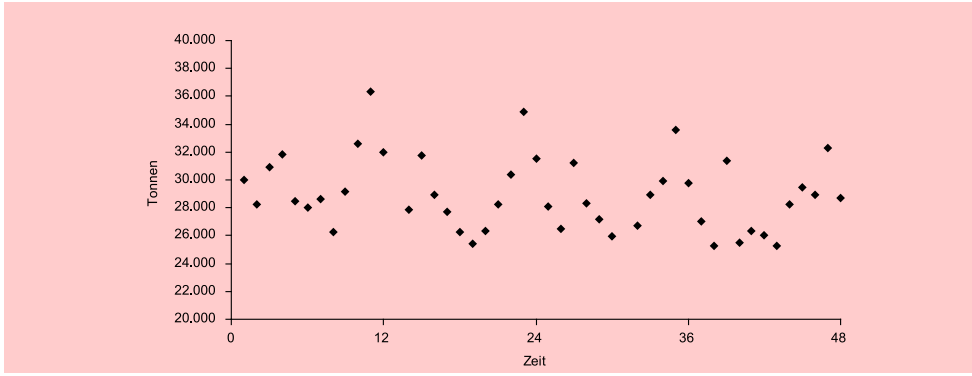


Abbildung 2.17: Zeitreihe des Marktabsatzvolumens von Margarine in Tonnen (Monatliche Daten von Jan. 2004 - Dez. 2007, Quelle: GfK ConsumerScan)

2.4.2 Ergebnisse

2.4.2.1 Extrapolationsmodelle

Der Manager beginnt zunächst mit der Anpassung eines linearen Trendmodells gemäß (5) an die Daten.

Linearer Trend

Modell 1: Linearer Trend

$$\hat{Y} = 29.836 - 41,7 \cdot \text{Zeit} \qquad (R^2 = 0,046)$$

Das lineare Trendmodell zeigt eine negative Entwicklung des Marktabsatzvolumens von Margarine in Deutschland über die Jahre 2004 bis 2007. Durch den Trend, der in Abbildung 2.18 grafisch dargestellt ist, lässt sich die Streuung in den Absatzdaten allerdings nur in geringem Umfang erklären. Das Bestimmtheitsmaß beträgt lediglich 4,6 %.

Schwankungen

In Abbildung 2.18 sind die Streupunkte des Marktabsatzvolumens durch Linien verbunden. Man erkennt so, dass die starken Schwankungen eine gewisse Regelmäßigkeit aufweisen, die sich im Abstand von 12 Monaten wiederholt. Daraus ist zu folgern, dass sie primär saisonal bedingt sind.

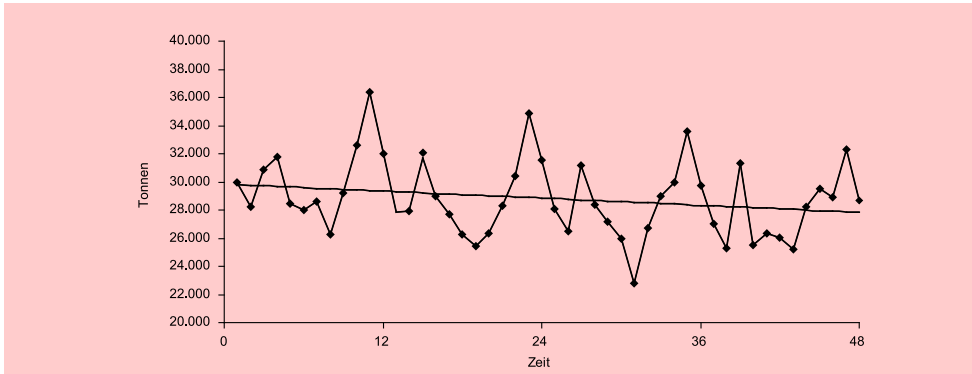


Abbildung 2.18: Entwicklung des Marktabsatzvolumens von Margarine (Jan. 2004 - Dez. 2007)

Modell 2: Trend + saisonale Dummies Saisonale Dummies

$$
\begin{aligned}
\hat{Y} = 29.439 \quad & - \quad 64,0 \text{ Zeit } (-5,3) * * \qquad (R^2 = 0,873) \\
& - \quad 1.194 \text{ Februar } (-1,5) \\
& + \quad 3.198 \text{ März } (+4,0) * * \\
& + \quad 634 \text{ April } (+0,8) \\
& - \quad 547 \text{ Mai } (-0,7) \\
& - \quad 1.343 \text{ Juni } (-1,7) \\
& - \quad 2.319 \text{ Juli } (-2,9) * * \\
& - \quad 892 \text{ August } (-1,1) \\
& + \quad 1.270 \text{ September } (+1,6) \\
& + \quad 2.816 \text{ Oktober } (+3,5) * * \\
& + \quad 6.706 \text{ November } (+8,3) * * \\
& + \quad 2.981 \text{ Dezember } (+3,7) * *
\end{aligned}
$$

Durch die Einbeziehung von monatlichen Dummy-Variablen zur Erfassung der saisonalen Schwankungen lässt sich der Erklärungsanteil des Modells auf 87,3 % erhöhen. Erklärungsanteil

Die t-Werte der Regressionskoeffizienten sind (hier und im Folgenden) jeweils in Klammern angegeben. Werte mit einem Signifikanzniveau von 5 % sind durch einen Stern (*) und Werte mit einem Signifikanzniveau von 1 % sind durch zwei Sterne (**) gekennzeichnet. Signifikanzniveau

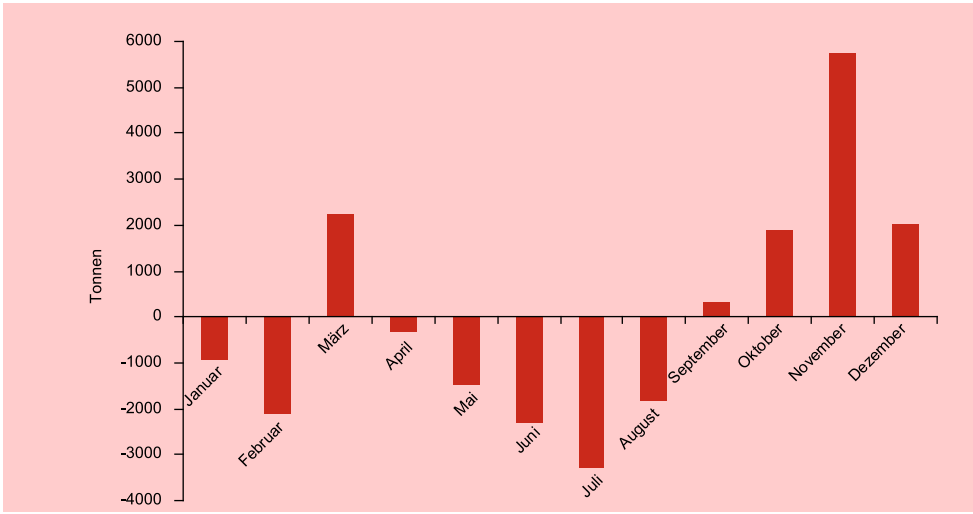


Abbildung 2.19: Saisonfigur im Margarinekonsum (2004–2007)

Aus den Regressionskoeffizienten der monatlichen Dummy-Variablen lässt sich die in Abbildung 2.19 dargestellte Saisonfigur des Marktabsatzvolumens von Margarine ableiten. Die Werte wurden hierzu um den Mittelwert zentriert. Auffällig sind die

starken positiven Ausschläge im März und ganz besonders im November, die die umsatzstärksten Monate bilden. Der Rückgang in den Sommermonaten hängt vermutlich mit den höheren Temperaturen und dem dadurch bedingten geringeren Kalorienbedarf zusammen. Dies soll im Folgenden näher untersucht werden.

Saisonfigur

Die zwei vorstehenden Regressionsmodelle machen ein Muster in den Schwankungen der Zeitreihe deutlich (Trend- und Saisonkomponente) und trennen dieses von den zufälligen Schwankungen. Mittels Extrapolation dieses Musters lassen sich Prognosen für beliebige zukünftige Perioden erstellen. Die Prognosen implizieren allerdings die Annahme, dass das Muster stabil ist, sich also in der Zukunft nicht verändert.

Prognose

Derartige Extrapolationsmodelle beschreiben lediglich den Verlauf der Zeitreihe, können ihn aber nicht erklären, d. h. sie berücksichtigen nicht die Faktoren, die den Verlauf der Zeitreihe beeinflussen bzw. ihr kausal zugrunde liegen. Prognosen lassen sich damit recht einfach erstellen, da man keine weiteren Informationen benötigt. Nachteilig ist, dass theoretische Überlegungen bezüglich kausaler Zusammenhänge, die die betrachtete Zeitreihe tangieren, außer acht bleiben.

Extrapolationsmodell

2.4.2.2 Strukturmodelle

Ein Vorteil der Regressionsanalyse gegenüber vielen anderen Verfahren der Zeitreihenanalyse ist darin zu sehen, dass sich die Modelle leicht durch Einbeziehung weiterer Variablen, die den Verlauf der Zeitreihe möglicherweise erklären, erweitern lassen und somit theoretische Überlegungen in die Modellbildung einbezogen werden können. Dies erfordert allerdings, dass für diese erklärenden Variablen entsprechende Zeitreihendaten vorliegen müssen. Die Prognose mittels derartiger *Strukturmodelle* wird überdies erschwert, da die Prognose der interessierenden Zeitreihenvariablen Y i. d. R. zunächst eine Prognose der erklärenden Variablen (Prädiktoren) erforderlich macht.

Strukturmodelle

Relevante Einflussfaktoren der Nachfrage nach Konsumgütern bilden z. B. die Größe der Bevölkerung und in vielen Fällen auch das Wetter. Ein Indikator des Wetters ist z. B. die mittlere Temperatur pro betrachteter Periode, in diesem Fall pro Monat. Nachfolgend soll zunächst der Einfluss der Temperatur auf den Margarinekonsum und sodann der Einfluss des Bevölkerungswachstums untersucht werden.

2.4.2.2.1 Einbeziehung der Temperatur

In Abbildung 2.20 ist der Temperaturverlauf (monatliche Durchschnittswerte) über den Zeitraum, für den die Margarine-Daten vorliegen, dargestellt. Daraus lässt sich ein leichter Anstieg erkennen.

Temperaturverlauf

Analog zur Saisonfigur des Margarinekonsums in Abbildung 2.19 ist in 2.21 die Saisonfigur des jährlichen Temperaturverlaufs dargestellt. Ein Vergleich der Abbildungen macht deutlich, dass sich die saisonalen Schwankungen des Margarinekonsums nur unvollständig durch die Temperaturschwankungen erklären lassen. Dies zeigt genauer die nachfolgende Regressionsanalyse.

Durch die Einbeziehung der Temperaturdaten erhöht sich der Erklärungsanteil des Modells von 87,3 % geringfügig auf 89,1 %. Die um den Temperatureinfluss bereinigten Regressionskoeffizienten der monatlichen Dummy-Variablen dagegen verändern sich z. T. erheblich. Sie sind in Abbildung 2.22 grafisch dargestellt. Dies deutet darauf hin, das die saisonalen Schwankungen des Margarinekonsums in starkem Maße im Verbraucherverhalten begründet liegen.

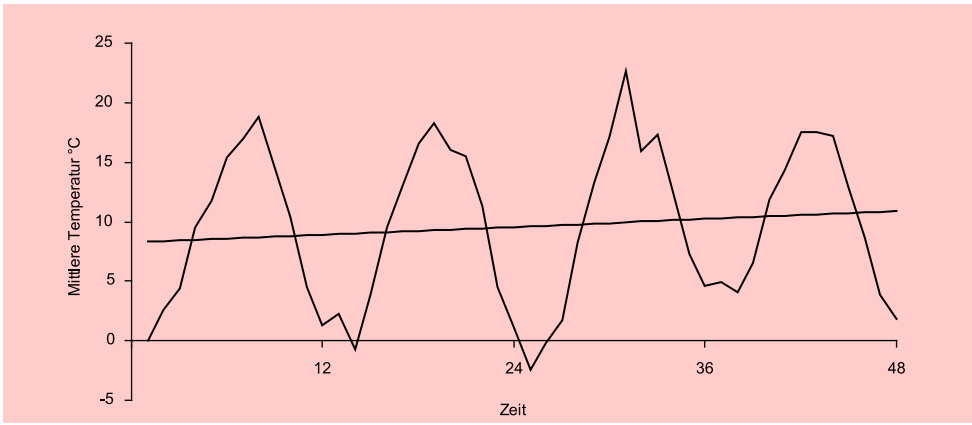


Abbildung 2.20: Temperaturverlauf Jan. 2004 - Dez. 2007 (monatliche Durchschnittswerte)

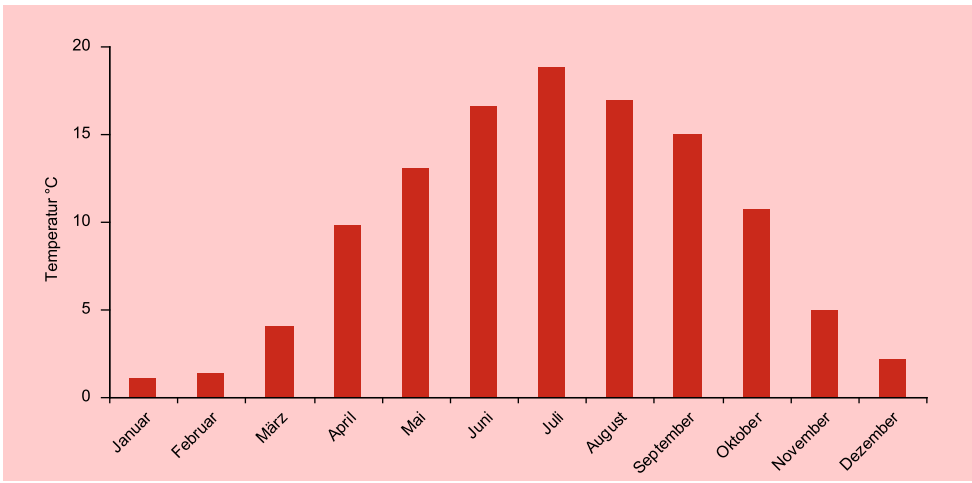


Abbildung 2.21: Saisonfigur der Temperatur (Mittel der Jahre 2004 - 2007)

Der Rückgang des Margarinekonsums in den Sommermonaten wird jetzt durch die Temperatur-Variable aufgefangen und zeigt sich in den Koeffizienten der Dummy-Variablen nur noch abgeschwächt. Dagegen wird der starke Einbruch im Januar und Februar, der nicht temperaturbedingt ist, jetzt noch deutlicher.

Modell 3: Trend + saisonale Dummies + Temperatur

$$\begin{aligned}\hat{Y} = 29.582 \quad &- \quad 57,5 \text{ Zeit } (-4,9) * * \qquad (R^2 = 0,891)\\ &- \quad 1.136 \text{ Februar } (-1,5)\\ &+ \quad 3.854 \text{ März } (+4,8) * *\\ &+ \quad 2.579 \text{ April } (+2,3) *\\ &+ \quad 2.142 \text{ Mai } (+1,6)\\ &+ \quad 2.148 \text{ Juni } (+1,3)\\ &+ \quad 1.668 \text{ Juli } (+0,9)\\ &+ \quad 2.663 \text{ August } (+1,6)\\ &+ \quad 4.372 \text{ September } (+2,9) * *\\ &+ \quad 4.938 \text{ Oktober } (+4,2) * *\\ &+ \quad 7.520 \text{ November } (+9,0) * *\\ &+ \quad 3.142 \text{ Dezember } (+4,1) * *\\ &- \quad 227,8 \text{ Temperatur } (-2,4) *\end{aligned}$$

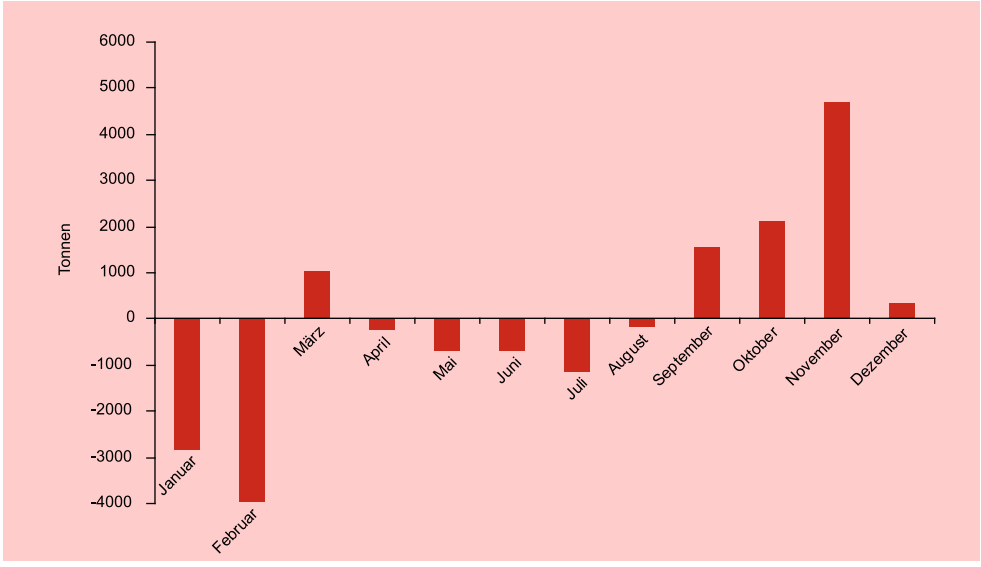


Abbildung 2.22: Saisonfigur im Margarinekonsum nach Elimination des Temperatureinflusses

2.4.2.2.2 Einbeziehung der Bevölkerungsentwicklung

Bevölkerungsgröße

Die Bevölkerungsgröße hat generell maßgeblichen Einfluss auf die Nachfrage nach Konsumgütern. Es soll hier deshalb der Einfluss auf den Margarinekonsum untersucht werden.

Abbildung 2.23 zeigt die Bevölkerungsentwicklung für den hier betrachteten Zeitraum. Während für den Margarinekonsum monatliche Daten vorliegen, sind Bevölkerungsdaten nur jährlich verfügbar, im vorliegenden Fall also nur vier Daten.

Jahr	Anzahl in Mio.
2004	82,500849
2005	82,437995
2006	82,314906
2007	82,208000*

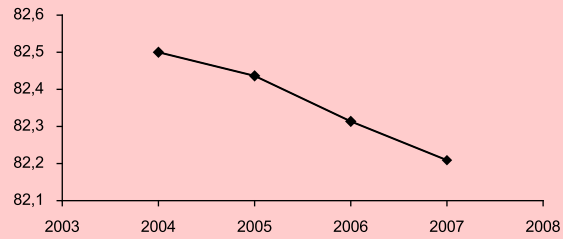


Abbildung 2.23: Bevölkerung Deutschlands. Quelle: Stat. Bundesamt, *Wert geschätzt

Im betrachteten Zeitraum von 2004 bis 2007 besteht eine sehr enge Korrelation zwischen der Variablen Zeit und der Bevölkerung (r = - 0,992), wie sich aus Abbildung 2.23 ersehen lässt. Ersetzt man in Modell 3 die Variable Zeit durch die Bevölkerungszahl, so erhält man folgendes Modell:

Modell 4: Saisonale Dummies + Temperatur + Bevölkerung

$$\begin{aligned}\hat{Y} = -510.754 \quad &- \quad 1.193 \text{ Februar } (-1,5) \qquad (R^2 = 0,884)\\ &+ \quad 3.746 \text{ März } (+4,5) **\\ &+ \quad 2.428 \text{ April } (+2,1) *\\ &+ \quad 1.941 \text{ Mai } (+1,4)\\ &+ \quad 1.899 \text{ Juni } (+1,1)\\ &+ \quad 1.367 \text{ Juli } (+0,7)\\ &+ \quad 2.300 \text{ August } (+1,3)\\ &+ \quad 3.947 \text{ September } (+2,5) *\\ &+ \quad 4.444 \text{ Oktober } (+3,6) **\\ &+ \quad 6.955 \text{ November } (+8,1) **\\ &+ \quad 2.512 \text{ Dezember } (+3,2) **\\ &- \quad 230,3 \text{ Temperatur } (-2,3) *\\ &+ \quad 6.547 \text{ Bevölkerung } (+4,6) *\end{aligned}$$

Der Einfluss der Bevölkerung erweist sich als hoch signifikant. Das Bestimmtheitsmaß verringert sich nur geringfügig von 89,1 % auf 88,4 %.

Die Modelle 3 und 4 eignen sich damit annähernd gleich gut für Prognosen, allerdings nur solange, wie die enge Korrelation zwischen Zeit und Bevölkerung bestehen bleibt. Ist dies nicht der Fall, so wird Modell 4 die besseren Prognosen liefern. Für seine Anwendung werden dann allerdings Prognosen der Bevölkerungsentwicklung benötigt. Mittels Kennziffern zu Fertilität, Sterblichkeit und Migration lassen sich allerdings recht genaue Bevölkerungsprognosen erstellen.[28]

Bevölkerungsprognose

[28] Siehe hierzu „Bevölkerung Deutschlands bis 2050“, Statistisches Bundesamt, Wiesbaden 2006.

2.4.2.3 Zusammenfassung

Die Analyse der Zeitreihe des Marktabsatzvolumens zeigt, dass der Margarinekonsum in Deutschland gekennzeichnet ist durch

- starke **saisonale Schwankungen** und
- einen leicht **abfallenden Trend** (siehe Abbildung 2.18).

Konsumverhalten

Die saisonalen Schwankungen im Margarinekonsum, die sich regelmäßig wiederkehrend im Verlauf eines Jahres zeigen, lassen sich z. T. auf die jahreszeitlichen Temperaturschwankungen (erhöhter Kalorienbedarf bei niedrigen Temperaturen) zurückführen. Allerdings lassen sich mit dem Temperatureinfluss nur knapp 25 % dieser Schwankungen erklären. Sehr viel stärker wirkt offenbar der Einfluss von Weihnachten und Ostern. Im Vorfeld dieser Festlichkeiten zeigt der Margarinekonsum jeweils im November und im März eine starke Zunahme, wobei der Konsum im November der weitaus höchste ist (vgl. Abbildung 2.19). Abbildung 2.22 zeigt die um den Temperatureinfluss bereinigten Schwankungen, die insbesondere den Einfluss von Weihnachten auf den Margarinekonsum deutlich werden lassen.

Der negative Trend im Margarinekonsum lässt sich weitgehend (zu über 90 %) durch den Rückgang der Bevölkerung erklären.

Prognoseeignung

Prognoseeignung der Modelle:

- Modell 1: Linearer Trend
  Infolge der starken Saisonalität im Margarinekonsum ist dieses Modell für monatliche Prognosen ungeeignet.
- Modell 2: Trend + saisonale Dummies
  Dieses Modell ist einfach anzuwenden und wird kurzfristig gute Prognosen liefern.

Temperaturentwicklung

- Modell 3: Trend + saisonale Dummies + Temperatur
  Da die Saisonalität im Margarinekonsum nicht parallel zur Temperaturentwicklung im Jahresablauf verläuft (vgl. Abbildung 2.19 und 2.21) und die Temperatur einen zusätzlichen Erklärungsbeitrag zu den Schwankungen im Margarinekonsum leistet, lässt sich Modell 2 durch Einbeziehung der Temperatur verbessern. Prognosen mit Modell 3 erfordern allerdings, dass dazu für den zu prognostizierenden Monat auch die mittlere Temperatur zu prognostizieren ist, was niemals ganz fehlerfrei gelingen kann.

Bevölkerungsentwicklung

- Modell 4: Saisonale Dummies + Temperatur + Bevölkerung
  Die Trendmodelle 1 bis 3 unterstellen, dass die enge Korrelation zwischen Zeit und Bevölkerungsentwicklung bestehen bleibt, d. h. dass der in den letzten Jahren nahezu linear abfallende Verlauf der Bevölkerung sich fortsetzen wird. Ist dies nicht der Fall, so wird Modell 4 die besseren Prognosen liefern. Bis zum Jahr 2002 war die Bevölkerungsentwicklung in Deutschland noch anwachsend. Die Prognosen des Statistischen Bundesamtes lassen aber erwarten, dass die Bevölkerung Deutschlands auch in den kommenden Jahrzehnten weiterhin schrumpfen wird. Da dieser Prozess nicht linear verlaufen wird, ist für langfristige Prognosen Modell 4 vorzuziehen.

2.4.3 SPSS-Kommandos

Die obigen Zeitreihenanalysen zum Margarinemarkt (Modelle 1 bis 4) wurden mit der Prozedur „REGRESSION“ (Menüpunkt „Analysieren/Regression/Linear“) durchgeführt. In Abbildung 2.24 ist die Syntaxdatei mit den SPSS-Kommandos für die Modelle 1 und 3 wiedergegeben. Zunächst erfolgt die Schätzung des linearen Trendmodells für das Volumen des Margarinemarktes (Modell 1). Daran anschließend erfolgt die Schätzung eines Modells mit Trend, saisonalen Dummies und Temperatureinfluss (Modell 3).[29]

Syntax

```
* MVA: Fallbeispiel Zeitreihenanalyse.
* DATENDEFINITION.
DATA LIST FREE / jahr monat zeit marktvolumen bevölkerung temperatur
  januar februar märz april mai juni juli august september oktober november
  dezember.

BEGIN DATA
2004   1  1  29964,62 82,5 -0,03 1 0 0 0 0 0 0 0 0 0 0 0
2004   2  2  28225,54 82,5  2,63 0 1 0 0 0 0 0 0 0 0 0 0
2004   3  3  30886,65 82,5  4,37 0 0 1 0 0 0 0 0 0 0 0 0
.....
2007 12 48  28702,51 82,2  1,78 0 0 0 0 0 0 0 0 0 0 0 1
END DATA.

* PROZEDUR.
* Zeitregression für das Marktvolumen des Margarinemarktes.
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT marktvolumen
  /METHOD=ENTER zeit
  /RESIDUALS DURBIN
  /SAVE PRED.

* Zeitregression für das Marktvolumen des Margarinemarktes mit Dummies und Temperatur.
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT marktvolumen
  /METHOD=ENTER zeit temperatur januar februar märz april mai juni juli august september oktober november
  dezember
  /RESIDUALS DURBIN
  /SAVE PRED.
```

Abbildung 2.24: SPSS-Job zur Zeitreihenanalyse

2.5 Anwendungsempfehlungen

Der Wunsch, in die Zukunft zu blicken, ist ein alter Menschheitstraum. Hiervon zeugen u.a. die hohe Anerkennung der biblischen Propheten oder der Ruhm des Orakels von Delphi, aber auch die heutige Popularität von Horoskopen und Wahrsagern. Der Begriff der hier behandelten Prognose ist allerdings streng zu trennen von anderen Aussagen über die Zukunft, wie Hellseherei, Prophetie oder Utopie.

Prognosen sind Aussagen über die Zukunft, die auf Informationen gestützt und um Objektivität bemüht sind. Es wäre aber eine Illusion, wollte man annehmen,

[29] Auf eine Wiedergabe der SPSS-Outputs sei hier verzichtet, da die Ergebnisse bereits im obigen Text enthalten sind. Zur allgemeinen Erläuterung der SPSS-Outputs vgl. die Ausführungen im vorhergehenden Kapitel zur Regressionsanalyse.

Prognose

dass Prognosen frei von Subjektivität sein können. Vielmehr sollten sie durch das Bemühen gekennzeichnet sein, die Subjektivität durch Verwendung formaler Modelle und Methoden einzuschränken.

Urteil des Untersuchers

Sowohl die Auswahl eines Prognoseverfahrens (wie hier z. B. die Regressionsanalyse) wie auch die Modellspezifikation sind abhängig vom Urteil des Untersuchers. Es wurden diverse empirische Vergleiche von alternativen Prognoseverfahren durchgeführt, um das beste Verfahren zu ermitteln. Hinsichtlich des größten deutschen Vergleichs, des „DGOR-Prognosevergleichs“[30], bemerkt Hüttner (1994), S. 351, dass für den „Sieg“ eher die Personen als die Verfahren entscheidend waren.

Für die Erstellung von Prognosen seien hier einige Hinweise gegeben:[31]

1. Das Prognoseproblem sollte immer in Verbindung mit dem dahinter stehenden Entscheidungsproblem gesehen werden. Ist vom Ergebnis der Prognose keine Entscheidung abhängig, braucht auch keine Prognose erstellt zu werden.

2. Es sollte überlegt werden, ob die Anwendung eines formalen Prognoseverfahrens Vorteile bringt. Dies ist z. B. zu verneinen, wenn es um die kurzfristige Prognose von Aktienkursen geht oder um die Prognose einer volkswirtschaftlichen Entwicklung, die maßgeblich durch politische Entscheidungen beeinflusst wird.

3. Es muss geklärt werden, welche Daten benötigt und welche Datenquellen verwendet werden sollen. Bei Wirkungsprognosen müssen die relevanten Prädiktorvariablen identifiziert werden. Neben theoretischen Überlegungen ist hier oft auch Kreativität gefordert.

4. Die Länge der Zeitreihe sollte dem Prognosehorizont angepasst sein. Langfristige Prognosen erfordern lange Zeitreihen, während für kurzfristige Prognosen kürzere Zeitreihen hinreichend sind.

5. Vor der Durchführung von Analysen sollten die Daten visualisiert werden, damit sich der Untersucher vom Verlauf der Zeitreihe ein Bild machen kann.

6. Für die Spezifikation eines Modells sind gesunder Menschenverstand, Sachkenntnis und theoretische Überlegungen genau so wichtig wie Methodenkenntnisse. Der Untersucher sollte z. B. eine Vorstellung davon haben, welches Prognoseergebnis zu erwarten ist bzw. welche Ergebnisse plausibel sind (z. B. progressives oder degressives Wachstum, Schrumpfung).

7. Bei der Auswahl von Variablen oder der Auswahl eines Modelltyps sollte man sich nicht allein von statistischen Kriterien leiten lassen. Ein guter Fit im Stützbereich ist keine Gewähr für Prognosegüte. Häufig werden Indikatoren nicht auf Basis einer kausalen Beziehung zur Prognosevariablen ausgewählt, sondern auf Basis ihrer Korrelation mit der Prognosevariablen. Diese aber kann zufällig bedingt und somit bedeutungslos für die Zukunft sein.

8. Kausale Zusammenhänge sollten genutzt werden, wenn immer möglich. Zu bedenken ist aber, dass in Strukturmodellen die Prognosen der Wirkungen auch Prognosen der Ursachen erfordern.

[30] Vgl. dazu Schwarze/Weckerle (1982).
[31] Vgl. dazu auch Armstrong, J. (2002), S. 680 ff.

9. Simplen Modellen ist gegenüber komplexen Modellen der Vorzug zu geben, solange empirische Befunde nicht dagegen sprechen.
10. Neben Punktprognosen sollten Intervallprognosen erstellt werden, damit das Risiko der mit der Prognose verbundenen Entscheidung abgeschätzt werden kann.
11. Die Schätzung des Modells sollte aktualisiert werden, sobald neue Daten verfügbar sind.
12. Die Verwendung alternativer Datenquellen sowie alternativer Prognoseverfahren und deren Kombination hat sich als nützlich erwiesen.[32]

2.6 Mathematischer Anhang

Nachfolgend sei unter Verwendung der Matrizenschreibweise die Berechnung des Multivariaten Prognosefehlers angegeben.

Prognose für Periode t': Prognosefehler

$$\hat{y}_{t'} = b_0 + \sum_{j=1}^{J} b_j x_{t'j}$$

Prognosefehler:

$$s_p(t') = s\sqrt{1 + \frac{1}{T} + (\mathbf{x}_{t'} - \overline{\mathbf{x}})'(\dot{\mathbf{X}}'\dot{\mathbf{X}})^{-1}(\mathbf{x}_{t'} - \overline{\mathbf{x}})}$$

mit

s = Standardfehler der Regression
$\mathbf{x}_{t'}$ = J-Vektor der Werte der Prädiktorvariablen für Periode t'
$\overline{\mathbf{x}}$ = J-Vektor der Mittelwerte der Prädiktorvariablen
$\dot{\mathbf{X}}$ = (K x J)-Matrix mit den Abweichungen der Prädiktorvariablen von ihrem Mittelwert

Prognoseintervall für Vertrauenswahrscheinlichkeit $1 - \alpha$: Prognoseintervall

$$\hat{y}_{t'} - t_{\alpha/2} s_p \leq y_{t'} \leq \hat{y}_{t'} + t_{\alpha/2} s_p$$

mit

$t_{\alpha/2}$ = Quantil der Verteilung (Student-Verteilung) für die Vertrauenswahrscheinlichkeit $1 - \alpha$ bei zweiseitigem Test und $T - J$ Freiheitsgraden.

Wurden nichtlineare Transformationen f_j (j = 1, ..., J) der Prädiktoren bei der Regressionsrechnung berücksichtigt, so sind die entsprechend transformierten Prädiktoren in die Prognosefunktion einzusetzen. Wurde auch die abhängige Variable transformiert, so ergibt sich allgemein der Prognosewert durch: Transformation

$$z = b_0 + \sum_{j=1}^{J} b_j f(x_{t'j})$$

[32] Vgl. dazu Armstrong, J. (2002), S. 417 ff., Bates/Granger (1969), Hüttner (1994).

$$\hat{y}_{t'} = f_0^{-1}(z)$$

Das Prognoseintervall ergibt sich dann durch

$$f_0^{-1}(\hat{y}_{t'} - t_{\alpha/2} s_p) \leq y_{t'} \leq f_0^{-1}(\hat{y}_{t'} + t_{\alpha/2} s_p)$$

Das Prognoseintervall ist bei nichtlinearer Transformation der abhängigen Variablen asymmetrisch.

Literaturhinweise

A. Basisliteratur zur Zeitreihenanalyse

Armstrong, J. (Hrsg.) (2002), Principles of Forecasting, Boston (MA) u. a.

Hamilton, J. (1994), Time Series Analysis, Princeton.

Hanke, J. E./Wichern, D. W. (2009), Business Forecasting, 9. Auflage, Upper Saddle River (N.J.) u. a.

Makridakis, S./Wheelwright, S./ Hyndman, R. (1998), Forecasting: Methods and Applications, 3. Auflage, New York u. a.

Mertens, P./Rässler, S. (Hrsg.) (2012), Prognoserechnung, 7. Auflage, Heidelberg.

Rinne, H./Specht, K. (2002), Zeitreihen: Statistische Modellierung, Schätzung und Prognose, München.

B. Zitierte Literatur

Armstrong, J. (2002), Principles of Forecasting, Boston (MA) u. a.

Backhaus, K./Erichson, B./Weiber, R. (2015), Fortgeschrittene Multivariate Analyseverfahren, 3. Auflage, Berlin/Heidelberg.

Backhaus, K./Simon, W. (1981), Indikatorprognosen im Investitionsgüter-Marketing, in: *Die Betriebswirtschaft*, Vol. 41, S. 419–422.

Bates, J./Granger, C. (1969), The Combination of Forecasts, in: *Operational Research Quarterly*, Vol. 20, Nr. 4, S. 451–468.

Greene, W. (2018), Econometric Analysis, 8. Auflage, Essex.

Hamilton, J. (1994), Time Series Analysis, Princeton.

Hammann, P./Erichson, B. (2000), Marktforschung, 4. Auflage, Stuttgart.

Hanke, J./Reitsch, A. (2009), Business Forecasting, 9. Auflage, Upper Saddle River (N.J.) u.a.

Hanssens, D./Parsons, L./Schultz, R. (2003), Market Response Models. Econometric and Time Series Analysis, 2. Auflage, Boston (MA) u. a.

Hüttner, M. (1994), Vergleich und Auswahl von Prognoseverfahren für betriebswirtschaftliche Zwecke, in: Mertens, P. (Hrsg.), Prognoserechnung, 5. Auflage, Heidelberg, S. 349–363.

IBM Cooperation (o.J.), SPSS Statistics Base 23, Chicago.

Kmenta, J. (1997), Elements of Econometrics, 2. Auflage, New York.

Lambin, L. (1969), Measuring the Profitability of Advertising: An Empirical Study, in: *Journal of Industrial Economics*, Vol. 17, Nr. 2, S. 86–103.

Mahajan, V./Muller, E./Bass, F. (1990), New Product Diffusion in Marketing: A Review and Directions for Research, in: *Journal of Marketing*, Vol. 54, Nr. 1, S. 1–26.

Makridakis, S./Wheelwright, S./Hyndman, R. (1998), Forecasting: Methods and Applications, 3. Auflage, New York u. a.

Meffert, H./Steffenhagen, H. (1977), Marketing-Prognosemodelle, Stuttgart.

Mertens, P. (1994), Prognoserechnung, 5. Auflage, Heidelberg.

Niederhübner, N. (1994), Indikatorprognosen, in: Mertens, P. (Hrsg.), Prognoserechnung, 5. Auflage, Heidelberg, S. 205–212.

P. Mertens and S. Rässler (2012), Prognoserechnung, 7. Auflage, Heidelberg.

Palda, K. (1984), The Measurement of Cumulative Advertising Effects, Englewood Cliffs (N.J.).

Ramanathan, R. (2002), Introductory Econometrics with Applications, 5. Auflage, Fort Worth (TX).

Rinne, H./Specht, K. (2002), Zeitreihen: Statistische Modellierung, Schätzung und Prognose, München.

Schneeberger, H. (1994), Punkt-, Intervallprognose und Test auf Strukturbruch, in: Mertens, P. (Hrsg.), Prognoserechnung, 5. Auflage, Heidelberg.

Schneeweiß, H. (1990), Ökonometrie, 4. Auflage, Heidelberg.

Schwarze, J./Weckerle, J. (1982), Prognoseverfahren im Vergleich, Braunschweig.

Studenmund, A. (2017), Using Econometrics: A Practical Guide, 7. Auflage, Boston (MA).

Weiber, R. (1992), Diffusion von Telekommunikation, Wiesbaden.

Wooldridge, J. (2016), Introductory Econometrics: A modern Approach, 6. Auflage, Boston.

3 Varianzanalyse

K. Backhaus, B. Erichson, W. Plinke, R. Weiber, *Multivariate Analysemethoden*,
https://doi.org/10.1007/978-3-662-56655-8_4

3.1 Problemstellung

Die Varianzanalyse ist ein Verfahren, das die Wirkung einer (oder mehrerer) unabhängiger Variablen auf eine (oder mehrere) abhängige Variablen untersucht. Für die unabhängige Variable wird dabei lediglich Nominalskalierung verlangt, während die abhängige Variable metrisches Skalenniveau aufweisen muss. Die Varianzanalyse ist das wichtigste Analyseverfahren zur Auswertung von *Experimenten*. Typische Anwendungsbeispiele zeigt Abbildung 3.1.

1.	Welche Wirkung haben verschiedene Formen der Bekanntmachung eines Kinoprogramms (z. B. Internet, Plakat, Zeitungsannonce) auf die Besucherzahlen? Um dieses zu erfahren, wendet ein Kinobesitzer eine Zeit lang jeweils nur eine Form der Bekanntmachung an.
2.	Welche Wirkung haben zwei Marketinginstrumente jeweils isoliert und gemeinsam auf die Zielvariable? Ein Konfitürenhersteller geht z. B. von der Vermutung aus, dass der Markenname und der Absatzweg einen wichtigen Einfluss auf den Absatz haben. Deshalb testet er drei verschiedene Markennamen in zwei verschiedenen Absatzwegen.
3.	Es soll die Wahrnehmung von Konsumenten untersucht werden, die sie gegenüber zwei alternativen Verpackungsformen für die gleiche Seife empfinden. Deshalb werden die Probanden gebeten, auf drei Ratingskalen die Attraktivität der Verpackung, die Gesamtbeurteilung des Produktes und ihre Kaufbereitschaft anzugeben.
4.	Ein Landwirtschaftsbetrieb will die Wirksamkeit von drei verschiedenen Düngemitteln im Zusammenhang mit der Bodenqualität überprüfen. Dazu werden der Ernteertrag und die Halmlänge bei gegebener Getreidegattung auf Feldern verschiedener Bodenbeschaffenheit, die jeweils drei verschiedene Düngesegmente haben, untersucht.
5.	In einer medizinischen Querschnittsuntersuchung wird der Einfluss unterschiedlicher Diäten auf das Körpergewicht festgestellt.
6.	In mehreren Schulklassen der gleichen Ausbildungsstufe wird der Lernerfolg verschiedener Unterrichtsmethoden festgestellt.

Abbildung 3.1: Anwendungsbeispiele

Gemeinsam ist allen Beispielen, dass ihnen eine *Vermutung über die Wirkungsrichtung* der Variablen zugrunde liegt. Wie in der Regressionsanalyse, die einen Erklärungszusammenhang der Art

$$Y = f(X_1, X_2, ..., X_j, ..., X_J) \tag{3.1}$$

über metrische Variable herstellt, formuliert auch die Varianzanalyse einen solchen Zusammenhang, allein mit dem Unterschied, dass es sich bei den Variablen $X_1, X_2, ..., X_J$ um kategoriale Variablen handelt. Die genannten Beispiele verdeutlichen dies: So wird im ersten Beispiel angenommen, dass die Werbung als unabhängige Variable mit den drei Ausprägungen „Internet“, „Plakat“ und „Zeitungsannonce“ einen Einfluss auf die Zahl der Kinobesucher hat. Die Ausprägungen der unabhängigen Variablen beschreiben dabei stets alternative Zustände. Demgegenüber ist die abhängige Variable, hier die Zahl der Kinobesucher, metrisch skaliert.

Skalenniveau

Gemeinsam ist weiterhin allen Anwendungsbeispielen, dass sie experimentelle Situationen beschreiben: Feldexperimente z. B. im ersten und zweiten, ein Laborexperiment im dritten Fall. Die Varianzanalyse ist das klassische Verfahren zur Analyse von Experimenten mit Variablen des bezeichneten Skalenniveaus.

Experimente

Die genannten Beispiele unterscheiden sich durch die Zahl der Variablen. So wird im ersten Beispiel die Wirkung *einer* unabhängigen Variablen (Werbeart) auf *eine* abhängige Variable (Besucherzahl) untersucht. Im zweiten Beispiel wird demgegenüber die Wirkung von *zwei* unabhängigen Variablen (Markenname und Absatzweg)

auf *eine* abhängige Variable (Absatz) analysiert. Im dritten Beispiel gilt das Interesse schließlich der Wirkung einer unabhängigen Variablen (Verpackungsform) auf drei abhängige Variable (Attraktivität der Verpackung, Gesamtbeurteilung des Produktes und Kaufbereitschaft).

Die abhängige Variable wird in der Varianzanalyse auch *Zielvariable* genannt und die unabhängigen kategorialen Variablen werden als *Faktoren* oder *Treatments* bezeichnet. Die Ausprägungen der Faktoren werden als *Faktorstufen* (Kategorien, Gruppen) bezeichnet. Die unterschiedlichen Typen der Varianzanalyse lassen sich nach der Zahl der Faktoren differenzieren. Wenn nur *eine* abhängige Variable betrachtet wird, so wird generell von *univariater Varianzanalyse* gesprochen (sog. ANOVA; Analysis of Variance). Je nach Anzahl der unabhängigen Variablen (Faktoren) kann die Bezeichnung dann konkretisiert werden in einfaktorieller, zweifaktorieller usw. Varianzanalyse. Varianzanalysen mit mehr als einer *abhängigen* Variablen werden als multivariate oder mehrdimensionale Varianzanalysen bezeichnet (sog. MANOVA; Mutivariate Analysis of Variance). Abbildung 3.2 gibt einen Überblick.

Faktoren (Treatments)
Faktorstufen

Zahl der abhängigen Variablen	Zahl der unabhängigen Variablen	Bezeichnung des Verfahrens
Univariate Varianzanalysen (ANOVA)		
1	1	Einfaktorielle oder einfache Varianzanalyse
1	zwei oder mehrere	Mehrfaktorielle oder multiple Varianzanalyse
Multivariate Varianzanalysen (MANOVA)		
Mindestens zwei	eine oder mehrere	

Abbildung 3.2: Typen der Varianzanalyse

3.2 Vorgehensweise

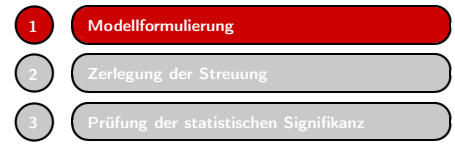
Das Grundprinzip der Varianzanalyse wird im Folgenden zunächst am Beispiel einer univariaten Varianzanalyse mit einer abhängigen und einer unabhängigen Variablen (einfaktorielles Modell) verdeutlicht. Dabei folgen wir einer Vorgehensweise in drei Schritten, die in Abbildung 3.3 dargestellt ist. Im Verlauf erweitern wir unsere Überlegungen auf die zweifaktorielle Varianzanalyse (zwei unabhängige Variable), wobei wir die schrittweise Gliederung des Ablaufs beibehalten. Abschließend werden ausgewählte Erweiterungen im Verfahren der Varianzanalyse besprochen.

ANOVA

Abbildung 3.3: Ablaufschritte der Varianzanalyse

3.2.1 Einfaktorielle Varianzanalyse

3.2.1.1 Modellformulierung



Um den Kern der Varianzanalyse herauszuarbeiten, betrachten wir zunächst die folgende Problemsituation: Der Leiter einer Supermarktkette will die Wirkung verschiedener Arten der Warenplatzierung überprüfen. Er wählt dazu Margarine in der Becherverpackung aus, wobei ihm drei Möglichkeiten der Regalplatzierung offen stehen:

1. Platzierung im Normalregal der Frischwarenabteilung
2. Platzierung im Normalregal der Frischwarenabteilung und Zweitplatzierung im Fleischmarkt
3. Platzierung im Kühlregal der Frischwarenabteilung.

Einfaktorielles Design

Anschließend wird folgendes experimentelle Design entworfen: Aus den insgesamt vorhandenen Supermärkten werden drei weitgehend vergleichbare Supermärkte des Unternehmens ausgewählt. In einem Zeitraum von 5 Tagen wird in jedem der drei Supermärkte jeweils eine Form der Margarine-Präsentation durchgeführt. Die Auswirkungen der Maßnahmen werden jeweils in der Größe „kg Margarineabsatz pro 1.000 Kassenvorgänge“ erfasst. Abbildung 3.4 zeigt die Ergebnisse. Sind die Teilstichproben (Gruppen) gleich groß, so erleichtert das die Analyse. Man spricht dann von einem *balancierten Design*.

	Ausgangsdaten y_{gk}				
	Tag 1	Tag 2	Tag 3	Tag 4	Tag 5
Supermarkt 1 ,,Normalregal “	47	39	40	46	45
Supermarkt 2 ,,Zweitplatzierung “	68	65	63	59	67
Supermarkt 3 ,,Kühlregal “	59	50	51	48	53

Abbildung 3.4: Ausgangsdaten im Fallbeispiel – Margarineabsatz in drei Supermärkten in Abhängigkeit der Regalplatzierung

Boxplot

Wir erhalten drei Teilstichproben mit jeweils fünf Beobachtungswerten. Am Anfang einer Analyse sollte immer eine Veranschaulichung der Daten stehen. Zum Vergleich mehrerer Stichproben eignen sich insbesondere Boxplots, wie sie Abbildung 3.5 für die vorliegenden Daten zeigt. Jeder Boxplot kennzeichnet eine der drei Teilstichproben und zeigt deren Lage und Streuung an. Die horizontale Linie zeigt die Lage des Medians und die Box gibt den Bereich an, in dem 50% der Beobachtungen liegen. Die Antennen (whiskers) markieren die Spannweite der Daten (maximaler und minimaler Wert) mit Ausnahme von Ausreißern. Im rechten Boxplot für das Kühlregal wird die Beobachtung Nr. 11 (y = 59) als ausreißerverdächtig angezeigt.

Aus Abbildung 3.5 ist ersichtlich, dass deutliche Unterschiede bezüglich der Absatzmengen in den drei Gruppen bestehen, und es ist augenscheinlich, dass diese Unterschiede durch die unterschiedlichen Regalplatzierungen bedingt sind. Man kann

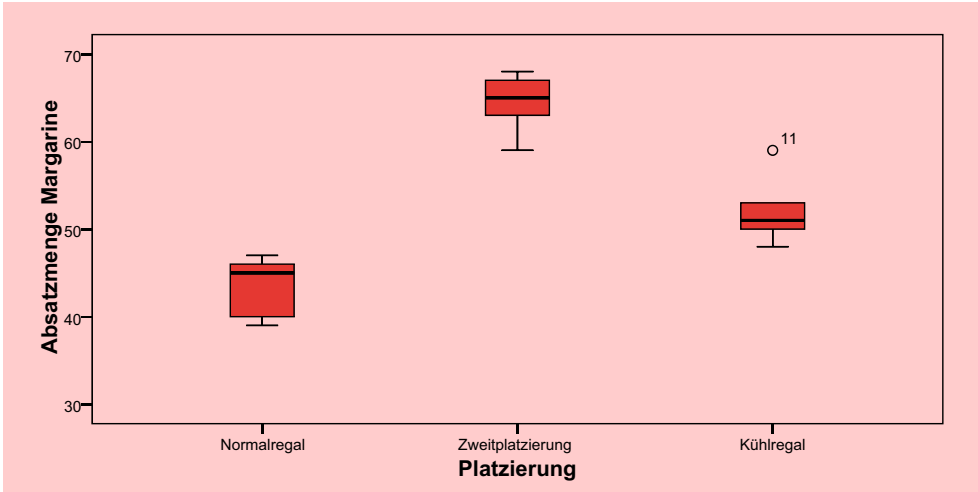


Abbildung 3.5: Boxplots der Daten

	Mittelwert pro Supermarkt
Supermarkt 1 „Normalregal“	$\overline{y}_1 = 43,4$
Supermarkt 2 „Zweitplatzierung“	$\overline{y}_2 = 64,4$
Supermarkt 3 „Kühlregal“	$\overline{y}_3 = 52,2$
Gesamtmittelwert	$\overline{y} = 53,33$

Abbildung 3.6: Mittelwerte des Margarineabsatzes in drei Supermärkten

daraus folgern, dass die Art der Regalplatzierung einen Einfluss auf den Margarineabsatz hat. Vorerst aber wollen wir dies nur als eine Hypothese betrachten, die nachfolgend mittels Varianzanalyse geprüft werden soll.

Die Mittelwerte der Gruppen und den Gesamtmittelwert der Daten zeigt Abbildung 3.6. Diese Werte werden für die Durchführung der Varianzanalyse benötigt. Man kann auch sagen, dass die Varianzanalyse die Differenzen zwischen den Mittelwerten analysiert. Dabei spielen die Varianzen der Beobachtungswerte um diese Mittelwerte eine entscheidende Rolle.

Da die Absatzmengen auch bei gleicher Regalplatzierung voneinander abweichen, müssen neben der Regalplatzierung noch andere Einflussgrößen vorhanden sein. Im Marktgeschehen gibt es immer vielfältige Einflüsse, die sich größtenteils nicht beobachten lassen. Diese werden im Modell der Varianzanalyse durch Zufallsgrößen (Störgrößen) ϵ_{gk} repräsentiert, die in jeder Beobachtung y_{gk} enthalten sind.[1] Die Störgrößen überlagern das Geschehen und machen eine statistische Analyse erforderlich. Das Modell der Varianzanalyse unterscheidet daher (wie auch die Regressionsanalyse) zwischen einer systematischen und einer stochastischen Komponente. Zwei alternative Formulierungen sind möglich.

[1] Dies erfolgt ganz analog zum Modell der Regressionsanalyse in Kapitel 1.

Modellformulierung A

$$y_{gk} = \mu_g + \epsilon_{gk} \tag{3.2}$$

mit:

y_{gk} = Beobachtungswert k ($k = 1, 2, ..., K$) in Faktorstufe g ($g = 1, 2, ..., G$)
μ_g = Mittelwert für Faktorstufe g in der Grundgesamtheit (Erwartungswert)
ϵ_{gk} = Störgrößen

Die unbekannten wahren Mittelwerte μ_g lassen sich durch die Mittelwerte der Beobachtungswerte in Abbildung 3.6 schätzen. In den Abweichungen zwischen den Mittelwerten der Faktorstufen (Gruppen) schlagen sich die Effekte der unterschiedlichen Regalplatzierung nieder (systematische Komponente).

Bezüglich der Störgrößen ϵ_{gk} wird wie üblich angenommen, dass sie voneinander unabhängig und normalverteilt sind mit gleicher Varianz.[2] Es wird also angenommen, dass in den Gruppen ungefähr gleich starke Störeinflüsse herrschen (stochastische Komponente).

Modellformulierung B

$$y_{gk} = \mu + \alpha_g + \epsilon_{gk} \tag{3.3}$$

mit:

μ = Gesamtmittelwert in der Grundgesamtheit (globaler Erwartungswert)
α_g = wahrer Effekt von Faktorstufe g ($g = 1, 2, ..., G$)

Es gilt: $\alpha_g = \mu_g - \mu$

Die zweite Modellformulierung enthält die Effekte in expliziter Form. Man spricht daher auch vom *Modell in Effektdarstellung*. Sie ergeben sich aus den Abweichungen der Gruppenmittelwerte vom Gesamtmittelwert. Da wir im Modell B jetzt $G + 1$ unbekannte Parameter bei nur G Kategorien haben, ist zwecks eindeutiger Bestimmung (Identifizierbarkeit) eine Nebenbedingung (Reparametrisierungsbedingung) erforderlich, z.B.:

$$\sum_{g=1}^{G} \alpha_g = 0 \tag{3.4}$$

Es wird also angenommen, dass die Effekte sich gegenseitig ausgleichen, wodurch lediglich deren Skalierung tangiert wird. Alternativ könnte man auch eine der Kategorien als Referenzkategorie wählen und deren Effekt null setzen.

[2] Die Annahme der Normalverteilung lässt sich durch den „zentralen Grenzwertsatz“ der Statistik stützen, da die Störgrößen meist die gemeinsame Wirkung sehr vieler und im Einzelnen relativ unbedeutender Einflussfaktoren repräsentieren, die voneinander weitgehend unabhängig sind.

Die Effekte lassen sich wie folgt schätzen. Sei a_g der Schätzwert für α_g, so gilt:

$$\mathrm{a}_g = (\bar{\mathrm{y}}_\mathrm{g} - \bar{\mathrm{y}}) \tag{3.5}$$

mit

$$\bar{y}_g = \frac{1}{K}\sum_{k=1}^{K} y_{gk} \qquad \text{Gruppenmittelwerte} \tag{3.6}$$

$$\bar{y} = \frac{1}{G \cdot K}\sum_{g=1}^{G}\sum_{k=1}^{K} y_{gk} \qquad \text{Gesamtmittelwert} \tag{3.7}$$

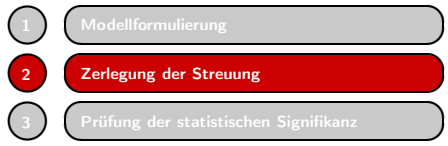
Es ergibt sich hier mit den Werten aus Abbildung 3.6:

$$\mathrm{a}_1 = (\bar{\mathrm{y}}_1 - \bar{\mathrm{y}}) = 43,40 - 53,33 = -9,93 \qquad \text{Normalregal}$$
$$\mathrm{a}_2 = (\bar{\mathrm{y}}_2 - \bar{\mathrm{y}}) = 64,40 - 53,33 = 11,07 \qquad \text{Zweitplatzierung}$$
$$\mathrm{a}_3 = (\bar{\mathrm{y}}_3 - \bar{\mathrm{y}}) = 52,20 - 53,33 = -1,33 \qquad \text{Kühlregal}$$

Die Summe der Effekte ist (bis auf Rundungsfehler) null. Den stärksten positiven Effekt erbringt die Zweitplatzierung während der Abverkauf bei Platzierung 1 (Normalregal) am geringsten ist. Das alles ist fast trivial. Mit der Bildung einiger Mittelwerte haben wir hier das Schätzproblem erledigt. Bei der Regressionsanalyse hatten wir dafür die Kleinst-Quadrate-Methode verwenden müssen. Aber auch der Mittelwert ist ein Kleinst-Quadrate-Schätzwert.[3]

Was wir noch nicht erledigt haben, ist die Frage, ob die ermittelten Effekte wirklich durch die Art der Regalplatzierung verursacht wurden. Wegen des Vorhandenseins von nicht beobachtbaren Einflussgrößen (Störgrößen) könnten möglicherweise die geschätzten Effekte auch rein zufällig entstanden sein. In der Klärung dieser Frage liegt das Kernproblem der Varianzanalyse. Hierzu ist eine Streuungszerlegung vorzunehmen, wie sie schon bei der Regressionsanalyse in Kapitel 1 durchgeführt wurde.

3.2.1.2 Zerlegung der Streuung



Das Prinzip der Varianzanalyse basiert auf einer Zerlegung der Abweichungen zwischen den beobachteten Werten y_{gk} und dem Gesamtmittelwert $\bar{y}$. Diese lassen sich jeweils aufspalten in einen systematischen Teil, der sich durch die Warenplatzierung erklären lässt, und einen nicht erklärbaren Teil, der zufällig bedingt ist. Abbildung 3.7 verdeutlicht dies. Betrachten wir z.B. den ersten Beobachtungswert der Gruppe 2 (Zweitplatzierung mit $y_{21} = 68$).

[3] Auch hier ließe sich die Schätzung mittels Regressionsanalyse unter Verwendung von Dummy-Variablen durchführen. Zur Schätzung von Modell A wären drei Dummy-Variablen für die Arten der Regalplatzierung zu bilden und ein Modell ohne konstantes Glied zu wählen: $y_i = \mu_1 x_{i1} + \mu_2 x_{i2} + ... + \mu_G x_{iG} + \varepsilon_i$ mit $x_{ig} = 1$ wenn Beobachtung i in Gruppe g fällt und 0 sonst $(i = 1, ..., G \cdot K)$.
Der Rechenaufwand wäre allerdings größer. Letztlich aber bildet die Varianzanalyse nur einen Spezialfall der Regressionsanalyse.

Die Abweichung vom Gesamtmittelwert beträgt $y_{21} - \bar{y} = 68 - 53,3 = 14,7$. Davon lässt sich die Abweichung $\bar{y}_2 - \bar{y} = a_2 = 11,1$ durch den Effekt der Regalplatzierung erklären, nicht aber die Abweichung $y_{21} - \bar{y}_2 = 68 - 64,4 = 3,6$. Es gilt:

$y_{21} - \bar{y}$	$= \bar{y}_2 - \bar{y}$	$+ y_{21} - \bar{y}_2$
$14,7$	$= 11,1$	$+3,6$
Gesamtabweichung	= erklärte Abweichung	+ nicht erklärte Abweichung.

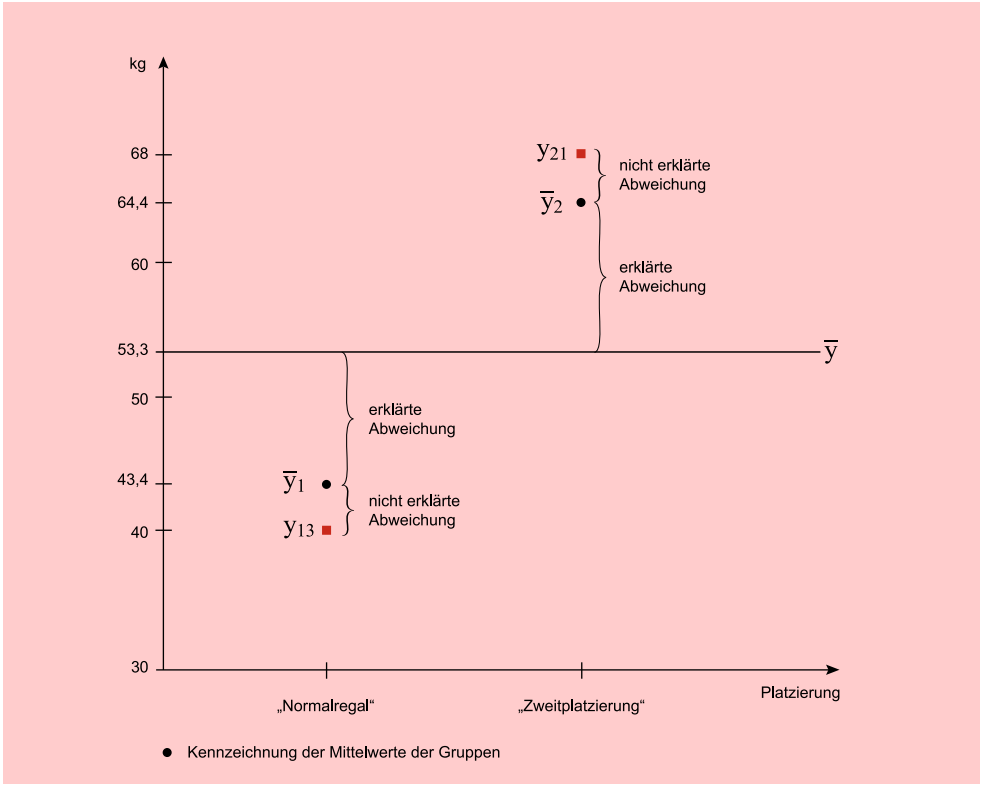


Abbildung 3.7: Erklärte und nicht erklärte Abweichungen bei „Normalregal“ und „Zweitplatzierung“ (y_{gk} aus Abbildung 3.4)

Die obige Gleichung gilt auch, wenn man die Elemente quadriert und über die Beobachtungen aufsummiert (SS = „sum of squares“). Man erhält damit die folgende *Zerlegung der Gesamtstreuung*:

Zerlegung der Gesamtstreuung

Gesamtstreuung	= erklärte Streuung	+ nicht erklärte Streuung
$\sum_{g=1}^{G} \sum_{k=1}^{K} (y_{gk} - \bar{y})^2$	$\sum_{g=1}^{G} K(\bar{y}_g - \bar{y})^2$	$\sum_{g=1}^{G} \sum_{k=1}^{K} (y_{gk} - \bar{y}_g)^2$
$SS_{t(otal)}$	$= SS_{b(etween)}$	$+ SS_{w(ithin)}$

Die Anwendung auf den Datensatz in Abbildung 3.4 ergibt die Datentabelle in Abbildung 3.8.

	SS_t $\sum_{g=1}^{G}\sum_{k=1}^{K}(y_{gk}-\overline{y})^2$	SS_b $\sum_{g=1}^{G}K(\overline{y_g}-\overline{y})^2$	SS_w $\sum_{g=1}^{G}\sum_{k=1}^{K}(y_{gk}-\overline{y}_g)^2$
Normalregal	$(47-53,\overline{3})^2 = 40,11$	$(43,4-53,\overline{3})^2 = 98,67$	$(47-43,4)^2 = 12,96$
	$+(39-53,\overline{3})^2 = 205,44$	$+(43,4-53,\overline{3})^2 = 98,67$	$(39-43,4)^2 = 19,36$
	$+(40-53,\overline{3})^2 = 177,78$	$+(43,4-53,\overline{3})^2 = 98,67$	$(40-43,4)^2 = 11,56$
	$+(46-53,\overline{3})^2 = 53,78$	$+(43,4-53,\overline{3})^2 = 98,67$	$(46-43,4)^2 = 6,76$
	$+(45-53,\overline{3})^2 = 69,44$	$+(43,4-53,\overline{3})^2 = 98,67$	$(45-43,4)^2 = 2,56$
Zweitplatzierung	$+(68-53,\overline{3})^2 = 215,11$	$+(64,4-53,\overline{3})^2 = 122,47$	$(68-64,4)^2 = 12,96$
	$+(65-53,\overline{3})^2 = 136,11$	$+(64,4-53,\overline{3})^2 = 122,47$	$(65-64,4)^2 = 0,36$
	$+(63-53,\overline{3})^2 = 93,44$	$+(64,4-53,\overline{3})^2 = 122,47$	$(63-64,4)^2 = 1,96$
	$+(59-53,\overline{3})^2 = 32,11$	$+(64,4-53,\overline{3})^2 = 122,47$	$(59-64,4)^2 = 29,16$
	$+(67-53,\overline{3})^2 = 186,78$	$+(64,4-53,\overline{3})^2 = 122,47$	$(67-64,4)^2 = 6,76$
Kühlregal	$+(59-53,\overline{3})^2 = 32,11$	$+(52,2-53,\overline{3})^2 = 1,28$	$(59-52,2)^2 = 46,24$
	$+(50-53,\overline{3})^2 = 11,11$	$+(52,2-53,\overline{3})^2 = 1,28$	$(50-52,2)^2 = 4,84$
	$+(51-53,\overline{3})^2 = 5,44$	$+(52,2-53,\overline{3})^2 = 1,28$	$(51-52,2)^2 = 1,44$
	$+(48-53,\overline{3})^2 = 28,44$	$+(52,2-53,\overline{3})^2 = 1,28$	$(48-52,2)^2 = 17,64$
	$+(53-53,\overline{3})^2 = 0,11$	$+(52,2-53,\overline{3})^2 = 1,28$	$(53-52,2)^2 = 0,64$
	$SS_t = 1287,33$	$SS_b = 1112,13$	$SS_w = 175,20$

Abbildung 3.8: Ermittlung der Abweichungsquadrate (sum of squares)

Mittels dieser Streuungszerlegung ist es jetzt einfach, die Güte des Modells zu beurteilen. Wir berechnen dazu, welcher Anteil der gesamten Streuung durch das Modell bzw. durch die Regalplatzierung erklärt wird:

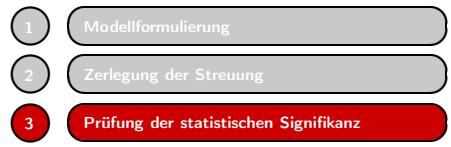
Eta-Quadrat

$$\text{Eta-Quadrat} = \frac{\text{erklärte Streuung}}{\text{gesamte Streuung}} = \frac{SS_b}{SS_t} = \frac{1.112,13}{1.287,33} = 0,864 \tag{3.8}$$

Mehr als 86% der Streuung in den Absatzmengen lassen sich also durch die unterschiedlichen Regalplatzierungen erklären. Nur knapp 14% bleiben unerklärt und müssen auf Störeinflüsse zurückgeführt werden. Eta-Quadrat ist eine normierte Größe, deren Wertebereich zwischen null und eins liegt. Es ist umso größer, je höher der Anteil der erklärten Streuung an der Gesamtstreuung ist. Es entspricht dem R-Quadrat der Regressionsanalyse.[4]

[4] Führt man die Schätzung mittels Regressionsanalyse unter Verwendung von Dummy-Variablen durch, wie oben angedeutet, so ist Vorsicht geboten, da R-Quadrat nicht korrekt bestimmt werden kann, wenn das Regressionsmodell kein konstantes Glied enthält. Um ein richtiges Ergebnis zu erzielen, ist ein Modell mit Referenzkategorie zu wählen.

3.2.1.3 Prüfung der statistischen Signifikanz



Ein hoher Wert für Eta-Quadrat besagt, dass das geschätzte Modell die Daten der Stichprobe gut erklärt. Das aber ist nur eine notwendige Bedingung dafür, dass dies auch für die Grundgesamtheit gilt. Um das zu prüfen, ist ein statistischer Test erforderlich, der auch den Umfang der Daten berücksichtigt. Eine Modellschätzung, die auf einer großen Stichprobe basiert, bietet größere Gewähr für allgemeine Gültigkeit, als eine Schätzung, die nur auf wenigen Daten basiert, auch wenn sie diese gut erklärt.

Die Varianzanalyse verwendet für die Prüfung der statistischen Signifikanz eines Modells die F-Statistik.

F-Statistik

$$F_{emp} = \frac{\text{erklärte Varianz}}{\text{nicht erklärte Varianz}} = \frac{SS_b/(G-1)}{SS_w/(G\cdot(K-1))} = \frac{MS_b}{MS_w} \tag{3.9}$$

Die F-Statistik[5] setzt zwei Varianzen ins Verhältnis, von denen die obere Varianz (im Zähler) die zu prüfenden experimentellen Effekte des Faktors (oder der Faktoren) enthält und die untere Varianz (im Nenner) die Störeinflüsse. Hieraus leitet sich der Name „Varianzanalyse“ ab. Je stärker die experimentellen Effekte sind, desto größer wird die F-Statistik. Sind die Störeinflüsse gering, so lassen sich schon kleinste Effekte als signifikant (durch den Faktor verursacht) nachweisen. Je größer aber die Störeinflüsse sind, desto größer wird die Varianz im Nenner und desto schwieriger wird der Nachweis der Signifikanz. Um eine Analogie zu gebrauchen: Je lauter die Umweltgeräusche sind, desto lauter muss man schreien, um verstanden zu werden. Der Nachrichtentechniker spricht hier von Signal-Rausch-Verhältnis (signal-to-noise ratio).

Die Varianzen (mittleren quadratischen Abweichungen) errechnen sich aus den Streuungen durch Division mit ihren jeweiligen Freiheitgraden. Abbildung 3.9 zeigt die Berechnungen.

Varianzquelle	SS	df	MS
zwischen den Faktorstufen	$\sum_{g=1}^{G} K(\overline{y}_g - \overline{y})^2 = 1112,13$	$G-1=2$	$\frac{SS_b}{G-1} = 556,07$
innerhalb der Faktorstufen	$\sum_{g=1}^{G}\sum_{k=1}^{K}(y_{gk} - \overline{y}_g)^2 = 175,2$	$G(K-1)=12$	$\frac{SS_w}{G(K-1)} = 14,6$
Gesamt	$\sum_{g=1}^{G}\sum_{k=1}^{K}(y_{gk} - \overline{y})^2 = 1287,33$	$G\cdot K-1=14$	$\frac{SS_t}{G\cdot K-1} = 91,95$

Abbildung 3.9: Zusammenstellung von Ergebnissen der einfaktoriellen Varianzanalyse

[5] Die F-Statistik wie auch die zugehörige F-Verteilung sind benannt nach dem genialen englischen Statistiker Sir Ronald A. Fisher (1890–1962), von dem auch das Konzept der Varianzanalyse stammt.

Die Zahl der *Freiheitsgrade* df (degrees of freedom) für die gesamte Streuung ergibt sich aus der Zahl der Beobachtungswerte vermindert um 1, weil der Mittelwert, von dem die Abweichungen berechnet wurden, aus den Beobachtungswerten selbst errechnet wurde. Demnach lässt sich immer einer der Beobachtungswerte aus den anderen $G \cdot K - 1$ Beobachtungswerten *und* dem geschätzten Mittelwert errechnen, d. h. er ist nicht mehr „frei". In unserem Beispiel haben wir 3 Faktorstufen mit je 5 Beobachtungen und damit insgesamt 15 Beobachtungen. df_t ist demnach $15 - 1 = 14$. Die nicht erklärte Streuung errechnet sich aus den Abweichung der 15 Beobachtungswerten von den drei Gruppenmittelwerten und es ergibt sich $df_w = 15 - 3 = 12$. Die erklärte Streuung errechnet sich aus den Abweichungen der drei Gruppenmittelwerte vom Gesamtmittelwert und es ergibt sich $df_b = 3 - 1 = 2$.

So wie sich erklärte Streuung und nicht erklärte Streuung zur Gesamtstreuung addieren, tun dies auch die zugehörigen Freiheitsgrade und es gilt: $df_t = df_b + df_w$.

Mit den Werten aus Abbildung 3.9 erhält man den folgenden empirischen F-Wert:

$$F_{emp} = \frac{MS_b}{MS_w} = \frac{556,07}{14,6} = 38,09$$

Entsprechend den beiden alternativen Modelformulierungen A und B lassen sich mit dem F-Test die folgenden zwei Nullhypothesen prüfen, die völlig gleichwertig sind:

(A) H_0: $\mu_1 = \mu_2 = ... = \mu_G$
(B) H_0: $\alpha_1 = \alpha_2 = ... = \alpha_G = 0$

Die zugehörigen Alternativhypothesen lauten:

(A) H_1: Die Mittelwerte sind nicht alle gleich
(B) H_1: mindestens zwei α_g sind $\neq 0$

Da sich die Effekte zu Null summieren sollen, kann nicht ein Effekt allein $\neq 0$ sein.

Bezogen auf unser Beispiel besagen die Nullhypothesen, dass sich die Absatzmengen bei den unterschiedlichen Regalplatzierungen nicht unterscheiden bzw. dass die Regalplatzierungen keine Wirkung auf die Absatzmengen haben.

Die Prüfung der Hypothesen erfolgt anhand eines Vergleichs des empirischen F-Wertes mit einem theoretischen F-Wert, der abhängig ist von der vom Untersucher gewählten Irrtumswahrscheinlichkeit α (Signifikanzniveau) und den Freiheitsgraden.

Abbildung 3.10 zeigt einen Ausschnitt aus der F-Tabelle für die Irrtumswahrscheinlichkeit $\alpha = 1\%$. Der gesuchte Tabellenwert findet sich in Spalte 2 (Zahl der Freiheitsgrade im Zähler) und Zeile 12 (Zahl der Freiheitsgrade im Nenner) und lautet $F_{tab} = 6,93$. Er entspricht dem 99%-Quantil der F-Verteilung mit 2 und 12 Freiheitsgraden.

Für die Durchführung des F-Tests gilt folgende Regel:

$F_{emp} > F_{tab} \rightarrow H_0$ wird verworfen $\rightarrow$ Zusammenhang ist signifikant
$F_{emp} \leq F_{tab} \rightarrow H_0$ wird nicht verworfen

Hier ergibt sich:

$$38,09 > 6,93 \rightarrow H_0 \text{ wird verworfen}$$

Mit einem Signifikanzniveau von $\alpha = 0,01$ kann daher hier gefolgert werden, dass die Regalplatzierung einen Einfluss auf die Absatzmenge hat.

Freiheitsgrade des Zählers/ Freiheitsgrade des Nenners	1	2	3	4	5
10	10,04	7,56	6,55	5,99	5,64
11	9,65	7,21	6,22	5,67	5,32
12	9,33	6,93	5,95	5,41	5,06
13	9,07	6,70	5,74	5,21	4,86
14	8,86	6,51	5,56	5,04	4,69

Abbildung 3.10: Ausschnitt aus der F-Tabelle (Signifikanzniveau 1 %)

p-Wert

Alternativ kann der F-Test durchgeführt werden, indem man den p-Wert (prob value) der F-Statistik ermittelt. Man kann dann auf Tabellen verzichten und gewinnt zusätzliche Information. Der p-Wert der F-Statistik ist definiert als die Wahrscheinlichkeit, dass eine F-verteilte Zufallsvariable (mit df1 und df2 Freiheitsgraden) größer ist als der empirisch ermittelte F-Wert. Er wird daher auch als *empirisches Signifikanzniveau* bezeichnet.[6] Das Entscheidungskriterium für den F-Test lautet:

$$p < \alpha \rightarrow H_0 \text{ wird verworfen}$$

Andernfalls, also falls $p \geq \alpha$, muss die Nullhypothese beibehalten werden. Für $F_{emp} = 38,09$ ergibt sich hier $p = 0,0000064$. H_0 ist damit zu verwerfen. Das Ergebnis ist natürlich identisch mit der oben gezeigten klassischen Testdurchführung. Der p-Wert ist hier praktisch null und zeigt damit, wie deutlich der Effekt der Regalplatzierung auf die Absatzmenge ist. Dies bestätigt das Bild, dass wir bereits anfangs mittels der Boxplots erhalten haben.

Der hier durchgeführte F-Test ist ein sog. Omnibus-Test, d.h. er prüft, ob es Unterschiede zwischen den Gruppen gibt, nicht aber, ob sich alle Gruppen voneinander unterscheiden. Hier ist das zwar der Fall, aber die Boxplots zeigen auch, dass der Unterschied zwischen Zweitplatzierung und Normalregal besonders groß ist, während der Unterschied zwischen Normalregal und Kühlregal am geringsten ist. Mit dem F-Test lassen sich separat auch alle Paare von Gruppen testen. Auf derartige *Multiple Vergleichstests* wird weiter unten eingegangen.

[6] Siehe dazu die Ausführungen in Kapitel 1, Abschnitt 1.2.3.2. In SPSS wird der p-Wert unter der Bezeichnung „Signifikanz“ oder „Sig.“ ausgegeben. In Excel kann man den p-Wert der F-Verteilung durch die Funktion FERT(x; df1; df2) berechnen. Es gilt damit FVERT$(38,09; 2; 12) = 0,0000064$.

Abschließend sei darauf hingewiesen, dass die Gültigkeit des F-Tests auf den obigen Modellannahmen basiert. So sollten die Beobachtungen unabhängig voneinander sein und ihre durch die Störgrößen bewirkten Varianzen sollten in den Gruppen annähernd gleich sein. In balancierten Designs hat eine Verletzung dieser Annahme allerdings nur geringe Auswirkungen auf das Testergebnis. Ein statistischer Test zur Prüfung auf gleiche Varianzen, den SPSS anbietet, ist der *Levene-Test*. Zur Prüfung auf Normalverteilung kann ein Q-Q-Diagramm (Quantil-Quantil-Plot) verwendet werden.[7] Der F-Test ist allerdings recht unempfindlich gegenüber Verletzungen der Normalverteilungsannahme. Je größer der Stichprobenumfang, desto mehr verliert sie an Bedeutung.

Und schließlich sollte die Störgröße keine systematischen Einflussfaktoren enthalten. Sollten weitere systematische Einflussfaktoren vorhanden sein, so gehen sie automatisch in die Störgröße ein. Um das zu verhindern, muss dann das Modell erweitert werden. Das ist Gegenstand der folgenden Ausführungen.

3.2.2 Zweifaktorielle Varianzanalyse

3.2.2.1 Modellformulierung

Es ist effizienter, wenn man gleichzeitig zwei oder mehrere Faktoren untersucht, anstatt für jeden Faktor eine separate Untersuchung durchzuführen. Bei gleichzeitiger Variation von zwei oder mehr Faktoren spricht man von *Faktoriellen Designs*. Die Erweiterung der Varianzanalyse auf zwei unabhängige Variablen heißt *zweifaktorielle Varianzanalyse*.
Abgesehen von Effizienzeffekten kann die Erweiterung des Modells der Varianzanalyse auf mehre Faktoren weitere Vorteile erbringen:

- Eventuelle Wechselwirkungen (Interaktionseffekte) zwischen den Faktoren lassen sich erfassen
- Die nichterklärte Varianz lässt sich verringern und damit der Nachweis von Faktorwirkungen erleichtern.

Wir kommen auf unser Beispiel zurück. Der Supermarkt-Manager möchte wissen, ob neben der Regalplatzierung auch die Verpackungsart („Becher“ oder „Papier“) Einfluss auf den Absatz hat. Dazu wird das Experiment entsprechend erweitert. Bei drei Platzierungsarten und zwei Verpackungsarten ergeben sich 3 x 2 experimentelle Kombinationen der Faktorstufen. Wir sprechen auch von einem 3 x 2-faktoriellen Design. Demnach werden jetzt sechs annähernd gleiche Supermärkte ausgesucht. Abbildung 3.11 zeigt die erweiterte Datenmatrix mit den sechs Zellen (Teilstichproben á 5 Beobachtungen).

[7] SPSS bietet diese Analysen unter dem Menüpunkt „Deskriptive Statistiken“ an.

Weiterhin möchte der Manager in Erfahrung bringen, ob zwischen den Faktoren Verpackung und Warenplatzierung eventuelle Wechselwirkungen (Interaktionen) bestehen. So mag z. B. die Vermutung gerechtfertigt erscheinen, dass der durchschnittliche Absatz von Margarine in Becherform anders auf die Variation der Platzierung reagiert als die Papierverpackung, etwa, weil ein Weichwerden der Margarine im „Normalregal“ eher auffällt als im Kühlregal.

Platzierung		Verpackung ,,Becher “	,,Papier “
,,Normalregal “	Tag 1	47	40
	Tag 2	39	39
	Tag 3	40	35
	Tag 4	46	36
	Tag 5	45	37
,,Zweitplatzierung “	Tag 1	68	59
	Tag 2	65	57
	Tag 3	63	54
	Tag 4	59	56
	Tag 5	67	53
,,Kühlregal “	Tag 1	59	53
	Tag 2	50	47
	Tag 3	51	48
	Tag 4	48	50
	Tag 5	53	51

Abbildung 3.11: kg Margarineabsatz pro 1.000 Kassenvorgänge in sechs Supermärkten in Abhängigkeit von der Platzierung und der Verpackung

Visualisierung

Eine einfache und sehr anschauliche Methode, um das Vorhandensein von Interaktion zu prüfen, ist ein Plot der Faktorstufenmittelwerte. Abbildung 3.12 zeigt die Werte des Beispiels.[8]

Keine Interaktionen liegen vor, wenn die Verbindungslinien der Mittelwerte (die hier nur zur Verdeutlichung eingezeichnet sind) parallel laufen. Nichtparallele Verläufe sind ein klares Indiz für das Vorhandensein und die Stärke von Interaktionen. Im vorliegenden Fall bietet sich ein Anhaltspunkt für eine schwache Interaktion von Verpackung und Platzierung, da der Wirkungsunterschied zwischen Becher und Papier im Kühlregal im Analyseergebnis nahezu verschwindet, möglicherweise, weil dort von den Käufern ein Unterschied nicht wahrgenommen wird.

[8] Die in Abbildung 3.11 abgebildeten Daten werden auch im Rahmen des Fallbeispiels in Abschnitt 3.3 verwendet. Abbildung 3.12 wurde deshalb mit SPSS, Prozedur „Allgemeines lineares Modell / Univariat“ und dort über das Dialogfenster „Diagramme“ (vgl. Abbildung 3.21) angefordert. In der Dialogbox zu „Diagramme“ ist dann der Faktor REGAL als „Horizontale Achse“ und der Faktor VERPACK unter „Separate Linien“ einzutragen und anschließend über „Hinzufügen“ in das Feld „Diagramme“ einzutragen.

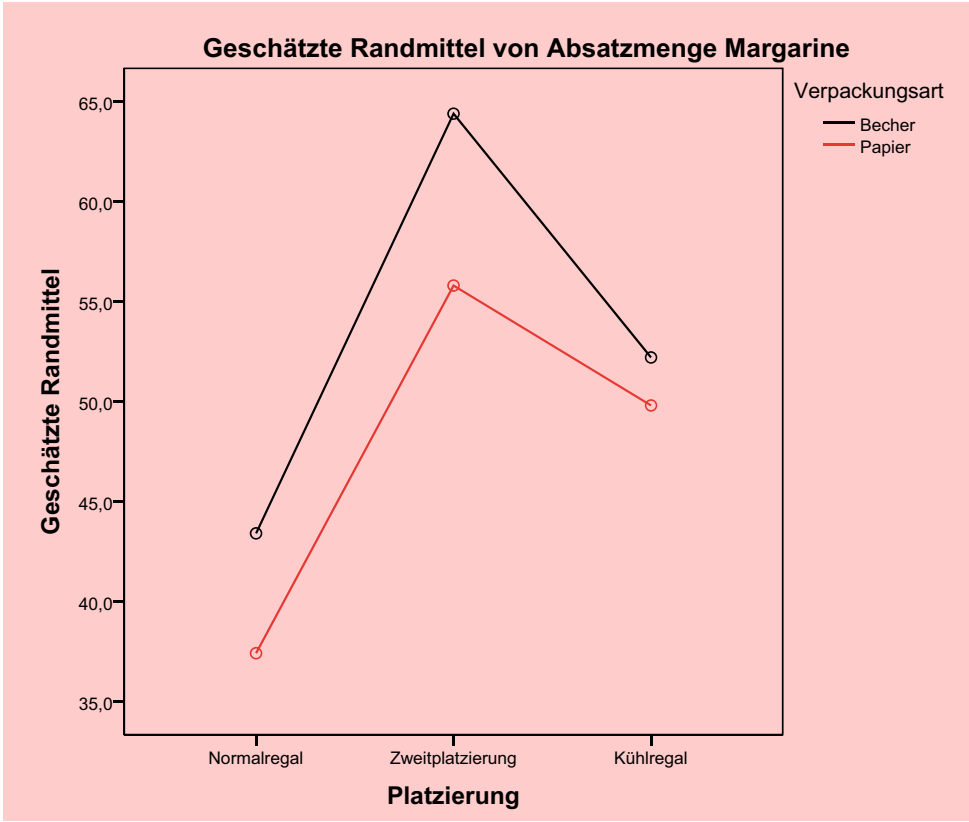


Abbildung 3.12: Graphische Analyse von Interaktionen (Werte entnommen aus Abbildung 3.13)

Modell der zweifaktoriellen Varianzanalyse

Das Modell der zweifaktoriellen Varianzanalyse mit Interaktionseffekten hat folgende Form:

$$y_{ghk} = \mu + \alpha_g + \beta_h + (\alpha\beta)_{gh} + \varepsilon_{ghk} \tag{3.10}$$

mit

y_{ghk}	= Beobachtungswert
μ	= Gesamtmittelwert der Grundgesamtheit
α_g	= wahrer Effekt von Platzierung g ($g = 1, 2, 3$)
β_h	= wahrer Effekt von Verpackung h ($h = 1, 2$)
$(\alpha\beta)_{gh}$	= wahrer Interaktionseffekt von Platzierung g und Verpackung h
ϵ_{ghk}	= Störgröße

Auch hier vereinbaren wir wieder zwecks eindeutiger Bestimmung (Identifizierbarkeit) der Effekte, dass diese sich jeweils zu null addieren. Die isolierten Effekte der Faktoren bezeichnet man bei der mehrfaktoriellen Varianzanalyse als Haupteffekte (main effects) zur Unterscheidung von den Interaktionseffekten. Ihre Schätzwerte errechnen sich wie bei der einfaktoriellen Varianzanalyse durch die Differenzen zwischen den Gruppenmittelwerten und dem Gesamtmittelwert. Abbildung 3.13 zeigt die für die Berechnung notwendigen Werte. Sie enthält die Gruppenmittelwerte für Platzierung als Zeilenmittelwerte und die Gruppenmittelwerte für Verpackung als Spaltenmittelwerte.

Berechnung der Haupteffekte

$$\mathrm{a}_g = (\bar{\mathrm{y}}_{\mathrm{g.}} - \bar{\mathrm{y}}) \tag{3.11}$$

$$\mathrm{b}_h = (\bar{\mathrm{y}}_{\mathrm{.h}} - \bar{\mathrm{y}}) \tag{3.12}$$

mit

$$\bar{y}_{g.} = \frac{1}{H \cdot K} \sum_{h=1}^{H} \sum_{k=1}^{K} y_{ghk} \qquad \text{(Gruppenmittelwerte g)}$$

$$\bar{y}_{.h} = \frac{1}{G \cdot K} \sum_{g=1}^{G} \sum_{k=1}^{K} y_{ghk} \qquad \text{(Gruppenmittelwerte h)}$$

$$\bar{y} = \frac{1}{G \cdot H \cdot K} \sum_{g=1}^{G} \sum_{h=1}^{H} \sum_{k=1}^{K} y_{ghk} \qquad \text{(Gesamtmittelwert)}$$

Für die Effekte der drei Platzierungsarten erhält man hier:

$$a_1 = (\bar{y}_1 - \bar{y}) = 40,4 - 50,5 = -10,1 \qquad \text{(Normalregal)}$$

$$a_2 = (\bar{y}_2 - \bar{y}) = 60,1 - 50,5 = 9,6 \qquad \text{(Zweitplatzierung)}$$

$$a_3 = (\bar{y}_3 - \bar{y}) = 51,0 - 50,5 = 0,5 \qquad \text{(Kühlregal)}$$

Die Summe der drei Effekte ist wieder null.

Analog ergeben sich für die Effekte der beiden Verpackungsarten:

$$b_1 = (\bar{y}_1 - \bar{y}) = 53,33 - 50,5 = 2,83 \qquad \text{(Becher)}$$

$$b_2 = (\bar{y}_2 - \bar{y}) = 47,67 - 50,5 = -2,83 \qquad \text{(Papier)}$$

Der Effekt der Verpackungsart ist in Abbildung 3.12 als mittlerer Abstand zwischen den beiden Polygonzügen sichtbar.

H G	h=1		h=2		$\sum_h \sum_k y_{ghk}$	
$g = 1$	47 39 40 46 45	$43,4$	40 39 35 36 37	$37,4$		$\overline{y}_{(g=1)} = \frac{404}{10} = 40,4$
$\sum y_{1hk}$	(217)		(187)		404	
$g = 2$	68 65 63 59 67	$64,4$	59 57 54 56 53	$55,8$		$\overline{y}_{(g=2)} = \frac{601}{10} = 60,1$
$\sum y_{2hk}$	(322)		(279)		601	
$g = 3$	59 50 51 48 53	$52,2$	53 47 48 50 51	$49,8$		$\overline{y}_{(g=3)} = \frac{510}{10} = 51,0$
$\sum y_{3hk}$	(261)		(249)		510	
$\sum_g \sum_k y_{ghk}$	800		715		1.515	
	$\overline{y}_{(h=1)} = \frac{800}{15} = 53,\overline{3}$		$\overline{y}_{(h=2)} = \frac{715}{15} = 47,\overline{6}$		$\overline{y} = \frac{1515}{30} = 50,5$	

Abbildung 3.13: Berechnung der Mittelwerte für Zeilen, Spalten und Zellen

3.2.2.2 Berechnung der Interaktionseffekte

Die Schätzung der Interaktionseffekte erfolgt durch

$$(ab)_{\text{gh}} = \bar{y}_{\text{gh}} - \hat{y}_{\text{gh}} \tag{3.13}$$

mit

$\bar{y}_{gh} = \frac{1}{K} \sum_{k=1}^{5} y_{ghk}$ = beobachteter Mittelwert in Zelle (g,h)
$\hat{y}_{\text{gh}}$ = Schätzwert für den Mittelwert von Zelle (g,h) ohne Interaktion

Der Schätzwert $\hat{y}_{gh}$ ist derjenige Wert, der für die Zelle (g,h) zu erwarten wäre, wenn keine Interaktion vorläge. Er errechnet sich aus den Gruppenmitteln und dem Gesamtmittelwert wie folgt:

$$\hat{\text{y}}_{gh} = \bar{\text{y}}_{\text{g.}} + \bar{\text{y}}_{\text{.h}} - \bar{\text{y}} \tag{3.14}$$

Wir betrachten z.B. die Zelle für $g = 3$ und $h = 2$. Der beobachtete Mittelwert beträgt $\bar{y}_{32} = 49,8$ (Abbildung 3.13). Dieser Wert enthält den Interaktionseffekt, falls

ein solcher vorhanden ist. Für den Schätzwert ohne Interaktion erhält man:

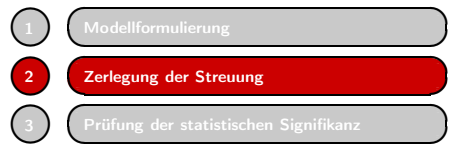
$$\hat{y}_{32} = 51,00 + 47,66 + 50,50 = 48,16$$

Der Interaktionseffekt ergibt sich damit durch:

$$(ab)_{gh} = 49,8 - 48,16 = 1,64$$

Bedingt durch die Interaktion ergibt sich hier für den Absatz von Margarine in Papier ein höherer Wert, wenn sie im Kühlregal angeboten wird. Entsprechend ist in Abbildung 3.12 der rechte Punkt des unteren Polygonzuges durch den Interaktionseffekt nach oben verschoben.

3.2.2.3 Zerlegung der Streuung



Zur Beurteilung der Güte des Modells und Prüfung der Signifikanz der Effekte ist wiederum eine Zerlegung der Gesamtstreuung der Daten vorzunehmen, wie sie schon bei der einfachen Varianzanalyse erfolgte. Zur Vereinfachung der Notation bezeichnen wir die beiden Faktoren mit A und B. Das Prinzip der Streuungszerlegung für die zweifaktorielle Varianzanalyse ist in Abbildung 3.14 dargestellt.

Wie bei der einfachen Varianzanalyse wird die Gesamtstreuung in eine erklärte Streuung und eine nicht erklärte Streuung aufgeteilt. Die erklärte Streuung wird dabei jetzt weiter in drei Komponenten aufgeteilt, die sich aus dem Einfluss des Faktors A, dem Einfluss des Faktors B und der Interaktion von Faktor A und B ergeben. Man erhält damit folgende Zerlegung der Gesamtstreuung:

$$SS_t = SS_A + SS_B + SS_{AxB} + SS_w \tag{3.15}$$

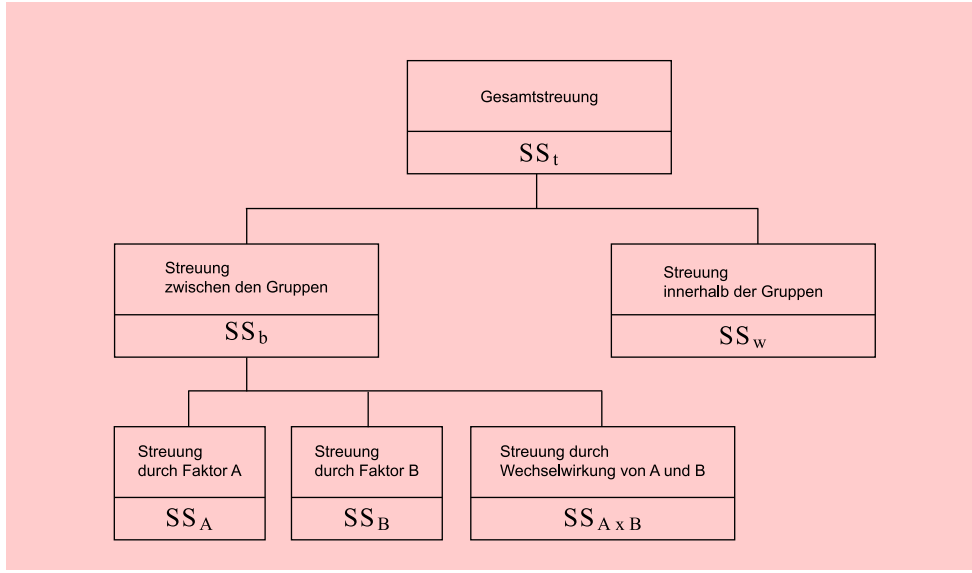


Abbildung 3.14: Aufteilung der Gesamtstreuung im faktoriellen Design mit 2 Faktoren

Für die Gesamtstreuung ergibt sich in unserem Beispiel:

$$SS_t = \sum_{g=1}^{G}\sum_{h=1}^{H}\sum_{k=1}^{K}(y_{ghk} - \bar{y})^2 = 2471,50 \tag{3.16}$$

Die durch die isolierten Wirkungen (Haupteffekte) von Faktor A (Platzierung) und Faktor B (Verpackung) erzeugten Streuungen errechnen sich aus den Abweichungen der Zeilen- bzw. Spaltenmittelwerte vom Gesamtmittelwert (vgl. Abbildung 3.13):

$$SS_A = 2 \cdot 5 \cdot [(40,4 - 50,5)^2 + (60,1 - 50,5)^2 + (51,0 - 50,5)^2] = 1.944,20$$
$$SS_B = 3 \cdot 5 \cdot [(53,\overline{3} - 50,5)^2 + (47,\overline{6} - 50,5)^2] = 240,83$$

SS_A	=	$H \cdot K \cdot \sum_{g=1}^{G}(\overline{y}_g - \overline{y})^2$
SS_B	=	$G \cdot K \cdot \sum_{h=1}^{H}(\overline{y}_h - \overline{y})^2$
G	=	Zahl der Ausprägungen des Faktors A
H	=	Zahl der Ausprägungen des Faktors B
K	=	Zahl der Elemente in Zelle (g,h)
$\overline{y}_g$	=	Zeilenmittelwert
$\overline{y}_h$	=	Spaltenmittelwert

Abbildung 3.15: Durch Haupteffekte erklärte Streuung im zweifaktoriellen Design

Die durch die Interaktionseffekte erzeugte Streuung ergibt sich durch Summation der quadrierten Abweichungen zwischen den Zellenmittelwerten und den Schätzwerten, die ohne Interaktion zu erwarten wären:

$$SS_{AxB} = K \cdot \sum_{g=1}^{G}\sum_{h=1}^{H}(\overline{y}_{gh} - \hat{y}_{gh})^2 \tag{3.17}$$

Die Zellenmittelwerte sind aus Abbildung 3.13 zu entnehmen. Für die ohne Interaktion zu erwartenden Mittelwerte erhält man folgende Schätzwerte:

$$\hat{y}_{11} = 40,4 + 53,\overline{3} - 50,5 = 43,2\overline{3}$$
$$\hat{y}_{12} = 40,4 + 47,\overline{6} - 50,5 = 37,5\overline{6}$$
$$\hat{y}_{21} = 60,1 + 53,\overline{3} - 50,5 = 62,9\overline{3}$$
$$\hat{y}_{22} = 60,1 + 47,\overline{6} - 50,5 = 57,2\overline{6}$$
$$\hat{y}_{31} = 51,0 + 53,\overline{3} - 50,5 = 53,8\overline{3}$$
$$\hat{y}_{32} = 51,0 + 47,\overline{6} - 50,5 = 48,1\overline{6}$$

Damit ergibt sich für die durch die Wechselwirkungen erklärte Streuung:

$$SS_{AxB} = 5 \cdot \left\{ \begin{array}{c} (43,4 - 43,2\overline{3})^2 + (37,4 - 37,5\overline{6})^2 \\ +(64,4 - 62,9\overline{3})^2 + (55,8 - 57,2\overline{6})^2 \\ +(52,2 - 53,8\overline{3})^2 + (49,8 - 48,1\overline{6})^2 \end{array} \right\}$$
$$= 48,47$$

Analog zu Abbildung 3.14 gilt:

$$SS_b = SS_A + SS_B + SS_{AxB} \tag{3.18}$$

Die Sum of Squares SS_b sind die Abweichungen zwischen den Gruppenmitteln und dem Gesamtmittel:

$$SS_b = K \cdot \sum_{g=1}^{G} \sum_{h=1}^{H} (\overline{y}_{gh} - \overline{y})^2 \tag{3.19}$$

Zu unserem Beispiel ergibt sich:

$$\begin{aligned} SS_b &= 5 \cdot \left\{ (43,4 - 50,5)^2 + ... + (49,8 - 50,5)^2 \right\} \\ &= 2.233,5 \end{aligned}$$

Die SS_{AxB} können nun auch bestimmt werden aus:

$$\begin{aligned} SS_{AxB} &= SS_b - SS_A - SS_B \\ &= 2.233,5 - 240,83 - 1.944,20 \\ &= 48,47 \end{aligned} \tag{3.20}$$

Die Reststreuung, die sich als „Streuung innerhalb der Zellen“ analog zu SS_W bei der einfachen Analyse manifestiert, ist definiert als:

$$SS_w = \sum_{g=1}^{G} \sum_{h=1}^{H} \sum_{k=1}^{K} (y_{ghk} - \overline{y}_{gh})^2 \tag{3.21}$$

Reststreuung

Sie ist die Streuung, die weder auf die beiden Faktoren noch auf Interaktionseffekte zurückzuführen ist, d. h. es handelt sich um zufällige Einflüsse auf die abhängige Variable. Die Beispielsrechnung ergibt (vgl. Abbildung 3.13):

$$\begin{aligned} SS_w &= (47 - 43,4)^2 + ... + (45 - 43,4)^2 \\ &\quad + (40 - 37,4)^2 + ... + (37 - 37,4)^2 \\ &\quad + (68 - 64,4)^2 + ... + ... \\ &\quad + (53 - 49,8)^2 + ... + (51 - 49,8)^2 \\ &= 238 \end{aligned}$$

In Analogie zu Abbildung 3.14 lässt sich die Reststreuung auch indirekt über die Zerlegung der Gesamtstreuung berechnen:

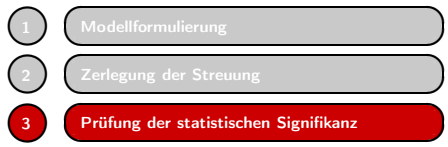
$$\begin{aligned} SS_w &= SS_t - SS_A - SS_B - SS_{AxB} = SS_t - SS_b \\ &= 2.471,5 - 2.233,5 = 238 \end{aligned} \tag{3.22}$$

Zusammenfassend kann auf Basis der Streuungszerlegung jetzt wieder die Güte des Modells beurteilt werden:

$$\text{Eta-Quadrat} = \frac{\text{erklärte Streuung}}{\text{gesamte Streuung}} = \frac{SS_b}{SS_t} = \frac{2.233,5}{2.471,5} = 0,904$$

Mittels des erweiterten Modells können jetzt 90,4% der gesamten Streuung erklärt werden. Eine einfache Varianzanalyse mit dem Faktor Regalplatzierung würde für den vorliegenden Datensatz 78,7% der Streuung erklären. Die nicht erklärte Streuung kann damit durch Erweiterung des Modells von 21,3% auf 9,6% reduziert werden.

3.2.2.4 Prüfung der statistischen Signifikanz



Im zweifaktoriellen Fall erfolgt die statistische Prüfung auf unterschiedliche Wirkungen der beiden Faktoren durch einen Vergleich der Mittelwerte in allen Zellen. Wenn alle Mittelwerte annähernd gleich sind, so deutet dies die darauf hin, dass die Faktoren keine Wirkung haben (Nullhypothese). Die Alternativhypothese besagt, dass zumindest eine der Faktorstufe einen Einfluss besitzt.

Der globale Signifikanztest für das zweifaktorielle Modell ist damit (bis auf die unterschiedliche Anzahl von Freiheitsgraden) identisch mit dem Test für das einfache Modell:

$$F_{emp} = \frac{\text{erklärte Varianz}}{\text{nicht erklärte Varianz}} = \frac{SS_b/(G \cdot H - 1)}{SS_w/(G \cdot H \cdot K - G \cdot H)} = \frac{MS_b}{MS_w} \tag{3.23}$$

Mit obigen Werten erhält man:

$$F_{emp} = \frac{2.233,5/5}{238,0/24} = \frac{446,7}{9,917} = 45,05$$

Bei einer Vertrauenswahrscheinlichkeit von 99% lässt sich hier aus der F-Tabelle der Wert $F_{tab} = 3,90$ entnehmen. Das Ergebnis ist also hoch signifikant und die Nullhypothese kann zurückgewiesen werden. Der zugehörige p-Wert ist praktisch null.

Damit lassen sich weitere Fragestellungen untersuchen, die die isolierte Analyse einzelner Faktoren bzw. deren Interaktionen betreffen. In diesen Fällen lautet die Nullhypothese, dass der jeweils untersuchte Faktor keine Wirkung hat oder dass keine Wechselwirkungen vorhanden sind. In Abbildung 3.16 sind die für die Testdurchführung benötigten Ergebnisse der zweifaktoriellen Varianzanalyse zusammengefasst. Für die Freiheitsgrade der verschiedenen Streuungen gilt:

$$\begin{aligned} df_A &= G - 1 \\ df_B &= H - 1 \\ df_{AxB} &= (G-1) \cdot (H-1) \\ df_w &= G \cdot H \cdot (K-1) \\ df_t &= G \cdot H \cdot K - 1 \end{aligned}$$

Varianzquelle	SS	df	MS
Haupteffekte			
Platzierung	1.944, 2000	2	972, 1000
Verpackung	240, 8333	1	240, 8333
Interaktion			
Platzierung/Verpackung	48, 4667	2	24, 2333
Reststreuung	238	24	9, 9167
Total	2.471, 50	29	85, 224

Abbildung 3.16: Streuungen und Varianzen der zweifaktoriellen Varianzanalyse (ANOVA-Tabelle)

Abbildung 3.17 zeigt die Ergebnisse der spezifischen F-Tests mit einer Vertrauenswahrscheinlichkeit von 99%. Die Varianz der Reststreuung ist in allen Fällen dieselbe. Je größer der Stichprobenumfang ist und desto mehr Streuung durch die Faktoren des Modells erklärt werden kann, desto kleiner wird sie und desto schärfer werden die Tests.

Quelle der Varianz	df (Zähler)	df (Nenner)	F_{tab}	F_{emp}
Verpackung	1	24	$7,82$	$24,2856$
Platzierung	2	24	$5,61$	$98,0265$
Interaktion				
Verpackung/Platzierung	2	24	$5,61$	$2,4437$

Abbildung 3.17: Spezifische F-Tests im zweifaktoriellen Design

Das Ergebnis zeigt, dass für beide Faktoren die jeweilige Nullhypothese verworfen werden kann, für die Interaktion dagegen nicht. Verpackung und Platzierung haben also isoliert betrachtet jeweils eine Wirkung auf den Absatz, eine gemeinsame Wirkung von Verpackung und Platzierung zeigt sich aufgrund des F-Tests als nicht signifikant. Dies muss nicht heißen, dass in Wirklichkeit kein Zusammenhang vorliegt, sondern nur, dass die Nullhypothese aufgrund der vorliegenden Ergebnisse nicht verworfen werden kann (vgl. die graphische Analyse der Interaktionen in Abbildung 3.12).

3.2.3 Ausgewählte Erweiterungen der Varianzanalysen

Ungleich besetzte Zellen

Ungleich besetzte Zellen

In der bisherigen Darstellung sind wir davon ausgegangen, dass jede Zelle mit einer gleich großen Zahl von Beobachtungswerten besetzt ist. Eine erste Erweiterung der Analyse liegt in der Einbeziehung von ungleich besetzten Zellen. Es ergibt sich eine Anpassung in den oben definierten Formeln zur Zerlegung der Streuung. Am Prinzip der Streuungszerlegung ändert sich allerdings nichts. Es kommt lediglich zu einer Gewichtung der einzelnen Beobachtungswerte.

Mehrere Faktoren

Eine andere Erweiterung, die ebenfalls am Prinzip der Streuungszerlegung festhält, ist die Einbeziehung von mehr als zwei Faktoren in die Analyse. So ergeben sich beispielsweise bei der dreifaktoriellen Varianzanalyse prinzipiell keine Unterschiede zur zweifaktoriellen. Durch das Hinzutreten des dritten Faktors ergibt sich lediglich eine differenziertere Zerlegung der Streuung. Die Gesamtstreuung teilt sich nunmehr wie in Abbildung 3.18 dargestellt auf.

Die Besonderheit gegenüber der zweifaktoriellen Varianzanalyse liegt darin, dass jetzt zwei verschiedene Ebenen möglicher Wechselwirkungen entstehen: Es gibt die Wechselwirkung zwischen jeweils *zwei* Faktoren und zusätzlich die Wechselwirkung zwischen allen drei Faktoren. Werden mehr als drei Faktoren in die Analyse einbezogen, ergeben sich entsprechend mehr Ebenen der Analyse von Interaktionen zwischen den Faktoren. In diesen Fällen sind die Interaktionen jedoch kaum noch inhaltlich interpretierbar.

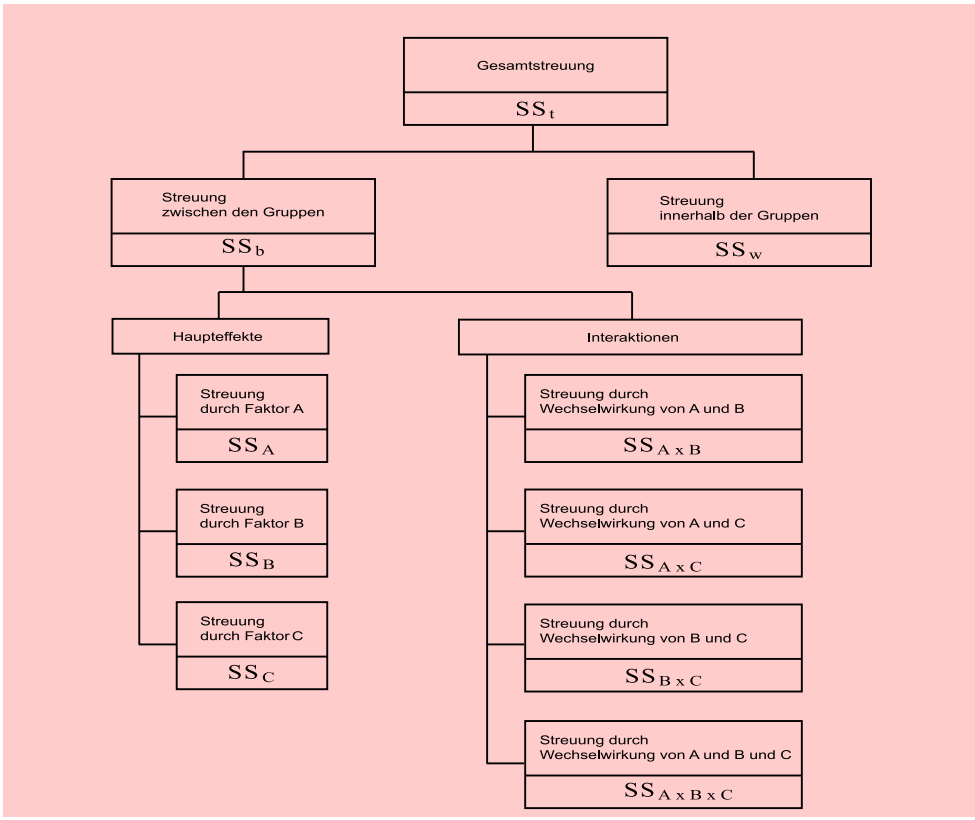


Abbildung 3.18: Aufteilung der Gesamtstreuung im dreifaktoriellen Design

Multiple Vergleichstests (Kontrastanalyse und Post-hoc-Tests)

Multiple Vergleichstests

Mit Hilfe der Ergebnisse einer Varianzanalyse kann festgestellt werden, ob ein Faktor bzw. mehrere Faktoren (Treatments) einen signifikanten Effekt auf eine abhängige Variable aufweisen. Dabei handelt es sich aber um eine sog. „Omnibushypothese“, da sich *nicht* feststellen lässt, *welche Stufen* eines bzw. mehrerer betrachteter Faktoren einen signifikanten Einfluss auf die abhängige Variable ausüben und *wie groß* diese Effekte sind. Zeigt also der F-Test, dass ein Faktor einen signifikanten Einfluss auf die abhängige Variable besitzt, so kann aus einem solchen Ergebnis *nicht* geschlossen werden, dass *alle* Faktorstufenmittelwerte (Gruppenmittelwerte) unterschiedlich sind und damit alle betrachteten Faktorstufen über einen bedeutsamen Einfluss auf die abhängige Variable verfügen. Vielmehr können durchaus mehrere Gruppenmittelwerte gleich sein und der Unterschied z. B. nur an einer Stelle begründet liegen. Für den Anwender ist die genaue Kenntnis der Unterschiede häufig aber von großem Interesse. Zur Analyse solcher Unterschiede sind zwei Situationen zu unterscheiden:

a) Der Anwender verfügt bereits *vor* der Analyse (ex ante; a priori) über theoretisch oder sachlogisch begründete Hypothesen, wo genau Mittelwertunterschiede in den Faktorstufen begründet liegen. Ob solche *vermuteten* Unterschiede (*Kontraste*), tatsächlich existieren, lässt sich dann mit Hilfe einer Kontrastanalyse überprüfen.

b) Der Anwender hat *keine* begründeten Hypothesen zu möglichen Wirkunterschieden in den Faktorstufen und möchte deshalb *nach* einem signifikanten F-Test (ex post) wissen, *wo* sich empirisch signifikante Mittelwertunterschiede zeigen. Um dies zu prüfen, kann er auf sog. Post-hoc-Tests zurückgreifen.

Kontrastanalyse

Eine *Kontrastanalyse* wäre anzuwenden, wenn im bisherigen Beispiel z. B. Marktstudien einen deutlichen Effekt von Zweitplatzierungen auf den Margarineabsatz nachgewiesen hätten und der Supermarktbetreiber deshalb davon ausgeht, dass die besondere Bedeutung der Zweitplatzierung auch in seinem Fall gilt. Zur Prüfung wäre dann der Margarineabsatz bei Zweitplatzierung im Vergleich zur Platzierung im Normal- und Kühlregal. Die Faktorstufen Normal- und Kühlregal würden zu diesem Zweck zu einer Gruppe zusammengefasst.

Post-hoc-Tests

Im Vergleich zur Kontrastanalyse werden *Post-hoc-Tests* erst dann durchgeführt, wenn der F-Test einer Varianzanalyse zu einem signifikanten Ergebnis führte und der Anwender anschließend (ex post; a posteriori) wissen möchte, welche Faktorstufen Unterschiede in den Mittelwerten begründen. Hypothesen hierzu bestehen allerdings nicht. Um dies herauszufinden, wäre vordergründig eine einfache Lösung, mit Hilfe eines t-Tests jeweils zwei Faktorstufen zu kombinieren und auf signifikante Unterschiede zwischen den Mittelwerten zu testen. Die Problematik, die sich dabei ergibt,

Alpha-Fehler-Inflation

liegt jedoch in der sog. *Alpha-Fehler-Inflation* (Alpha-Fehler-Kumulierung), die sich wie folgt verdeutlichen lässt:

Bonferoni-Korrektur

Sind z. B. fünf Faktorstufen vorhanden, so wären bereits „fünf-über-zwei“ = 10 verschiedene t-Tests durchzuführen, um alle paarweisen Kombinationen der Faktorstufen auf Mittelwertunterschiede zu testen. Es wird hier von *multiplen Tests* gesprochen, da dieselbe Nullhypothese mit mehreren Tests untersucht wird. Aufgrund dieser Einzeltests kommt es zu einer Kumulierung des Alpha-Fehlers (Fehler 1. Art: Wahrscheinlichkeit, die Nullhypothese abzulehnen, obwohl sie korrekt ist). Mit der Anzahl der Testwiederholungen steigt die Wahrscheinlichkeit (Gefahr), dass ein Unterschied als signifikant erscheint, auch wenn in Wirklichkeit keiner der Unterschiede signifikant ist. Bei $\alpha = 0,5$ und 10 Testwiederholungen beträgt diese Wahrscheinlichkeit schon rund 40%. Der Alpha-Fehler muss deshalb so korrigiert werden, dass im Ergebnis über die Vergleichstests die gewünschte Irrtumswahrscheinlichkeit (z. B. 5%) erhalten bleibt (am bekanntesten ist die *Bonferoni-Korrektur*, bei der α durch die Zahl der Testwiederholungen dividiert wird).

Die Möglichkeit der Vermeidung einer Alpha-Fehler-Inflation bieten die sog. *Post-hoc-Tests*, denen die Nullhypothese zu Grunde liegt, dass kein Unterschied zwischen *zwei* Gruppenmittelwerten besteht.[9] In der Literatur werden vielfältige Varianten von Post-hoc-Tests diskutiert, von denen in SPSS in den Prozeduren zur Varianzanalyse (einfaktorielle, univariate und multivariate Varianzanalyse) allein jeweils 18 Tests implementiert sind. Die verschiedenen Testverfahren lassen sich z. B. danach unterscheiden, ob Varianzhomogenität angenommen werden darf, ob die Fallzahl in den Gruppen gleich ist und ob es sich um eher konservative Tests handelt oder nicht.[10]

[9] Ihren Namen verdanken Post-hoc-Tests dem Umstand, dass sie im Rahmen der Varianzanalyse erst *nach* einem signifikanten F-Test (ex post) durchgeführt werden und vorab keine sachlogischen Hypothesen zu spezifischen Mittelwertunterschieden vorliegen. Ihre Anwendung ist also explorativ, d. h. hypothesengenerierend und erfolgt ad hoc. Demgegenüber sind Kontrastanalysen konfirmatorisch, d. h. hypothesenprüfend.

[10] Einen guten Überblick zu alternativen Post-hoc-Tests und deren inhaltlichen Unterschieden liefert Werner (1997), S. 322 ff. Darstellungen zu den in SPSS implementierten Post-hoc-Tests findet der Leser bei Janssen/Laatz (2017), S. 353 ff. sowie über die Hilfefunktion in SPSS.

Unvollständige Versuchspläne

In unserem Beispiel der Supermarktkette sind wir bisher stets davon ausgegangen, dass ein *vollständiger Versuchsplan* vorliegt, d. h. alle $G \cdot H$ Faktorstufenkombinationen sind besetzt und werden in die Analyse einbezogen. Dieses kann aus verschiedenen Gründen nicht möglich (z. B. fehlende Daten oder inhaltliche Gründe) oder nicht wünschenswert sein, da es zu unnötigen und daher kostspieligen Beobachtungen führt. So kann es z. B. unsinnig sein, bei weiteren Faktorstufen der Verpackung und der Platzierung Kombinationen wie „Lose Ware“ und „Zweitplatzierung“ zu bilden, da lose Ware allein in der Fachabteilung durch Bedienungspersonal verkauft werden kann. Wenn nicht alle Zellen besetzt sind, sind bestimmte Vorkehrungen hinsichtlich der Versuchsanordnung[11] und -auswertung[12] zu treffen.

(Un)vollständiger Versuchsplan

Kovarianzanalyse

Eine Erweiterung der Varianzanalyse liegt in der Einbeziehung von sog. Kovariaten in die Analyse ((M)ANCOVA, (Multivariate) Analysis of Covariance). Kovariate sind metrisch skalierte unabhängige, d. h. erklärende Variablen in einem faktoriellen Design. Häufig ist dem Forscher bewusst, dass es außer den Faktoren noch weitere Einflussgrößen auf die abhängige Variable gibt, deren Einbeziehung sinnvoll und notwendig sein kann. Wenn in unserem Margarine-Beispiel auch die Preise variieren, dann würde die Reststreuung nicht nur zufällige, sondern auch systematische Einflüsse enthalten.[13] Indem der Preis als Kovariate eingeführt wird, kann ein Teil der Gesamtvarianz möglicherweise auf die Variation des Preises zurückgeführt werden, was sich bei Nichterfassung in einer erhöhten Reststreuung (SS_W) ausdrücken würde.

(M)ANCOVA

Kovariaten

Üblicherweise geht die Varianzanalyse bei einem Untersuchungsdesign mit Kovariaten („Kovarianzanalyse“) so vor, dass zunächst der auf die Kovariaten entfallende Varianzanteil ermittelt wird. Dieses entspricht im Prinzip einer vorgeschalteten Regressionsanalyse. Die Beobachtungswerte der abhängigen Variablen werden um den durch die Regressionsanalyse ermittelten Einfluss korrigiert und anschließend der Varianzanalyse unterzogen.[14]

Kovarianzanalyse

Mehrdimensionale Varianzanalyse

Die *mehrdimensionale Varianzanalyse* erlaubt ein Design mit mehr als einer abhängigen Variablen und mehreren Faktoren und Kovariaten. Diese Analyse führt zu einem allgemeinen linearen Modellansatz, der in der Lage ist, nicht nur die Varianzanalyse, sondern auch die Regressionsanalyse und weitere multivariate Verfahren auf ihren gemeinsamen (linearen) Kern zurückzuführen. Eine Darstellung des Algorithmus der mehrdimensionalen Varianzanalyse geht über eine Einführung weit hinaus, sodass hier auf Spezialliteratur verwiesen wird.[15]

Mehrdimensionale Varianzanalyse

[11] Es handelt sich dabei um ein sog. reduziertes Design.

[12] Vgl. Rasch et al. (2014), S. 17 ff.

[13] Zwecks Vermeidung von verzerrten Schätzwerten sollte der Untersucher versuchen, die Preise konstant zu halten und so die Analyse zu vereinfachen. Falls er die Preise experimentell variiert, um deren Effekte zu ermitteln, so sollten die Variationen unabhängig (unkorreliert) von den anderen experimentellen Variablen erfolgen.

[14] Vgl. dazu Diehl (1983); Hair et al. (2010), S. 455 ff. Dadurch wird rechnerisch die Absatzmenge um den Einfluss der Kovariaten bereinigt.

[15] Vgl. zum Allgemeinen Linearen Modell Hartung/Elpert (2007), S. 655 ff., (zur multivariaten Varianzanalyse ebendort S. 667 ff.); Bortz/Schuster (2010); Fahrmeir/Hamerle/Tutz (1996), S. 239 ff., (zur multivariaten Varianzanalyse ebendort S. 228 ff.).

3.3 Fallbeispiel

3.3.1 Problemstellung

Der Betreiber einer Supermarktkette geht aufgrund seiner Erfahrungen davon aus, dass der Margarineabsatz (MENGE) durch die Faktoren Verpackung (VERPACK) und Regalplatzierung (REGAL) sowie die metrisch skalierten Größen Verkaufspreis (PREIS) und durchschnittliche Temperatur im Supermarkt (TEMP) erklärt werden kann. Zur Prüfung seiner Vermutung greift er auf die in Abbildung 3.11 aufgeführten Daten zurück und erweitert diese um Messungen der beiden Kovariaten PREIS und TEMP. Mit der sich ergebenden Datenmatrix (vgl. Abbildung 3.19) möchte er folgende Einzelfragen beantworten:

1. Besitzen die Verpackungsform und die Regalplatzierung einen signifikanten Einfluss auf den Margarineabsatz?
2. *Fall a*: Studien haben ergeben, dass die Zweitplatzierung von Margarine in besonderer Weise den Absatz erhöhen kann. Der Supermarktbetreiber möchte deshalb prüfen, ob dieser Effekt auch in seinem Fall Gültigkeit besitzt.
   Fall b: Der Supermarktbetreiber hat –im Gegensatz zu Fall a– keine Vorstellungen davon, welche Faktorstufen der Regalplatzierung einen besonderen Einfluss auf den Margarineabsatz besitzen. Sollte aber die Varianzanalyse einen signifikanter Einfluss für den Faktor Regalplatzierung bestätigen, so möchte er wissen, ob alle drei Faktorstufen (Normal-, Zweit-, Kühlregal) über einen Einfluss auf die Absatzmenge verfügen und wie stark diese Effekte sind.
3. Kann durch zusätzliche Aufnahme der metrischen Größen PREIS und TEMP die Erklärung des Margarineabsatzes verbessert werden?

Verpackung		„Becher“			„Papier“		
Platzierung		Absatz	Preis	Temp.	Absatz	Preis	Temp.
„Normal-Regal“	Tag 1	47	1,89	16	40	2,13	22
	Tag 2	39	1,89	21	39	2,13	24
	Tag 3	40	1,89	19	35	2,13	21
	Tag 4	46	1,84	24	36	2,09	21
	Tag 5	45	1,84	25	37	2,09	20
„Zweit-Regal“	Tag 1	68	2,09	18	59	2,09	18
	Tag 2	65	2,09	19	57	1,99	19
	Tag 3	63	1,99	21	54	1,99	18
	Tag 4	59	1,99	21	56	2,09	18
	Tag 5	67	1,99	19	53	2,09	18
„Kühl-Regal“	Tag 1	59	1,99	20	53	2,19	19
	Tag 2	50	1,98	21	47	2,19	20
	Tag 3	51	1,98	23	48	2,19	17
	Tag 4	48	1,89	24	50	2,13	18
	Tag 5	53	1,89	20	51	2,13	18

Abbildung 3.19: Datenmatrix des Fallbeispiels

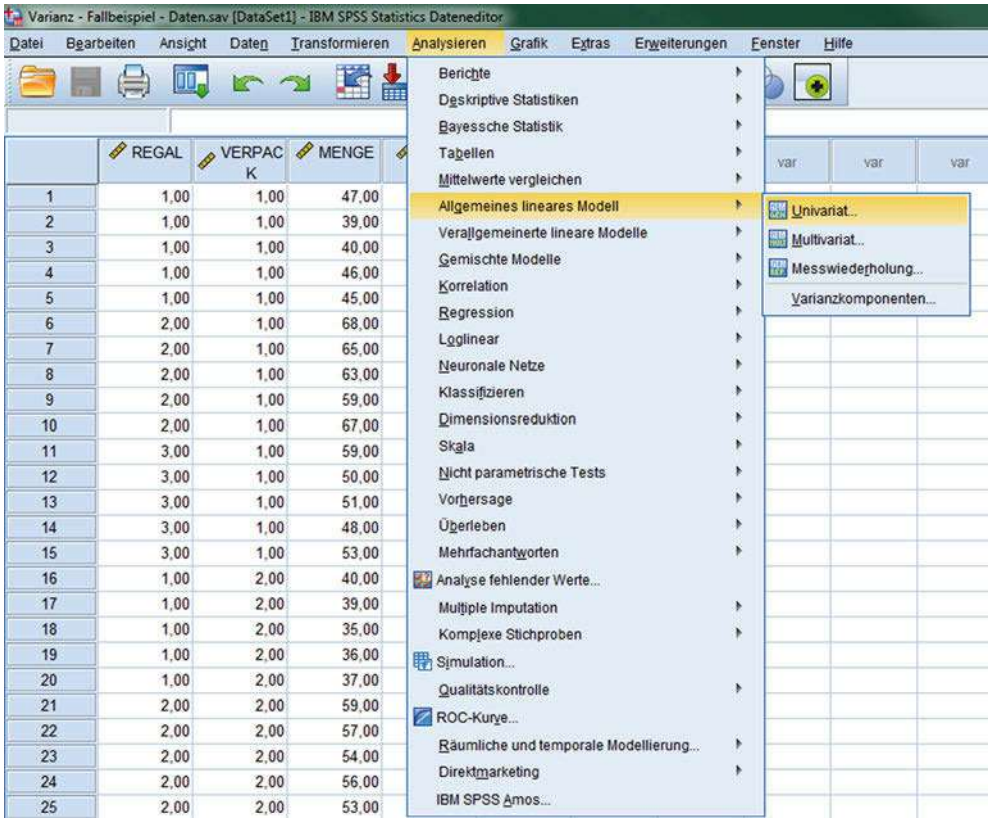


Abbildung 3.20: Daten-Editor mit Auswahl des Analyseverfahrens „Univariat“

Analyse mit Hilfe von SPSS

Die *erste Frage* kann mit Hilfe einer *zweifaktoriellen Varianzanalyse* beantwortet werden. Zu diesem Zweck ist in SPSS unter dem Hauptmenü „Analysieren“ der Unterpunkt „Allgemeines lineares Modell“ und dort die Prozedur „Univariat“ aufzurufen (vgl. Abbildung 3.20).

Im erscheinenden Dialogfeld „Univariat“ sind die abhängige Variable (Absatzmenge) und die beiden unabhängigen, nominal skalierten Variablen (Platzierung und Verpackungsart) aus der Liste auszuwählen und in das Feld „Feste Faktoren“ zu übertragen (vgl. Abbildung 3.21).

Weiterhin können über den Unterpunkt „Optionen“ (vgl. Abbildung 3.22) diverse Statistiken und Kenngrößen ausgewählt werden. Für das Fallbeispiel wurden „Deskriptive Statistiken“, „Homogenitätstests“ und „Schätzungen der Effektgrößen“ ausgewählt.

Frage zwei erfordert im *Fall* a die Durchführung einer Kontrastanalyse und im *Fall b* die Durchführung eines Post-hoc-Tests:
Da der Supermarktbetreiber im *Fall a* nur an einer *Kontrastanalyse* für den Faktor „Regalplatzierung“ interessiert ist, greifen wir bei der Kontrastanalyse auf die einfaktorielle Varianzanalyse zurück. Zu diesem Zweck ist unter dem Hauptmenü „Analysieren“ der Unterpunkt „Mittelwerte vergleichen“ und dort die Prozedur „Einfaktorielle Varianzanalyse“ aufzurufen. Nach Drücken des Felds „Kontraste“ öffnet sich das zugehörige Dialogfenster (vgl. Abbildung 3.23). Im Fall der einfaktoriellen Varianz-

Abbildung 3.21: Dialogfeld „Univariat“

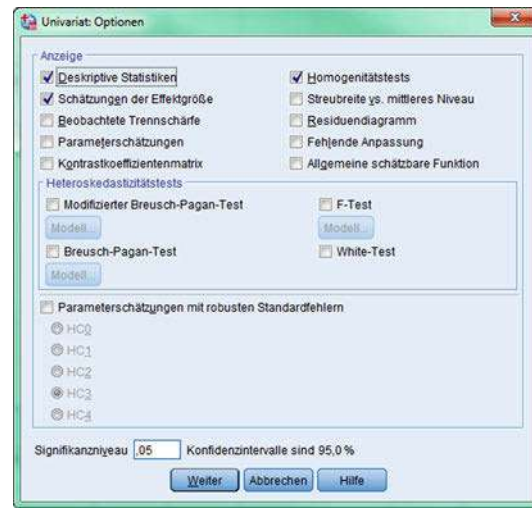


Abbildung 3.22: Dialogfeld „Univariat Optionen“

Kontrast- bzw. Lambda-Koeffizient

analyse vergleicht die Kontrastanalyse die interessierende Kontrastvariable (Faktorstufe) mit den übrigen Faktorstufen, die zu diesem Zweck zu einer Gruppe zusammengefasst werden. Dies wird durch die Festlegung der sog. *Kontrast-Koeffizienten* erreicht, die häufig auch als *Lambda-Koeffizienten* bezeichnet werden. Um die Faktorstufe „Zweitplatzierung“ gegenüber den beiden anderen Faktorstufen möglichst gut zu kontrastieren, wählt der Margarinehersteller hier einen Kontrast-Koeffizienten von -1. Für die verbleibenden Faktorstufen setzt er aufgrund sachlogischer Überlegungen den Kontrast-Koeffizienten für die „Normalplatzierung“ sowie den für die „Platzierung im Kühlregal“ jeweils auf $+0{,}5$. Dadurch wird erreicht, dass die „Zweitplatzierung“ als eigenständige Gruppe betrachtet wird und die Faktorstufen „Normalregal“ und „Kühlregal“ zu einer Gruppe zusammengefasst werden, wobei deren Mittelwerte gleichgewichtig mit jeweils 0,5 in die Berechnung des Mittelwertes für die neue Gruppe „Normal- und Kühlregal“ sowie die Bildung des Gesamtmittelwertes eingehen.[16]

[16] Die absolute Höhe der Ausprägungen der Lambda-Koeffizienten ist unerheblich. Sie geben lediglich das Gewichtungsverhältnis der Mittelwerte (hier 1:1) an. Zu beachten ist, dass die Koeffizienten der zu kontrastierenden Faktorstufen gegensätzliche Vorzeichen aufweisen müssen und die Summe aller Kontrast-Koeffizienten insg. Null ergibt.

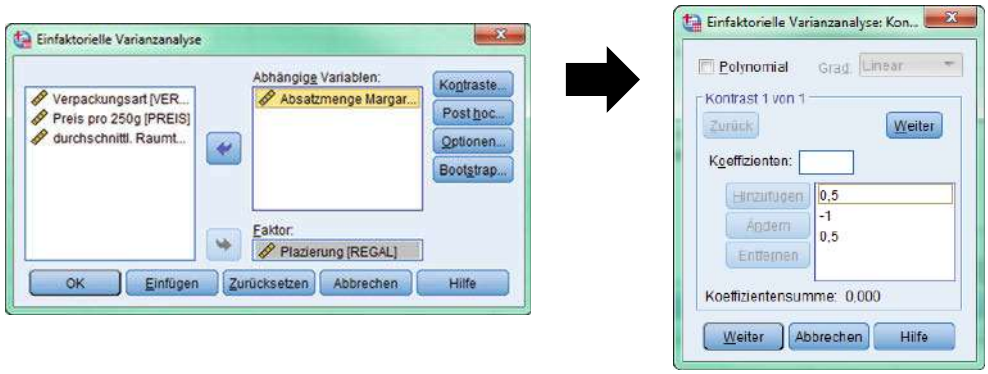


Abbildung 3.23: Dialogfenster „Kontraste“ der einfaktoriellen Varianzanalyse

Im Dialogfeld „Kontraste“ wurden diese Werte jeweils in das Feld „Koeffizienten“ eingetragen und durch Drücken von „Hinzufügen“ für die Analyse übernommen.

Post-hoc-Tests

Zur Beantwortung der Frage im *Fall b* können *Post-hoc-Tests* bei der Prozedur „Einfaktorielle Varianzanalyse“ über den Unterpunkt „Post-hoc“ angefordert werden. In dem sich dann öffnenden Dialogfenster (vgl. Abbildung 3.24) werden unterschiedliche Post-hoc-Tests zur Auswahl angeboten.[17] Insgesamt stehen vierzehn verschiedene Post-hoc-Tests zur Verfügung, wenn Varianzgleichheit in den Gruppen unterstellt und vier Testmöglichkeiten, wenn keine Varianzgleichheit angenommen werden kann. Ob Varianzgleichheit gegeben ist, kann über einen Homogenitätstest geprüft werden, der im Dialogfeld „Optionen“ angefordert werden kann.

Abbildung 3.24: Dialogfenster „Post-hoc-Mehrfachvergleiche“

[17] Kontrastanalysen und Post-hoc-Tests können sowohl mit Hilfe der Prozedur „Einfaktorielle Varianzanalyse“ als auch mit der Prozedur „Univariat“ durchgeführt werden. Während die Möglichkeiten für Post-hoc-Tests in beiden Prozeduren identisch sind, bezieht sich die Kontrastanalyse bei der einfaktoriellen Varianzanalyse allein auf Unterschiede zwischen den Faktorstufen des betrachteten Faktors, während bei der univariaten Varianzanalyse Unterschiede zwischen den betrachteten Faktoren über die Faktorstufen hinweg (sog. Randmittelwerte) betrachtet werden. Außerdem ist zu beachten, dass im Fall der univariaten Varianzanalyse für den Faktor VERPACK *kein* Post-hoc-Test durchgeführt werden kann, da dieser nur zwei Faktorstufen besitzt und Post-hoc-Tests nur für Faktoren mit drei und mehr Faktorstufen definiert sind.

Tukey-HSD-Test
Scheffé-Test

Für das Fallbeispiel wurden der Tukey- und der Scheffé-Test ausgewählt, die zu den gebräuchlichsten Post-hoc-Tests zählen, wobei beide Tests Varianzhomogenität voraussetzen: Der *Tukey-HSD-Test* wird vor allem bei paarweisen Mittelwertvergleichen (wie im Fallbeispiel) empfohlen und als sehr robust eingestuft.[18] Der *Scheffé-Test* ist ebenfalls robust, wird aber gegenüber dem Tukey-Test als *konservativer* bezeichnet.[19] Auch der Scheffé-Test wird in der Literatur als empfehlenswert hervorgehoben und gilt auch für den sog. unbalancierten Fall, d. h. die Stichprobenumfänge in den Gruppen dürfen unterschiedlich sein. Es ist zu beachten, dass die Prozedur „Univariat“ Post-hoc-Tests nur zulässt, wenn *keine* Kovariaten in die Analyse einbezogen sind. Da im Fallbeispiel nur der Faktor REGAL über mehr als zwei Faktorstufen verfügt, hätte Frage 2 auch mit Hilfe der Prozedur „Einfaktorielle Varianzanalyse“ beantwortet werden können, die in SPSS unter dem Hauptmenü „Analysieren“ in Unterpunkt „Mittelwerte vergleichen“ implementiert ist (vgl. Abbildung 3.20). Auch diese Prozedur bietet dann die Durchführung von Post-hoc-Tests an, wobei die Anzahl der Testmöglichkeiten mit denen der Prozedur „Univariat“ identisch sind.

Zur Beantwortung von *Frage 3* sind im Dialogfeld „Univariat“ (vgl. Abbildung 3.21) zusätzlich die metrisch skalierten Variablen PREIS und TEMP in das Feld „Kovariate“ einzufügen. Nach der Übertragung wird automatisch der Unterpunkt „Post hoc“ ausgeblendet, da Post-hoc-Tests nur für Analysen ohne Kovariate definiert sind.
Nach erfolgten Einstellungen in den Untermenüs gelangt der Anwender durch den Button „Weiter“ jeweils wieder zurück zur Prozedur „Univariat“ bzw. „Einfaktorielle Varianzanalyse“, und die Durchführung der Analysen kann durch Drücken von „OK“ gestartet werden.

3.3.2 Ergebnisse

Zweifaktorielle Varianzanalyse ohne Kovariate

Abbildung 3.25 enthält die Ergebnisse der deskriptiven Statistiken, wobei für die jeweiligen Platzierungsarten der durchschnittliche Margarineabsatz in kg pro 1.000 Kassenvorgänge sowie die zugehörige Standardabweichung und die Fallzahlen (N) für die beiden Verpackungsarten angegeben sind.

Für Fragestellung 1 zeigt Abbildung 3.26 die sog. Varianztabelle (Tests der Zwischensubjekteffekte). Diese Tabelle enthält die bereits erhaltenen Werte aus Abschnitt 3.2.2.2 (vgl. Abbildung 3.16 und 3.17).

Zusätzlich sind in Abbildung 3.26 zu den empirischen F-Werten die zugehörigen p-Werte („Signifikanz“) angegeben. Ist der p-Wert kleiner, als das vorgegebene Signifikanzniveau α, so kann die Nullhypothese verworfen werden. Das Nachschlagen in einer Tabelle der F-Verteilung wird dem Benutzer so erspart.

Der Aufbau der Tabelle in Abbildung 3.26 spiegelt sehr deutlich das Grundprinzip der Varianzzerlegung wider: Es findet sich in der ersten Spalte (überschrieben mit „Quelle“) die Gesamtstreuung (Korrigierte Gesamtvariation = SSt) und ihre Zerlegung in die erklärte (Korrigiertes Modell = SS_b) und die nicht erklärte (Fehler = SS_w) Streuung. In der ersten Spalte wird ebenfalls die erklärte Streuung aufgeglie-

Prinzip der Varianzzerlegung

[18] Vgl. Leonhart (2013), S. 419ff. Statistische Tests werden als robust bezeichnet, wenn sie auch bei Verletzungen der Testvoraussetzungen (z. B. bestimmte Verteilungsannahmen, Stichprobenumfänge) noch verlässliche Ergebnisse erbringen. Die Abkürzung HSD steht für „Honestly Significant Difference“ und wird im weiteren Verlauf nicht explizit ausgewiesen.

[19] Statistische Tests werden als *konservativ* bezeichnet, wenn Monte-Carlo-Simulationen (Simulationen mit Zufallszahlen) gezeigt haben, dass die Nullhypothese (hier: Gruppenmittelwerte sind nicht unterschiedlich) häufiger angenommen wird.

Deskriptive Statistiken

Abhängige Variable: Absatzmenge Margarine

Plazierung	Verpackungsart	Mittelwert	Std.-Abweichung	N
Normalregal	Becher	43,4000	3,64692	5
	Papier	37,4000	2,07364	5
	Gesamt	40,4000	4,22164	10
Zweitplazierung	Becher	64,4000	3,57771	5
	Papier	55,8000	2,38747	5
	Gesamt	60,1000	5,36346	10
Kühlregal	Becher	52,2000	4,20714	5
	Papier	49,8000	2,38747	5
	Gesamt	51,0000	3,46410	10
Gesamt	Becher	53,3333	9,58918	15
	Papier	47,6667	8,20859	15
	Gesamt	50,5000	9,23169	30

Abbildung 3.25: Deskriptive Statistiken zum Fallbeispiel

Tests der Zwischensubjekteffekte

Abhängige Variable: Absatzmenge Margarine

Quelle	Quadratsumme vom Typ III	df	Mittel der Quadrate	F	Sig.	Partielles Eta-Quadrat
Korrigiertes Modell	2233,500[a]	5	446,700	45,045	,000	,904
Konstanter Term	76507,500	1	76507,500	7715,042	,000	,997
REGAL	1944,200	2	972,100	98,027	,000	,891
VERPACK	240,833	1	240,833	24,286	,000	,503
REGAL * VERPACK	48,467	2	24,233	2,444	,108	,169
Fehler	238,000	24	9,917			
Gesamt	78979,000	30				
Korrigierte Gesamtvariation	2471,500	29				

a. R-Quadrat = ,904 (korrigiertes R-Quadrat = ,884)

Abbildung 3.26: Ergebnisse der zweifaktoriellen Varianzanalyse ohne Kovariate

dert in die durch die beiden Haupteffekte jeweils einzeln erklärte Streuung (REGAL, VERPACK) sowie die durch die Interaktionseffekte (REGAL*VERPACK) erklärte Streuung. Die übrigen Angaben lassen die Bildung der empirischen F-Statistik (F) nachvollziehen. Sie zeigen die jeweiligen Freiheitsgrade (df) sowie die ermittelten Varianzen (Mittel der Quadrate). Die letzte Spalte weist die über die Option „Schät-

Eta

zungen der Effektgröße“ angeforderte Eta-Statistik aus, welche die Erklärungskraft der einzelnen Faktoren (REGAL, VERPACK) sowie des Interaktionseffektes (REGAL*VERPACK) im Hinblick auf die abhängige Variable angibt. Es handelt sich hierbei um sog. *partielle Eta*^{2}*-Werte*, d. h., der berechnete Erklärungsanteil wird um die Einflüsse der übrigen im Modell enthaltenen Faktoren bereinigt. Für einen beliePartielle Eta^2-Werte bigen Faktor i ergibt sich das Partielle Eta^2 nach der Formel:

$$\text{Partielles Eta}_i^2 = \frac{df_i \cdot F_i}{df_i \cdot F_i + df_{Fehler}} \tag{3.24}$$

Unter Rückgriff auf die in Abbildung 3.26 ausgewiesenen Freiheitsgrade (df) und F-Statistiken (F) können die einzelnen partiellen Eta^2-Werte nun für die Faktoren REGAL und VERPACK sowie für den Interaktionsterm REGAL*VERPACK ($R*V$) nachvollzogen werden:

$$\begin{aligned}\text{Partielles Eta}_{REGAL}^2 &= \frac{df_{REGAL} \cdot F_{REGAL}}{df_{REGAL} \cdot F_{REGAL} + df_{Fehler}} \qquad (3.25)\\ &= \frac{2 \cdot 98,027}{2 \cdot 98,027 + 24} = 0,891\end{aligned}$$

$$\begin{aligned}\text{Partielles Eta}_{VERPACK}^2 &= \frac{df_{VERPACK} \cdot F_{VERPACK}}{df_{VERPACK} \cdot F_{VERPACK} + df_{Fehler}} \qquad (3.26)\\ &= \frac{1 \cdot 24,286}{1 \cdot 24,286 + 24} = 0,503\end{aligned}$$

$$\begin{aligned}\text{Partielles Eta}_{R*V}^2 &= \frac{df_{R*V} \cdot F_{R*V}}{df_{R*V} \cdot F_{R*V} + df_{Fehler}} \qquad (3.27)\\ &= \frac{2 \cdot 2,444}{2 \cdot 2,444 + 24} = 0,169\end{aligned}$$

Die ermittelten partiellen Eta^2-Werte verdeutlichen, dass der Faktor REGAL mit 89,1 % einen größeren Varianzerklärungsanteil aufweist als der Faktor VERPACK (50,3 %). Durch den Interaktionsterm REGAL*VERPACK können 16,9 % der Varianz der abhängigen Variablen erklärt werden.[20]

Kontrastanalyse für den Faktor „Regalplatzierung“

Das Ergebnis zur Beantwortung von Frage 2, Fall a, ist in Abbildung 3.27 dargeKontrastwert-Koeffizienten stellt. Die Matrix der Kontrast-Koeffizienten zeigt nochmals die von uns im Fallbeispiel vorgenommenen Gewichtungen. Der Mittelwertvergleich zwischen der Faktorstufe „Zweitplatzierung“ und der Gruppe „Normal- und Kühlregal“ wird mit Hilfe eines t-Tests sowohl unter der Annahme „gleiche Varianzen“ als auch unter der Annahme „keine gleichen Varianzen“ durchgeführt. Der Kontrastwert spiegelt den Unterschied zwischen den zwei betrachteten Gruppenmittelwerten wider und berechnet sich im Kontrastwert Fallbeispiel wie folgt:[21]

$$\text{Kontrast} = 0,5 \cdot 40,4 + (-1 \cdot 60,1) + 0,5 \cdot 51,0 = -14,4$$

[20] Es sei angemerkt, dass der Interaktionsterm in der Spalte „Sig.“ einen Wert von 0,108 aufweist. Das heißt, für eine Vertrauenswahrscheinlichkeit $> 89,2\,\%$ ist der Einfluss des Interaktionsterms als *nicht* signifikant einzustufen.

[21] Die Gruppenmittelwerte der Faktorstufen können Abbildung 3.25 (jeweils in der Zeile Gesamtsumme) entnommen werden.

Kontrast-Koeffizienten

Kontrast	Plazierung: Normalregal	Zweitplazierung	Kühlregal
1	,5	-1	,5

Kontrast-Tests

		Kontrast	Kontrastwert	Std.-Fehler	T	df	Sig. (2-seitig)
Absatzmenge Margarine	Varianzen sind gleich	1	-14,4000	1,71156	-8,413	27	,000
	Varianzen sind nicht gleich	1	-14,4000	1,90321	-7,566	13,789	,000

Abbildung 3.27: Ergebnisse der Kontrastanalyse

Beide t-Tests führen zum gleichen Ergebnis und sind mit einem p-Wert von 0,000 hoch signifikant. Das bedeutet, dass die Vermutung des Supermarktbetreibers bestätigt werden kann und die Zweitplatzierung von Margarine gegenüber der Platzierung im Normal- oder Kühlregal den Margarineumsatz deutlich erhöht.

Post-hoc-Tests für den Faktor „Regalplatzierung“

Sowohl die Varianzanalyse insgesamt als auch die für das Fallbeispiel angeforderten Post-hoc-Tests setzen Varianzgleichheit in den Faktorstufen (Gruppen) voraus. Diese Annahme kann mit Hilfe des Levene-Tests überprüft werden, der über das Untermenü „Optionen“ (vgl. Abbildung 3.22) angefordert wurde. Dem *Levene-Test* liegt die Nullhypothese zu Grund, dass die Fehlervarianz der abhängigen Variable über die Gruppen hinweg gleich sind. Abbildung 3.28 zeigt das Testergebnis und weist in der letzten Spalte die Signifikanz aus, die im Fallbeispiel Sig.=0,499 beträgt. Die Prüfgröße ist damit *nicht* signifikant, und eine Ablehnung der Nullhypothese wird empirisch nicht gestützt. Das bedeutet, dass keine signifikanten Unterschiede in den Fehlervarianzen zwischen den drei Faktorstufen bestehen. Es kann somit von *Varianzhomogenität* ausgegangen werden.

Levene-Test

Levene-Test auf Gleichheit der Fehlervarianzen[b]

Levene-Statistik	df1	df2	Sig.
,896	5	24	,499

Prüft die Nullhypothese, daß die Fehlervarianz der abhängigen Variablen über Gruppen hinweg gleich ist.

b. Design: Konstanter Term + REGAL + VERPACK + REGAL * VERPACK

Abbildung 3.28: Ergebnis des Levene-Tests auf Varianzhomogenität

Die Ergebnisse für die zwei angeforderten Post-hoc-Tests sind in Abbildung 3.29 dargestellt. Alle zwei Testvarianten führen im Fallbeispiel zum gleichen Ergebnis und weisen signifikante Unterschiede zwischen den paarweisen Mittelwertvergleichen auf. Es zeigt sich, dass alle paarweisen Vergleiche der Mittelwerte der drei Faktorstufen zu signifikanten Ergebnisse führen, was in der Abbildung in der Spalte „Sig.“ direkt abzulesen ist. Darüber hinaus ist eine Signifikanz auf einem Niveau von $\leq 5\%$ ebenfalls in der Spalte „Mittlere Differenz (I-J)“ durch einen Stern gekennzeichnet. Daraus lässt

Post-hoc-Tests

Mehrfachvergleiche

Abhängige Variable: Absatzmenge Margarine

	(I) Plazierung	(J) Plazierung	Mittlere Differenz (I-J)	Std.-Fehler	Signifikanz	95%-Konfidenzintervall Untergrenze	Obergrenze
Tukey-HSD	Normalregal	Zweitplazierung	-19,70000*	1,97634	,000	-24,6002	-14,7998
		Kühlregal	-10,60000*	1,97634	,000	-15,5002	-5,6998
	Zweitplazierung	Normalregal	19,70000*	1,97634	,000	14,7998	24,6002
		Kühlregal	9,10000*	1,97634	,000	4,1998	14,0002
	Kühlregal	Normalregal	10,60000*	1,97634	,000	5,6998	15,5002
		Zweitplazierung	-9,10000*	1,97634	,000	-14,0002	-4,1998
Scheffé-Prozedur	Normalregal	Zweitplazierung	-19,70000*	1,97634	,000	-24,8188	-14,5812
		Kühlregal	-10,60000*	1,97634	,000	-15,7188	-5,4812
	Zweitplazierung	Normalregal	19,70000*	1,97634	,000	14,5812	24,8188
		Kühlregal	9,10000*	1,97634	,000	3,9812	14,2188
	Kühlregal	Normalregal	10,60000*	1,97634	,000	5,4812	15,7188
		Zweitplazierung	-9,10000*	1,97634	,000	-14,2188	-3,9812

*. Die Differenz der Mittelwerte ist auf dem Niveau 0.05 signifikant.

Abbildung 3.29: Ergebnisse der Post-hoc-Tests

sich der Hinweis ableiten, dass alle Faktorstufen unterschiedliche Wirkungen auf den Margarineabsatz haben.

Die Mittelwertdifferenzen zeigen aber auch, dass der Unterschied im durchschnittlichen Margarineabsatz bei den Faktorstufen „Normalregal“ und „Zweitplatzierung“ am größten ist (vgl. die zugehörigen Mittelwerte in Abbildung 3.24). Hier liegt die Mittlere Differenz bei $(40,4 - 60,1 =) -19,7$.

Zweifaktorielle Varianzanalyse mit Kovariaten

Varianzanalyse mit Kovariaten

Zur Beantwortung von *Frage 3* sind im Dialogfeld „Univariat“ (vgl. Abbildung 3.21) zusätzlich die metrisch skalierten Variablen PREIS und TEMP in das Feld „Kova-

Tests der Zwischensubjekteffekte

Abhängige Variable: Absatzmenge Margarine

Quelle	Quadratsumme vom Typ III	df	Mittel der Quadrate	F	Sig.	Partielles Eta-Quadrat
Korrigiertes Modell	2247,511[a]	7	321,073	31,536	,000	,909
Konstanter Term	8,815	1	8,815	,866	,362	,038
PREIS	5,010	1	5,010	,492	,490	,022
TEMP	4,884	1	4,884	,480	,496	,021
REGAL	1207,881	2	603,941	59,319	,000	,844
VERPACK	82,605	1	82,605	8,113	,009	,269
REGAL * VERPACK	13,220	2	6,610	,649	,532	,056
Fehler	223,989	22	10,181			
Gesamt	78979,000	30				
Korrigierte Gesamtvariation	2471,500	29				

a. R-Quadrat = ,909 (korrigiertes R-Quadrat = ,881)

Abbildung 3.30: Zweifaktorielle Kovarianzanalyse mit 2 Kovariaten mittels Prozedur UNIVARIAT

riate“ einzufügen. Damit steht automatisch die Auswahl „Post-hoc“ nicht mehr zur Verfügung.

Durch den erneuten Aufruf der Prozedur und eine neue Analyse zeigt sich das in Abbildung 3.30 dargestellte Ergebnis.

Wiederum finden wir in der ersten Spalte der Tabelle die Zerlegung der Gesamtstreuung in die erklärte Streuung (Korrigiertes Modell) und in die Reststreuung (Fehler). Die mittleren Zeilen zeigen nunmehr in der ersten Spalte eine Aufteilung der durch die Kovariaten und durch die Faktoren erklärten Streuung (Korrigiertes Modell) in ihre jeweiligen Einzelbeiträge (PREIS, TEMP, REGAL, VERPACK, REGAL*VERPACK). Die übrigen Spalten enthalten wie oben die Freiheitsgrade (df), die Varianzen (Mittel der Quadrate), die empirischen F-Werte (F), das Signifikanzniveau der F-Statistik (Signifikanz) sowie die partiellen Eta-Quadrate. Der SPSS-Output verdeutlicht, dass für eine gegebene Vertrauenswahrscheinlichkeit von 95 % der Einfluss der Kovariaten PREIS und TEMP auf die abhängige Variable als nicht signifikant einzustufen ist. d. h. die anfängliche Vermutung, dass die nachgefragte Menge zusätzlich durch diese Faktoren erklärt werden kann, lässt sich nicht bestätigen.

3.3.3 SPSS-Kommandos

```
* MVA: Fallbeispiel Varianzanalyse.
* DATENDEFINITION.
DATA LIST FREE / regal verpack menge preis temp.
VARIABLE LABELS regal "Platzierung"
  /verpack "Verpackungsart"
  /menge "Absatzmenge Margarine"
  /preis "Preis pro 250g"
  /temp "durchschnittl. Raumtemperatur".
VALUE LABELS
  /regal 1 "Normalregal" 2 "Zweitplatzierung" 3 "Kühlregal"
  /verpack 1 "Becher" 2 "Papier".

BEGIN DATA
1 1 47 1,89 16
1 1 39 1,89 21
1 1 40 1,89 19
.....
3 2 51 2,13 18
END DATA.

* PROZEDUREN
* Zweifaktorielle Varianzanalyse für den Margarinemarkt ohne Kovariate
UNIANOVA MENGE BY REGAL VERPACK
  /METHOD=SSTYPE(3)
  /INTERCEPT=INCLUDE
  /POSTHOC=REGAL(TUKEY SCHEFFE)
  /PLOT=PROFILE(REGAL*VERPACK)
  /PRINT=ETASQ HOMOGENEITY DESCRIPTIVE
  /CRITERIA=ALPHA(.05)
  /DESIGN=REGAL VERPACK REGAL*VERPACK.

* Zweifaktorielle Varianzanalyse für den Margarinemarkt mit Kovariaten
UNIANOVA MENGE BY REGAL VERPACK WITH PREIS TEMP
  /METHOD=SSTYPE(3)
  /INTERCEPT=INCLUDE
  /PRINT=ETASQ HOMOGENEITY DESCRIPTIVE
  /CRITERIA=ALPHA(.05)
  /DESIGN=PREIS TEMP REGAL VERPACK REGAL*VERPACK.
```

Abbildung 3.31: SPSS-Job zur Varianzanalyse

3.4 Anwendungsempfehlungen

Voraussetzungen

Um das Instrument der Varianzanalyse anwenden zu können, müssen Voraussetzungen erfüllt sein, die sich sowohl auf die Eigenschaften der erhobenen Daten als auch auf die Auswertung der Daten beziehen. Aus wissenschaftstheoretischer Sicht ist es erforderlich, eine *Hypothese* über den Wirkungszusammenhang der unabhängigen Variablen (z. B. Platzierung) und der abhängigen Variablen (z. B. Absatzmenge) zu formulieren. Die theoretische Frage, die durch die Varianzanalyse beantwortet werden soll, darf sich nicht erst aus den Daten ergeben. Von der Qualität der Hypothese über den Wirkungszusammenhang hängt es ab, ob neben der *statistischen* Signifikanz des Ergebnisses auch eine inhaltlich relevante Aussage formuliert werden kann.

Manipulation Check

Bevor untersucht wird, welche Zusammenhänge zwischen unabhängigen und abhängigen Variablen bestehen, muss geprüft werden, ob die unabhängigen Variablen überhaupt in der vorgesehenen Weise in der Stichprobe realisiert sind.[22] Diese Kontrolle bezeichnet man auch als Manipulation Check. Sie stellt sicher, dass „die auf den abhängigen Variablen festgestellten Merkmalsausprägungen auch auf die unterschiedlichen Faktorstufen der unabhängigen Variablen zurückzuführen sind“[23].

Die Methode stellt bestimmte Anforderungen an die *Auswahl der Daten*. Während unabhängige Variable mit nominalem und auch metrischem Skalenniveau in die Untersuchung eingehen können, müssen die abhängigen Variablen metrisch skaliert sein.

Darüber hinaus erfordern sowohl die MANOVA als auch die Kovarianzanalyse (MANCOVA) die Einhaltung bestimmter Prämissen, deren Prüfung und Heilbarkeit bei gegebener Verletzung in Abbildung 3.32 überblicksartig dargestellt sind.

Schließlich müssen sich die Faktoren eindeutig voneinander unterscheiden, d. h. sie müssen wirklich verschiedene Einflussgrößen der abhängigen Variablen darstellen. Wird nämlich unter zwei vermeintlich unterschiedlichen Faktoren *derselbe* Zusammenhang erhoben (z. B. wenn als Faktoren Verpackung und Markierung gewählt werden, der Käufer beide aber unlösbar gemeinsam wahrnimmt), so lässt sich die Variation der abhängigen Variablen nicht mehr eindeutig auf einen der beiden Faktoren zurückführen.

Der Einstieg in die Varianzanalyse mit Hilfe des SPSS-Programms wird erleichtert, wenn der Anfänger nicht zu viele Faktoren und Kovariaten auf einmal in die Untersuchung einbezieht, da andernfalls die Interpretation der Ergebnisse erschwert wird. Das SPSS-Programm sieht über die Voreinstellungen (DEFAULT) der Prozedur hinaus eine Reihe von weiteren Optionen vor, die nur dann zur Anwendung kommen sollten, wenn der Anwender sich ein genaues Bild von der Wirkungsweise dieser Prozedur-Variationen gemacht hat.

Die Behandlung von Missing Values

Fehlende Werte

Als fehlende Werte (MISSING VALUES) bezeichnet man Variablenwerte, die von den Befragten entweder außerhalb des zulässigen Beantwortungsintervalles vergeben wurden oder überhaupt nicht eingetragen wurden. Im Datensatz können fehlende Werte der Merkmalsvariablen beim Einlesen mit dem Format Fix als Leerzeichen kodiert werden. Sie werden dann vom Programm automatisch durch einen sog. *System-missing value* ersetzt.

[22] Vgl. Kahn (2011), S. 687 ff.
[23] Eschweiler/Evanschitzsky/Woisetschläger (2007), S. 549.

	Prämisse	Prüfungsmethode	Verletzung heilbar über
MANOVA	Keine Ausreißer	Plausibilitätsprüfung der Einträge bei offenen Skalen	Eliminierung
	Randomisierte Zuordnung zu Gruppen	(ex-ante festgelegt)	
	Gruppengröße zumindest größer 20		
	Korrelation zwischen abhängigen Variablen	Signifikanzprüfung über Pearson's R	Anwendung mehrer unabhängiger ANOVAs
	Keine Multikollinearität der abhängigen Variablen	Prüfung der Toleranz	
	Multivariate Normalverteilung	Kolmogorov-Smirnov-Test	Gleichbesetzung der Zellen
	Varianzhomogenität	Levene-Test	Gleichbesetzung der Zellen
Kovarianzanalyse	Keine Beeinflussung der Kovariablen durch experimentelle Anordnung	Plausibilität	
	Kovariable auf intervall-skaliertem Datenniveau	(ex-ante festgelegt)	
	Korrelation der Kovariablen mit abhängigen Variablen	Signifikanzprüfung über Pearson's R	
	Keine Interaktion zwischen Kovariablen und Faktor	Multiple Regressionsanalyse	
	Homogenität der Regressionskoeffizienten	Regressionsanalyse	Gleichbesetzung der Zellen

Abbildung 3.32: Prämissenprüfung der (M)AN(C)OVA im Überblick
(Entnommen aus Eschweiler/Evanschitzsky/Woisetschläger (2007), S. 551.)

Alternativ können fehlende Werte im Datensatz auch durch einen anderen Wert, der unter den beobachteten Werten nicht vorkommt, ersetzt und vom Anwender gewählt werden. Es empfiehlt sich hierfür z. B. den Wert „-99“ zu verwenden. Die zugehörige Anweisung würde in der SPSS-Syntax für die Variabel MENGE im Fallbeispiel dann lauten:
MISSING VALUES Menge (-99).

Derartige vom Benutzer bestimmte fehlende Werte werden von SPSS als *User-missing values* bezeichnet. Für eine Variable lassen sich mehrere Missing Values angeben, z. B. 0 für „Ich weiß nicht“ und 9 für „Antwort verweigert“. Im Rahmen der hier aufgezeigten Varianzanalyse treten allerdings keine fehlenden Werte auf.

In der Voreinstellung werden alle Fälle, die einen fehlenden Wert bei einer oder mehreren Variablen aufweisen, aus den Berechnungen ausgeschlossen (LIST-WISE-Deletion). Durch Verwendung eines entsprechenden Befehls MISSING können aber alle User-missing-values in die Berechnungen eingeschlossen werden.

Literaturhinweise

A. Basisliteratur zur Varianzanalyse

Bleymüller, J./Weißbach R. (2015), Statistik für Wirtschaftswissenschaftler, 17. Auflage, München.

Bortz, J./Schuster, C. (2010), Statistik für Human- und Sozialwissenschaftler, 7. Auflage, Berlin/Heidelberg, Kapitel 12-16.

Fahrmeier, L./Heumann, C./Künstler, R./Pigeot, L./Tutz, G. (2016), Statistik – Der Weg zur Datenanalyse, 8. Auflage, Berlin u. a.

Hair, J./Black, W./Babin, B./Anderson, R. (2010), Multivariate Data Analysis, 7. Auflage, Englewood Cliffs (N.J.), Kapitel 8.

Herrmann, A./Landwehr, J. (2008), Varianzanalyse, in: Herrmann, A./Homburg, C./Klarmann, M. (Hrsg.), Handbuch Marktforschung - Methoden, Anwendungen, Praxisbeispiele, 3. Auflage, Wiesbaden, S. 579–606.

IBM Corporation (2017), IBM SPSS Statistics Base 25, ohne Ort.

Janssen, J./Laatz, W. (2017), Statistische Datenanalyse mit SPSS, 9. Auflage, Berlin u. a.

B. Zitierte Literatur

Bortz, J./Schuster, C. (2010), Statistik für Human- und Sozialwissenschaftler, 7. Auflage, Berlin/Heidelberg, Kapitel 12–16.

Diehl, J. (1983), Varianzanalyse, Frankfurt am Main.

Eschweiler, M./Evanschitzsky, H./Woisetschläger, D. (2007), Ein Leitfaden zur Anwendung varianzanalytisch ausgerichteter Laborexperimente, in: *Wirtschaftswissenschaftliches Studium*, Vol. 36, Nr. 12, S. 546–554.

Fahrmeir, L./Hamerle, A./Tutz, G. (1996), Multivariate statistische Verfahren, 2. Auflage, Berlin.

Hair, J./Black, W./Babin, B./Anderson, R. (2010), Multivariate Data Analysis, 7. Auflage, Englewood Cliffs (N.J.).

Hartung, J./Elpert, B. (2007), Multivariate Statistik, 7. Auflage, München u. a.

Janssen, J./Laatz, W. (2017), Statistische Datenanalyse mit SPSS, 9. Auflage, Berlin u. a.

Kahn, J. (2011), Validation in marketing experiments revisited, in: *Journal of Business Research*, Vol. 64, Nr. 7, S. 687–692.

Leonhart, R. (2013), Lehrbuch Statistik: Einstieg und Vertiefung, 3. Auflage, Bern.

Rasch, B./Friese, M./Hofmann, W./Naumann, E. (2014), Quantitative Methoden 2 – Einführung in die Statistik für Psychologen und Sozialwissenschaftler, 4. Auflage, Heidelberg.

Werner, J. (1997), Lineare Statistik, Weinheim.

4 Diskriminanzanalyse

K. Backhaus, B. Erichson, W. Plinke, R. Weiber, *Multivariate Analysemethoden*,
https://doi.org/10.1007/978-3-662-56655-8_5

4.1 Problemstellung

Die Diskriminanzanalyse ist ein multivariates Verfahren zur Analyse von Gruppenunterschieden. Sie ermöglicht es, die Unterschiedlichkeit von zwei oder mehreren Gruppen hinsichtlich einer Mehrzahl von Variablen zu untersuchen,[1] um Fragen folgender Art zu beantworten:

Fragen

- *„Unterscheiden sich die Gruppen signifikant voneinander hinsichtlich der Variablen?“*
- *„Welche Variablen sind zur Unterscheidung zwischen den Gruppen geeignet bzw. ungeeignet?“*

Beispielsweise kann es sich bei den Gruppen um Käufer verschiedener Marken, Wähler verschiedener Parteien oder Patienten mit verschiedenen Symptomen handeln. Untersuchen lässt sich sodann mittels Diskriminanzanalyse, ob sich die jeweiligen Gruppen hinsichtlich soziodemographischer, psychographischer oder sonstiger Variablen unterscheiden und welche dieser Variablen zur Unterscheidung besonders geeignet oder ungeeignet sind.

Die Anwendung der Diskriminanzanalyse erfordert, dass Daten für die *Merkmalsvariablen* der Elemente (Personen, Objekte) und deren *Gruppenzugehörigkeit* vorliegen.

Strukturen-prüfende Verfahren

Die Diskriminanzanalyse gehört, wie z. B. auch die Regressionsanalyse oder die Varianzanalyse, zur Klasse der *strukturen-prüfenden Verfahren*. Während die Merkmalsvariablen der Elemente metrisch skaliert sein müssen, lässt sich die Gruppenzugehörigkeit durch eine nominal skalierte Variable (Gruppierungsvariable) ausdrücken. Die Diskriminanzanalyse lässt sich damit formal als ein Verfahren charakterisieren, mit dem die *Abhängigkeit einer nominal skalierten Variable* (der Gruppierungsvariable) *von metrisch skalierten Variablen* (den Merkmalsvariablen der Elemente) untersucht wird. Während die Analyse von Gruppenunterschieden primär wissenschaftlichen Zwecken dient, ist ein weiteres Anwendungsgebiet der Diskriminanzanalyse von unmittelbarer praktischer Relevanz. Es handelt sich hierbei um die Bestimmung oder *Prognose der Gruppenzugehörigkeit* von Elementen (Klassifizierung). Die Fragestellung lautet:

„In welche Gruppe ist ein 'neues' Element, dessen Gruppenzugehörigkeit nicht bekannt ist, aufgrund seiner Merkmalsausprägungen einzuordnen?“

Anwendungsbeispiele

Ein illustratives, wenn auch in der praktischen Durchführung nicht ganz unproblematisches *Anwendungsbeispiel* bildet die Kreditwürdigkeitsprüfung.[2] Die Kreditkunden einer Bank lassen sich nach ihrem Zahlungsverhalten in „gute“ und „schlechte“ Fälle einteilen. Mit Hilfe der Diskriminanzanalyse kann sodann geprüft werden, hinsichtlich welcher Variablen (z. B. Alter, Familienstand, Einkommen, Dauer des gegenwärtigen Beschäftigungsverhältnisses oder der Anzahl bereits bestehender Kredite) sich die beiden Gruppen signifikant unterscheiden. Auf diese Weise lässt sich ein Katalog von relevanten (diskriminatorisch bedeutsamen) Merkmalen zusammenstellen. Die Diskriminanzanalyse ermöglicht es weiterhin, die Kreditwürdigkeit

[1] Soll geprüft werden, ob sich zwei Gruppen (Stichproben) hinsichtlich nur eines einzigen Merkmals signifikant unterscheiden, so kann dies durch einen t-Test, und bei mehr als zwei Gruppen mittels Varianzanalyse erfolgen (vgl. dazu Kapitel 3).

[2] Problematisch für die Anwendung der Diskriminanzanalyse bei der Kreditwürdigkeitsprüfung ist, dass die Datenbasis immer vorselektiert ist und daher in der Regel weit weniger „schlechte“ als „gute“ Fälle enthalten wird. Vgl. hierzu z. B.: Häußler (1979), S. B191-B210.

neuer Antragsteller zu überprüfen, wobei, wie noch zu zeigen ist, im Modell der Diskriminanzanalyse die Wahrscheinlichkeit einer *Fehlklassifikation* minimiert wird.

Ganz ähnliche Probleme, wie bei der Kreditwürdigkeitsprüfung, stellen sich z. B. auch dem Personalberater oder der Zulassungsbehörde, der (die) die Erfolgsaussichten von Bewerbern zu beurteilen hat; oder dem Arzt, der eine Frühdiagnose stellen muss; oder dem Archäologen, der einen Schädel gefunden hat und jetzt klären möchte, zu welchem Volksstamm sein Träger wohl gehört haben mag. Die Lösung derartiger Entscheidungsprobleme lässt sich mit Methoden zur Klassifizierung unterstützen oder auch automatisieren. Bei industriellen und kommerziellen Prozessen findet letzteres in zunehmendem Maße statt.

Generell geht es also bei praktischen Problemstellungen der Diskriminanzanalyse um die Feststellung oder Prognose der Gruppenzugehörigkeit von Personen oder Objekten. In Abbildung 4.1 sind einige Anwendungsbeispiele der Diskriminanzanalyse mit Angabe der jeweiligen Gruppierungsvariable und den Merkmalsvariablen zusammengestellt.[3]

Die Diskriminanzanalyse unterscheidet sich hinsichtlich ihrer Problemstellung grundsätzlich von sog. taxonomischen (gruppierenden) Verfahren, wie der Clusteranalyse (siehe dazu Kapitel 8 in diesem Buch), die von ungruppierten Daten ausgehen. Durch die Clusteranalyse werden Gruppen *erzeugt*, durch die Diskriminanzanalyse dagegen werden vorgegebene Gruppen *untersucht*. Beide Verfahren können sich damit sehr gut ergänzen.

Einordnung in Gruppen

In beiden Problembereichen wird von Klassifizierung gesprochen, wobei der Begriff mit unterschiedlicher Bedeutung verwendet wird. Zum einen wird damit die *Bildung von Gruppen* (Taxonomie), zum anderen die *Einordnung von Elementen in vorgegebene Gruppen* gemeint. Im Rahmen der Diskriminanzanalyse findet er mit letzterer Bedeutung Verwendung.

Zur Klassifizierung werden heute äußerst vielfältige Methoden verwendet, die sich hinsichtlich unterschiedlicher Kriterien (z.B. Genauigkeit, Geschwindigkeit, Verständlichkeit und Interpretierbarkeit) beurteilen lassen. Es existiert kein generell bestes Verfahren, und die Eignung hängt auch von den jeweiligen Daten, der spezifischen Problemstellung und den Anforderungen des Nutzers ab.

Eine grobe Einteilung der Verfahren zur Klassifizierung ist die folgende:[4]

- Statistische Verfahren
- Entscheidungsbäume und Regeln
- Neuronale Netze

Statistische Verfahren basieren auf einem zugrundeliegenden Verteilungsmodell und ermöglichen daher Signifikanzbeurteilungen der Ergebnisse. Klassische statistische Verfahren zur Klassifizierung sind die Diskriminanzanalyse und die Logistische Regression.[5] Die Diskriminanzanalyse, deren Grundgedanken von R. A. Fisher (1936)

[3] Auf zahlreiche Anwendungen der Diskriminanzanalyse verweist Lachenbruch (1975). Eine Bibliographie zu Anwendungen der Diskriminanzanalyse im Marketing-Bereich findet sich in: Green/Tull/Albaum (1988), S. 508 f.

[4] Siehe dazu Michie/Spiegelhalter/Taylor (1994); Lim/Loh/Shih (2000), S. 225; Hastie/Tibshirani/Friedman (2009).

[5] Die Logistische Regression wird im nachfolgenden Kapitel dieses Buches behandelt und Neuronale Netze werden in Backhaus/Erichson/Weiber (2015): Fortgeschrittene Multivariate Analysemethoden, behandelt. Zu Entscheidungsbäumen (decision trees) siehe Breiman et al. (1984).

entwickelt wurden, bildet das klassische Verfahren zur Klassifizierung. Zur Unterscheidung von neueren Varianten und Erweiterungen bezeichnet man die Diskriminanzanalyse nach R. A. Fisher auch als Lineare Diskriminanzanalyse (LDA).[6]

Problemstellung	Gruppierung	Merkmalsvariablen
Prüfung der Kreditwürdigkeit	Risikoklasse: -hoch -niedrig	Soziodemographische Merkmale (Alter, Einkommen etc.), Anzahl weiterer Kredite, Beschäftigungsdauer etc.
Auswahl von Außendienstmitarbeitern	Verkaufserfolg: -hoch -niedrig	Ausbildung, Alter, Persönlichkeitsmerkmale, körperliche Merkmale etc.
Analyse der Markenwahl beim Autokauf	Marke: -Mercedes -BMW -Audi etc.	Einstellung zu Eigenschaften von Autos, z. B.: Aussehen, Straßenlage, Geschwindigkeit, Wirtschaftlichkeit etc.
Wähleranalyse	Partei: -CDU -SPD -FDP -Grüne	Einstellung zu politischen Themen wie Abrüstung, Atomenergie, Tempolimit, Besteuerung, Wehrdienst, Mitbestimmung etc.
Diagnose bei Atemnot von Neugeborenen	Überleben: -ja -nein	Geburtsgewicht, Geschlecht, postmenstruales Alter, pH-Wert des Blutes etc.
Erfolgsaussichten von neuen Produkten	Wirtschaftlicher Erfolg: -Gewinn -Verlust	Neuigkeitsgrad des Produktes, Marktkenntnis des Unternehmens, Preis-/Leistungs-Verhältnis, technolog. Know-how etc.
Analyse der Diffusion von Innovationen	Adoptergruppen -Innovatoren -Imitatoren	Risikofreudigkeit, soziale Mobilität, Einkommen, Statusbewusstsein etc.

Abbildung 4.1: Anwendungsbeispiele der Diskriminanzanalyse

Die Lineare Diskriminanzanalyse (LDA) und die Logistische Regression (LR) gehören zu den gebräuchlichsten Verfahren zur Klassifizierung. Beide Verfahren haben große Ähnlichkeit mit der linearen Regressionsanalyse und dies gilt besonders für den Zwei-Gruppen-Fall. Sie basieren wie die Regressionsanalyse auf einem linearen Modell und sind daher relativ einfach anwendbar und gut interpretierbar.

Die Parameter der LDA lassen sich mittels KQ-Methode schätzen, während die LR eine ML-Schätzung erforderlich macht. Daraus folgt, dass die LDA sehr schnell ist, da sich das Schätzproblem analytisch lösen lässt. Dies ist allerdings nur bei sehr großen Datensätzen relevant. Die ML-Schätzung bei der LR macht einen iterativen Prozess erforderlich, bei dem u.U. Konvergenzprobleme auftreten können.

Als ein Vorteil der LR gilt, dass sie auf weniger Annahmen bezüglich der verwendeten Daten basiert als die LDA.[7] Sie gilt daher als flexibler und robuster und ist unempfindlicher gegenüber groben Ausreißern, als die LDA.

[6] Zu modernen Erweiterungen der LDA siehe Hastie/Tibshirani/Friedman (2009).

[7] Die LDA erwartet, dass die Merkmalsvariablen multivariat normalverteilt sind mit annähernd gleicher Kovarianzstruktur in den Gruppen. Die LR dagegen erwartet lediglich eine Multinomialverteilung der Gruppierungsvariable.

Sind allerdings die Annahmen der LDA erfüllt, dann nutzt diese mehr Information aus den Daten und liefert durchgängig effizientere Schätzwerte (mit kleinerer Varianz) als die LR.[8] Die ist besonders bei kleinem Stichprobenumfang (unter 50) von Vorteil. Die LDA ist aber recht unempfindlich gegenüber kleineren Abweichungen von den Annahmen. Die Erfahrung zeigt, dass bei großem Stichprobenumfang beide Verfahren ähnlich gute Ergebnisse liefern, auch wenn die Annahmen der LDA nicht erfüllt sind.[9]

In einer großen Untersuchung von Lim/Loh/Shih (2000) wurden 33 Algorithmen zur Klassifizierung an 32 Datensätzen getestet. Die Lineare Diskriminanzanalyse und die Logistische Regression gehörten dabei zur Gruppe der fünf besten Verfahren. Die Untersucher bemerkten dazu, es sei interessant, dass die alte LDA gegenüber neueren Verfahren so gut abschneide und Ergebnisse erbringt, deren mittlere Fehlerrate nahe bei den Besten liegt. Und dies sei umso erstaunlicher, als die Daten die Annahmen der LDA nicht erfüllten, da sie kategoriale Merkmalsvariablen enthielten.[10]

4.2 Vorgehensweise

Die Durchführung einer Diskriminanzanalyse lässt sich in sechs Teilschritte zerlegen, wie sie das folgende Ablaufdiagramm in Abbildung 4.2 darstellt.

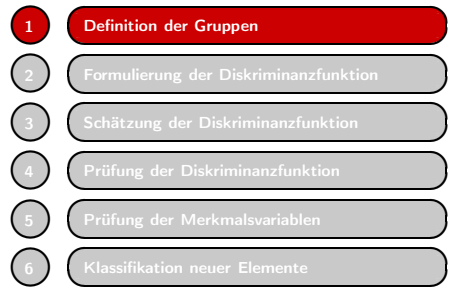
Abbildung 4.2: Ablaufschritte der Diskriminanzanalyse

Gemäß den Stufen dieses Schemas behandeln wir nachfolgend die Diskriminanzanalyse. Zur Illustration wählen wir ein kleines *Beispiel*. Ein Hersteller von Margarine möchte wissen, ob und in welchem Maße die Merkmale „Streichfähigkeit“ und „Haltbarkeit“ bei der Wahl einer Margarinemarke von Bedeutung sind. Insbesondere möchte er herausfinden, ob sich die Stammkäufer der von ihm hergestellten Marke hinsichtlich der Beurteilung dieser Merkmale von den Stammkäufern anderer Marken unterscheiden.

Beispiel

[8] Vgl. Hastie/Tibshirani/Friedman (2009), S. 128; Pohar/Blas/Turk (2004), S. 159.
[9] Vgl. dazu Michie/Spiegelhalter/Taylor (1994), S. 214; Hastie/Tibshirani/Friedman (2009), S. 128; Lim/Loh/Shih (2000), S. 216.
[10] Lim/Loh/Shih (2000), S. 225. Eine andere große Untersuchung, das STATLOG-Projekt von Michie/Spiegelhalter/Taylor (1994), erbrachte ähnliche Ergebnisse.

4.2.1 Definition der Gruppen



Die Durchführung einer Diskriminanzanalyse beginnt mit der Definition der Gruppen. Diese kann sich unmittelbar aus dem Anwendungsproblem ergeben (z. B. Gruppierung von Käufern nach Produktmarken oder von Wählern nach Parteien). Sie kann aber auch das Ergebnis einer vorgeschalteten Analyse sein. So lassen sich z. B. durch die Anwendung der Clusteranalyse Gruppen bilden, die sodann mit Hilfe der Diskriminanzanalyse untersucht werden.

Gruppenanzahl

Mit der Definition der Gruppen ist auch die Festlegung der Anzahl der Gruppen, die in einer Diskriminanzanalyse berücksichtigt werden sollen, verbunden. In unserem Beispiel könnte der Margarinehersteller z. B. für jede existierende Marke eine Gruppe bilden. Die Zahl der Gruppen wäre dann allerdings sehr groß und die Analyse sehr aufwändig.

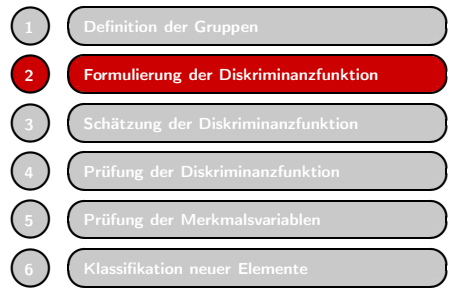
Bei der Definition der Gruppen ist auch das verfügbare Datenmaterial zu berücksichtigen, da die Fallzahlen in den einzelnen Gruppen nicht zu klein werden dürfen. Außerdem sollte die Anzahl der Gruppen nicht größer sein als die Anzahl der Merkmalsvariablen. Unter Umständen kann es daher erforderlich werden, mehrere Gruppen zu einer Gruppe zusammenzufassen.

Zwei-Gruppen-Fall

Den einfachsten Fall bildet die Analyse von zwei Gruppen, auf die wir uns hier zunächst beschränken wollen. So definiert im *Beispiel* unser Margarinehersteller zwei Gruppen A und B, eine Gruppe für die Stammkäufer der von ihm hergestellten Marke A und eine zweite Gruppe für die Stammkäufer der wichtigsten Konkurrenzmarke B. Alternativ hätte er in der zweiten Gruppe auch die Stammkäufer mehrerer oder aller Konkurrenzmarken zusammenfassen können.

Die Gruppen werden zweckmäßigerweise durch eine Gruppierungsvariable bzw. einen Gruppenindex g (g = 1,2,...,G) gekennzeichnet, wobei G die Zahl der Gruppen ist. Im Beispiel gilt damit G = 2 und g = 1, 2 bzw. hier g = A, B.

4.2.2 Formulierung der Diskriminanzfunktion



Im Rahmen der Diskriminanzanalyse ist eine Diskriminanzfunktion zu formulieren und zu schätzen, die sodann eine optimale Trennung zwischen den Gruppen und eine Prüfung der diskriminatorischen Bedeutung der Merkmalsvariablen ermöglicht.

Die *Diskriminanzfunktion* hat allgemein die folgende Form:

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_J X_J \tag{4.1}$$

mit

Y = Diskriminanzvariable
X_j = Merkmalsvariable j (j = 1,2,...,J)
b_j = Diskriminanzkoeffizient für Merkmalsvariable j
b_0 = Konstantes Glied

Die Parameter b_0 und b_j (j = 1,2,...,J) sind auf Basis von Daten für die Merkmalsvariablen zu schätzen. Für jedes Element i (i = 1,...,I_g) einer Gruppe g (g = 1,...,G) mit den Merkmalswerten X_{jgi} (j = 1,...,J) liefert die Diskriminanzfunktion einen Diskriminanzwert Y_{gi}.

Die Diskriminanzfunktion wird auch als kanonische Diskriminanzfunktion und die Diskriminanzvariable Y als kanonische Variable bezeichnet. Der Ausdruck *„kanonisch“* kennzeichnet, dass eine *Linearkombination* von Variablen vorgenommen wird. Wir hatten oben die Diskriminanzanalyse als ein Verfahren charakterisiert, mit dem die Abhängigkeit einer nominal skalierten Variable (der Gruppierungsvariablen oder dem Gruppenindex) von metrisch skalierten Variablen untersucht wird. Die sich ergebende Diskriminanzvariable aber ist eine metrische Variable, da sie durch eine arithmetische Verknüpfung von metrischen Variablen gebildet wird.

Kanonische Funktion

Die Formulierung der Diskriminanzfunktion erfordert die *Auswahl von Merkmalsvariablen*. Diese erfolgt zunächst hypothetisch, d. h. aufgrund von theoretischen oder sachlogischen Überlegungen werden solche Variablen ausgewählt, die mutmaßlich zwischen den Gruppen differieren und somit zur Unterscheidung der Gruppen oder Erklärung der Gruppenunterschiede beitragen können. Nach Schätzung der Diskriminanzfunktion lässt sich sodann die diskriminatorische Eignung der Variablen überprüfen.

Merkmalsauswahl

In unserem *Beispiel* beschränken wir uns auf den einfachsten Fall einer Diskriminanzanalyse, den mit zwei Gruppen und auch nur zwei Merkmalsvariablen. Der Margarinehersteller möchte wissen, ob und in welchem Maße die empfundene Wichtigkeit von *Streichfähigkeit* und *Haltbarkeit* bei der Wahl einer Margarinemarke von Bedeutung ist. Insbesondere möchte er herausfinden, ob sich die Stammkäufer der von ihm hergestellten Marke A hinsichtlich der Beurteilung dieser Merkmale von den Stammkäufern der Konkurrenzmarke B unterscheiden. Es gilt damit:

Gruppen (g = A,B):

A = Stammkäufer von Marke A

B = Stammkäufer von Marke B

Diskriminanzfunktion

$$Y = b_0 + b_1 X_1 + b_2 X_2$$

mit

X_1 = Wichtigkeit der Streichfähigkeit
X_2 = Wichtigkeit der Haltbarkeit

Centroid

Jede Gruppe g lässt sich kompakt durch ihren mittleren Diskriminanzwert, der als *Centroid* (Schwerpunkt) bezeichnet wird, beschreiben:

$$\bar{Y}_g = \frac{1}{I_g} \sum_{i=1}^{I_g} Y_{gi} \tag{4.2}$$

Die *Unterschiedlichkeit zweier Gruppen* $g = A,B$ lässt sich damit durch die Differenz

$$\left|\bar{Y}_A - \bar{Y}_B\right| \tag{4.3}$$

messen. Es wird später gezeigt, wie sich dieses Maß verfeinern und für die Messung der Unterschiedlichkeit von mehr als zwei Gruppen (Mehrgruppenfall) erweitern lässt.

Diskriminanzachse

Die Werte der Diskriminanzfunktion lassen sich auf einer sog. *Diskriminanzachse* abtragen. Einzelne Elemente sowie die Centroide der Gruppen lassen sich damit auf der Diskriminanzachse lokalisieren und die Unterschiede zwischen den Elementen und/oder Gruppen als *Distanzen* repräsentieren. In Abbildung 4.3 sind schematisch die Centroide der Gruppen A und B auf der Diskriminanzachse markiert.

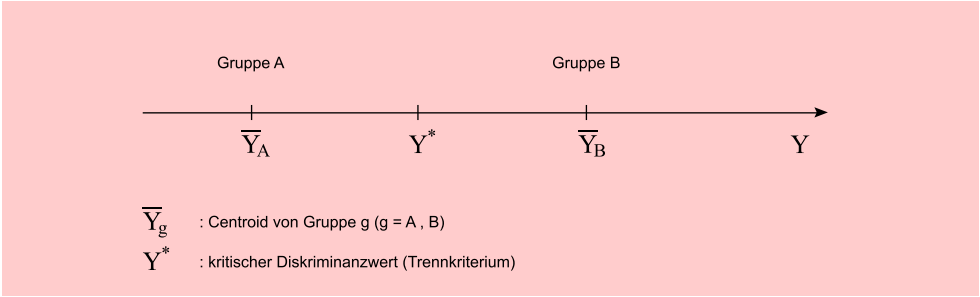


Abbildung 4.3: Diskriminanzachse

Diskriminanzwert

Neben den Gruppen-Centroiden ist auf der Diskriminanzachse in Abbildung 4.3 auch der *kritische Diskriminanzwert* Y^* markiert. Dieser ermöglicht eine Klassifizierung neuer Elemente. Die Einteilung eines Elementes i' mit dem Diskriminanzwert $Y_{i'}$ lässt sich damit wie folgt durchführen:

$$\begin{aligned} Y_{i'} &< Y^* \rightarrow \text{Gruppe A} \\ Y_{i'} &> Y^* \rightarrow \text{Gruppe B} \end{aligned} \tag{4.4}$$

In unserem *Beispiel* könnte der Margarinehersteller auf Basis der Urteilswerte $X_{1i'}$ und $X_{2i'}$ eines Käufers i' prognostizieren, ob dieser Stammkäufer der Marke A oder B ist. Durch Einsetzen in die Diskriminanzfunktion erhält er den Diskriminanzwert $Y_{i'}$. Die Diskriminanzfunktion laute:

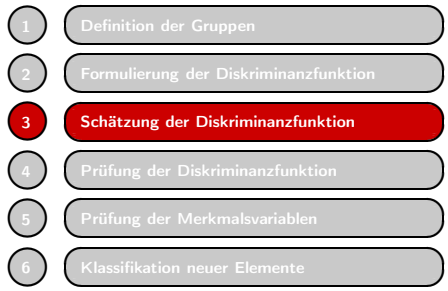
$$Y = -2 + 1,0X_1 - 0,5X_2$$

mit

$$Y^* = 0 \quad \text{(kritischer Wert)}$$

Für einen Käufer i' mit den Urteilswerten $X_{1i'} = 4$ und $X_{2i'} = 6$ erhält man den Diskriminanzwert $Y_{i'} = -1$. Folglich wäre zu prognostizieren, dass diese Person Stammkäufer der Marke A ist.

4.2.3 Schätzung der Diskriminanzfunktion



Die Schätzung der Diskriminanzfunktion (4.1) oder genauer gesagt der unbekannten Parameter b_j in der Diskriminanzfunktion soll so erfolgen, dass sie optimal zwischen den untersuchten Gruppen trennt. Dazu ist ein Kriterium erforderlich, welches die Unterschiedlichkeit der Gruppen misst. Dieses Kriterium wird als *Diskriminanzkriterium* bezeichnet. Die Schätzung erfolgt dann so, dass das Diskriminanzkriterium maximiert wird.

Diskriminanzkriterium

4.2.3.1 Das Diskriminanzkriterium

Als Maß für die Unterschiedlichkeit von Gruppen wurde bereits die Distanz zwischen den Gruppencentroiden eingeführt. Dieses Maß muss jedoch noch verfeinert werden.

Die Unterscheidung zwischen zwei Gruppen ist zwar einerseits umso besser möglich, je größer die Distanz ihrer Centroide ist, andererseits aber wird sie erschwert, wenn die Gruppen stark streuen. Dies zeigt Abbildung 4.4, in der zwei Paare von Gruppen mit gleichem Abstand der Centroide als Verteilungen über der Diskriminanzachse dargestellt sind. Die beiden Gruppen in der unteren Hälfte überschneiden sich stärker, da sie breiter streuen, und lassen sich daher weniger gut unterscheiden.

Unterschiedlichkeitsmaß

Ein besseres Maß der Unterschiedlichkeit (Diskriminanz) erhält man deshalb, wenn auch die Streuung der Gruppen berücksichtigt wird. Wählt man die Standardabweichung s der Diskriminanzwerte als Maß für die Streuung einer Gruppe, so lässt sich das folgende Diskriminanzmaß für zwei Gruppen A und B bilden:

$$\frac{\left|\bar{Y}_A - \bar{Y}_B\right|}{s} \tag{4.5}$$

Dieses Diskriminanzmaß ist allerdings nur unter folgenden *Prämissen* anwendbar:

Voraussetzungen

- nur zwei Gruppen
- annähernd gleiche Streuung s für beide Gruppen.

Diese Prämissen lassen sich aufheben, wenn man das folgende *Diskriminanzkriterium* verwendet:

$$\Gamma = \frac{\text{Streuung zwischen den Gruppen}}{\text{Streuung in den Gruppen}}$$

das sich wie folgt präzisieren lässt:

$$\Gamma = \frac{\sum_{g=1}^{G} I_g \left(\bar{Y}_g - \bar{Y}\right)^2}{\sum_{g=1}^{G} \sum_{i=1}^{I_g} \left(Y_{gi} - \bar{Y}_g\right)^2} = \frac{SS_b}{SS_w} \tag{4.6}$$

Die *Streuung zwischen den Gruppen* wird durch die quadrierten Abweichungen der Gruppencentroide vom Gesamtmittel gemessen und kann so für beliebig viele Gruppen erfolgen. Um unterschiedliche Gruppengrößen zu berücksichtigen, werden die Abweichungen jeweils mit der Gruppengröße I_g multipliziert.

Inner-/Inter-Gruppen-Streuung

Die *Streuung in den Gruppen* wird durch die quadrierten Abweichungen der Gruppenelemente vom jeweiligen Gruppencentroid gemessen.

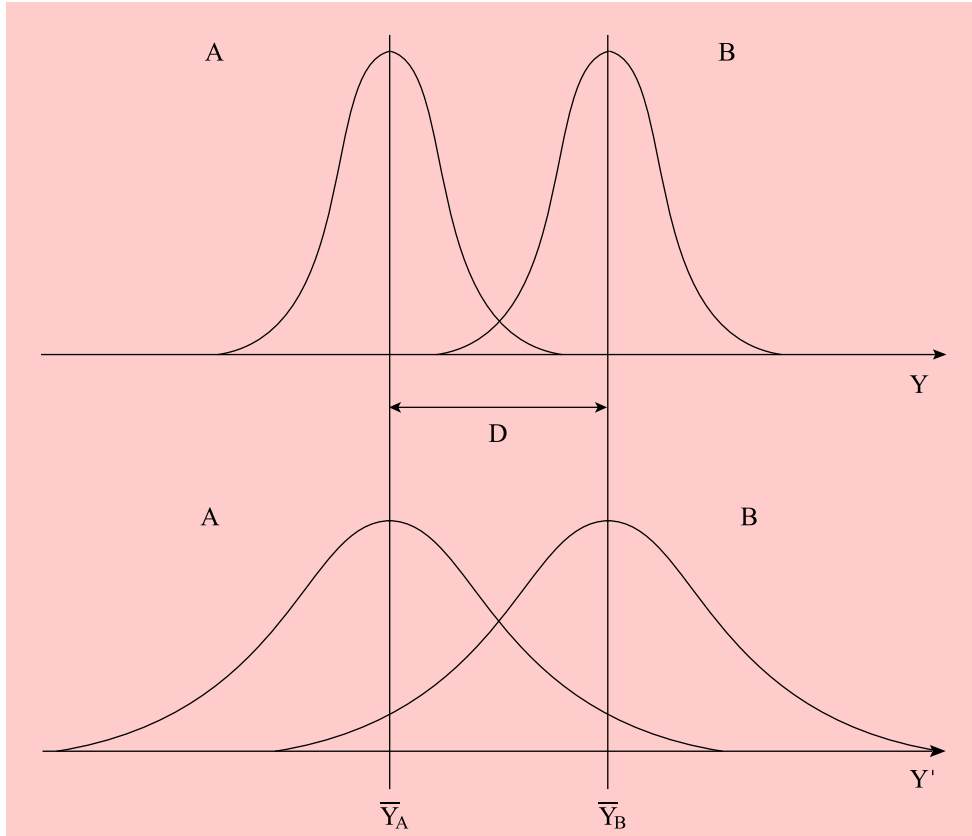


Abbildung 4.4: Gruppen (Verteilungen) mit unterschiedlicher Streuung

Die Streuung zwischen den Gruppen wird gewöhnlich durch SS_b (Sum of Squares *between*) und die Streuung in den Gruppen durch SS_w (Sum of Squares *within*) symbolisiert.

Die Streuung zwischen den Gruppen wird auch als (durch die Diskriminanzfunktion) *erklärte Streuung* und die Streuung in den Gruppen als *nicht erklärte Streuung* bezeichnet. Das Diskriminanzkriterium lässt sich damit auch als Verhältnis von erklärter zu nicht erklärter Streuung interpretieren.

Die *Gesamtstreuung* (Streuung aller Elemente im Gesamtmittel) errechnet sich durch:

$$SS = \sum_{g=1}^{G} \sum_{i=1}^{I_g} \left(Y_{gi} - \bar{Y}\right)^2 \tag{4.7}$$

Zerlegung der Gesamtstreuung

Wie schon in vorherigen Kapiteln bei der Behandlung der Regressionsanalyse oder der Varianzanalyse ausgeführt, gilt folgende *Zerlegung der Gesamtstreuung:*

$$\begin{aligned} SS &= SS_b + SS_w \\ \text{Gesamtstreuung} &= \text{Streuung zwischen den Gruppen} + \text{Streuung in den Gruppen} \\ &= \text{erklärte Streuung} + \text{nicht erklärte Streuung} \end{aligned} \tag{4.8}$$

Die Diskriminanzwerte selbst und damit auch deren Streuungen sind abhängig von den zu bestimmenden Koeffizienten b_j der Diskriminanzfunktion. Das konstante Glied b_0 spielt dabei keine Rolle. Es bewirkt lediglich eine Skalenverschiebung der Diskriminanzwerte, verändert aber nicht deren Streuung. Durch geeignete Wahl von b_0 kann man z. B. bewirken, dass der kritische Diskriminanzwert den Wert Null erhält.

Optimierung

Die Schätzung der Diskriminanzfunktion beinhaltet damit das folgende *Optimierungsproblem*:

$$\max_{b_1,\ldots,b_j} \{\Gamma\} \tag{4.9}$$

Wähle die Koeffizienten b_j ($j = 1,\ldots,J$) so, dass das Diskriminanzkriterium Γ maximal wird. Die mathematische Lösung dieses Optimierungsproblems wird im Anhang (Teil A) dieses Kapitels ausgeführt.[11]

4.2.3.2 Rechenbeispiel

Beispiel

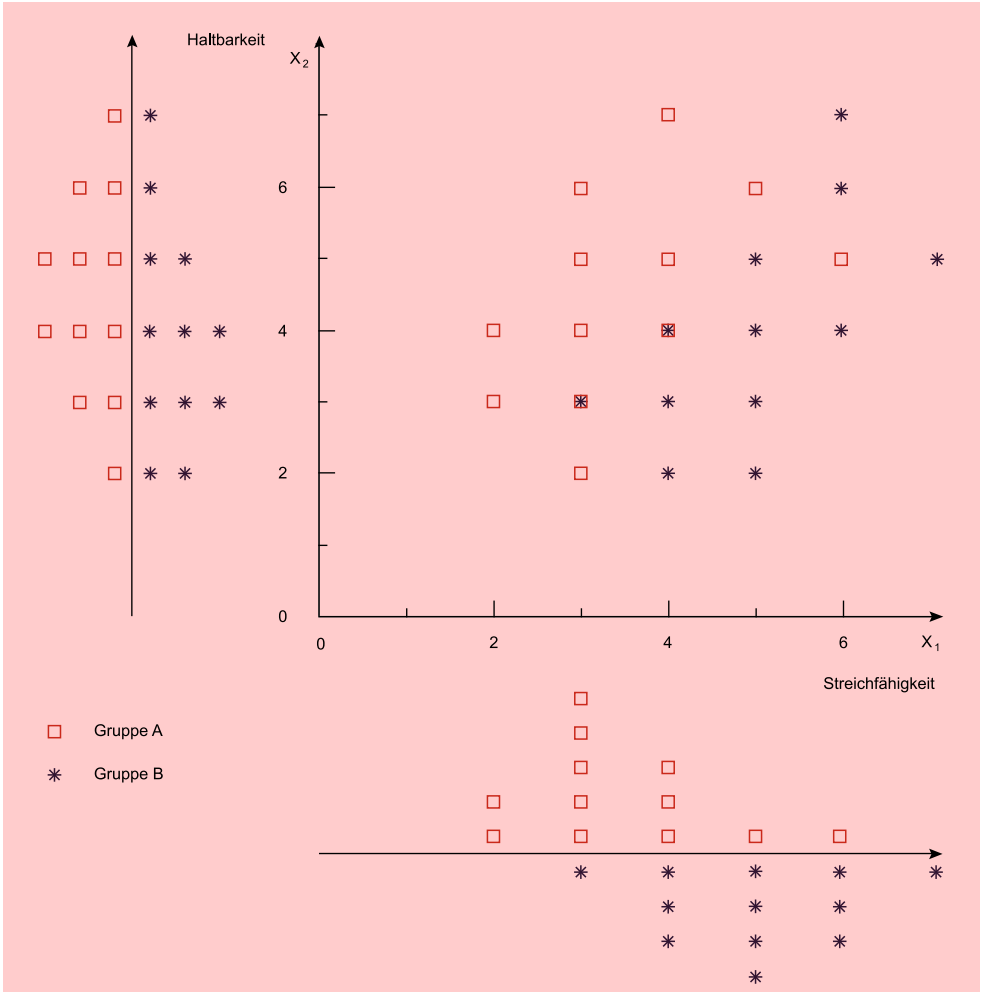
Die Schätzung der Diskriminanzfunktion soll nachfolgend an einem kleinen Rechenbeispiel demonstriert werden. Unser Margarinehersteller, der herausfinden möchte, welche Bedeutung die Merkmale „Streichfähigkeit“ und „Haltbarkeit“ für die Markenwahl haben, lässt jeweils 12 Stammkäufer der Marken A und B befragen. Jede der 24 Personen wird gebeten, die empfundene Wichtigkeit der beiden Merkmale auf einer siebenstufigen Rating-Skala zu beurteilen. Die Daten sind in Abbildung 4.5 wiedergegeben.

Stammkäufer von Marke A			Stammkäufer von Marke B		
Person	Streichfähigkeit	Haltbarkeit	Person	Streichfähigkeit	Haltbarkeit
i	X_{1Ai}	X_{2Ai}	i	X_{1Bi}	X_{2Bi}
1	2	3	13	5	4
2	3	4	14	4	3
3	6	5	15	7	5
4	4	4	16	3	3
5	3	2	17	4	4
6	4	7	18	5	2
7	3	5	19	4	2
8	2	4	20	5	5
9	5	6	21	6	7
10	3	6	22	5	3
11	3	3	23	6	4
12	4	5	24	6	6

Abbildung 4.5: Ausgangsdaten für das Rechenbeispiel (zwei Gruppen, zwei Variablen)

In Abbildung 4.6 ist das Ergebnis der Befragung als Streudiagramm dargestellt. Jede der 24 befragten Personen ist entsprechend der abgegebenen Urteilswerte im Raum der beiden Variablen als Punkt repräsentiert. Dabei sind die Käufer von Marke A durch Quadrate und die der Marke B durch Sterne markiert.

[11] Zur Mathematik der Diskriminanzanalyse vgl. insbesondere Tatsuoka (1988), S. 210 ff.; Cooley/ Lohnes (1971), S. 243 ff.; Kendall (1980).



Diskriminanz-Plot

Abbildung 4.6: Streuung der Urteilswerte in den beiden Gruppen

In Abbildung 4.6 sind außerdem die Häufigkeitsverteilungen (Histogramme) der Urteilswerte bezüglich jeder der beiden Variablen neben bzw. unter dem Streudiagramm dargestellt. Man ersieht daraus, dass die Stammkäufer von Marke B die Wichtigkeit der Streichfähigkeit tendenziell höher einstufen als die Stammkäufer von Marke A. Dagegen ergeben sich für die Käufer der Marke A im Durchschnitt etwas höhere Werte bei der Einstufung der Haltbarkeit. Infolge der erheblichen Überschneidungen der Häufigkeitsverteilungen aber ermöglicht keine der beiden Variablen eine gute Trennung zwischen den Käufergruppen. Offenbar aber besitzt die „Streichfähigkeit“ eine größere diskriminatorische Bedeutung.

Urteilswerte

Es soll jetzt geprüft werden, ob die Diskriminanzfunktion

$$Y = b_0 + b_1 X_1 + b_2 X_2$$

eine bessere Trennung zwischen den Gruppen ermöglicht.

Gruppenstreuung

Die Auswertungstabellen der Abbildungen 4.7 - 4.10 zeigen die Berechnung der Streuung der beiden Merkmalsvariablen in und zwischen den Gruppen. Mit diesen Werten lässt sich die optimale *Diskriminanzfunktion* berechnen (vgl. mathematischer Anhang A).
Sie lautet:

$$Y = -1,98 + 1,031X_1 - 0,565X_2 \tag{4.10}$$

Gruppe g:	Marke A		Marke B	
Variable j :	Streich-fähigkeit X_{1A}	Haltbar-keit X_{2A}	Streich-fähigkeit X_{1B}	Haltbar-keit X_{2B}
$\bar{X}_{jg} = \frac{1}{I_g}\sum_{i=1}^{I_g} X_{jgi}$	3,5	4,5	5,0	4,0
$SS_{jg} = \sum_{i=1}^{I_g}\left(X_{jgi} - \bar{X}_{jg}\right)^2$	15,0	23,0	14,0	26,0
$SC_{12g} = \sum_{i=1}^{I_g}\left(X_{1gi} - \bar{X}_{1g}\right)\cdot\left(X_{2gi} - \bar{X}_{2g}\right)$	9,0		12,0	

Abbildung 4.7: Gruppenspezifische Maße der Merkmalsvariablen
$\bar{X}_{jg}$ = Mittelwert von Variable j in Gruppe g
SS_{jg} = Quadratsumme der Abweichungen vom Mittelwert (Sum of Squares)
SC_{12} = Kreuzproduktsumme der Abweichungen (Sum of Cross-Products)

Durch Einsetzen der Daten aus Abbildung 4.5 in die Diskriminanzfunktion (4.10) erhält man die Diskriminanzwerte in Abbildung 4.11. Beispielsweise ergibt sich für den ersten Stammkäufer der Marke A:

$$Y = -1,98 + 1,031 \cdot 2 - 0,565 \cdot 3 = -1,614$$

Mit diesen Werten und unter Anwendung der Formel (4.6) erhält man für das Diskriminanzkriterium Γ den Wert:

$$\Gamma = \frac{SS_b}{SS_w} = \frac{20,07}{22,0} = 0,912$$

Zum Vergleich berechnen wir beispielhaft, welche Werte sich für das Diskriminanzkriterium mit anderen Werten der Koeffizienten ergeben würden. In Abbildung 4.12 sind einige Werte der Koeffizienten b_1 und b_2 mit dem jeweiligen Wert für das Diskriminanzkriterium zusammengestellt. Der Wert von b_0 hat keinen Einfluss auf das Diskriminanzkriterium und kann hier somit auch auf Null gesetzt werden. Die Koeffizienten in Abbildung 4.12 wurden zwecks besserer Übersicht so normiert, dass ihre Absolutwerte sich zu eins addieren:

Normierung

$$|b_1| + |b_2| = 1$$

Variable j :		Streichfähigkeit X_1	Haltbarkeit X_2
W_{jj}	$= \sum_{g=1}^{G} \sum_{i=1}^{I_g} \left(X_{jgi} - \bar{X}_{jg}\right)^2$	$15 + 14 = 29$	$23 + 26 = 49$
	$= SS_{jA} + SS_{jB}$		
W_{12}	$= \sum_{g=1}^{G} \sum_{i=1}^{I_g} \left(X_{1gi} - \bar{X}_{1g}\right) \cdot \left(X_{2gi} - \bar{X}_{2g}\right)$	$9 + 12 = 21$	
	$= SC_{12A} + SC_{12B}$		

Abbildung 4.8: Innergruppen-Streuungsmaße der Merkmalsvariablen
W_{jj} = Within Sum of Squares
W_{12} = Within Sum of Cross Products

Variable j :	Streichfähigkeit X_1	Haltbarkeit X_2
$\bar{X} = \frac{1}{I} \sum_{i=1}^{I} X_{ji}$	$4,25$	$4,25$

Abbildung 4.9: Gesamtmittelwerte der Merkmalsvariablen

Variable j :	Streichfähigkeit X_1	Haltbarkeit X_2
$B_{jj} = \sum_{g=1}^{G} I_g \left(\bar{X}_{jg} - \bar{X}_j\right)^2$	$12(3,5 - 4,25)^2$ $+12(5,0 - 4,25)^2$ $= 13,5$	$12(4,5 - 4,25)^2$ $+12(4,0 - 4,25)^2$ $= 1,5$
$B_{12} = \sum_{g=1}^{G} I_g (\bar{X}_{1g} - \bar{X}_1) \cdot (\bar{X}_{2g} - \bar{X}_2)$	$12(3,5 - 4,25)(4,5 - 4,25)$ $+12(5,0 - 4,25)(4,0 - 4,25)$ $= -4,5$	

Abbildung 4.10: Zwischengruppen-Streuungsmaße der Merkmalsvariablen
B_{jj} = Between Sum of Squares
B_{12} = Between Sum of Cross Products

Erläuterung

Abbildung 4.12 verdeutlicht, dass sich keine Kombination von Koeffizienten finden lässt, die einen höheren Wert als 0,912 für das Diskriminanzkriterium liefert. Die Koeffizienten 0,646 und -0,354 sind proportional zu den Koeffizienten in (4.10), d. h. sie unterscheiden sich von diesen nur um einen konstanten Faktor:

$$\frac{1,031}{0,646} = \frac{-0,565}{-0,354} = 1,6$$

Person i	Marke A Y_{Ai}	Person i	Marke B Y_{Bi}
1	-1,614	13	0,914
2	-1,148	14	0,448
3	1,381*	15	2,412
4	-0,117	16	-0,583*
5	-0,018	17	-0,117*
6	-1,810	18	2,044
7	-1,712	19	1,013
8	-2,179	20	0,350
9	-0,215	21	0,252
10	-2,277	22	1,479
11	-0,583	23	1,946
12	-0,681	24	0,816
$\bar{Y}_A$	-0,914	$\bar{Y}_B$	0,914
SS_{wA}	12,8	SS_{wB}	9,2

Abbildung 4.11: Diskriminanzwerte der Markenbeurteilungen sowie deren Mittelwerte und Standardabweichungen

Diskriminanzkoeffizienten		Diskriminanzkriterium
b_1	b_2	Γ
1	0	0,466
0	1	0,031
0,5	0,5	0,050
0,5	-0,5	0,667
0,6	-0,4	0,885
0,646	-0,354	0,912*
0,7	-0,3	0,882
0,8	-0,2	0,735
0,9	-0,1	0,582

Abbildung 4.12: Werte des Diskriminanzkriteriums für unterschiedliche Werte der Diskriminanzkoeffizienten

Sie liefern daher ebenfalls den maximalen Wert für das Diskriminanzkriterium.

Für die Werte $b_1 = 1$ und $b_2 = 0$ ist die Diskriminanzvariable identisch mit Variable 1 (Streichfähigkeit) und für die Werte $b_1 = 0$ und $b_2 = 1$ ist sie identisch mit Variable 2 (Haltbarkeit). Abbildung 4.12 zeigt also in den ersten beiden Zeilen die isolierte Diskriminanz der beiden Merkmalsvariablen. Wie schon aus Abbildung 4.6 ersichtlich, besitzt die Variable Streichfähigkeit eine erheblich größere Trennschärfe für die Markenwahl als die Variable Haltbarkeit. Bei optimaler Verknüpfung der beiden Variablen aber lässt sich die Trennschärfe fast verdoppeln.

Um die isolierte Diskriminanz der beiden Merkmalsvariablen zu bestimmen, kann auf die Auswertungstabellen der Abbildungen 4.7 - 4.10 zurückgegriffen werden. Abbildung 4.13 zeigt das Ergebnis.

Streichfähigkeit X_1	Haltbarkeit X_2
$SS_b = B_{11} = 13,5$	$SS_b = B_{22} = 1,5$
$SS_w = W_{11} = 29,0$	$SS_w = W_{22} = 49,0$
$\Gamma_1 = \frac{13,5}{29,0} = 0,466$	$\Gamma_2 = \frac{1,5}{49,0} = 0,031$

Abbildung 4.13: Isolierte Diskriminanz der beiden Merkmalsvariablen

Die Variable Haltbarkeit weist eine niedrige Streuung zwischen den Gruppen (erklärte Streuung) und eine hohe Streuung in den Gruppen (nichterklärte Streuung) auf. Ihre Diskriminanz ist daher mit 0,031 minimal.

4.2.3.3 Geometrische Ableitung

Die Diskriminanzfunktion bildet geometrisch gesehen eine Ebene (für $J = 2$) bzw. Hyperebene (für $J > 2$) *über* dem Raum, der durch die J Merkmalsvariablen gebildet wird. Sie lässt sich aber auch als eine Gerade im Raum (Koordinatensystem) der Merkmalsvariablen repräsentieren, die als *Diskriminanzachse* bezeichnet wird.

Diskriminanzachse

Für die Diskriminanzfunktion

$$Y = b_0 + b_1 X_1 + b_2 X_2$$

bildet die Diskriminanzachse eine Gerade der Form

$$X_2 = \frac{b_2}{b_1} \cdot X_1 \tag{4.11}$$

Sie verläuft durch den Nullpunkt des Koordinatensystems und ihre Steigung bzw. Neigung wird durch das Verhältnis der Diskriminanzkoeffizienten bestimmt. Die Diskriminanzachse ist so mit einer Skala zu versehen, dass die Projektion eines beliebigen Punktes (X_1, X_2) gerade den zugehörigen Diskriminanzwert Y liefert. Abbildung 4.14 zeigt die der optimalen Diskriminanzfunktion (4.10) zugehörige Diskriminanzachse im Raum der beiden Merkmalsvariablen. Es sind außerdem die Häufigkeitsverteilungen der Diskriminanzwerte beider Gruppen dargestellt. Wie man sieht, weisen die Häufigkeitsverteilungen der Diskriminanzwerte eine geringere Überschneidung auf, als die Häufigkeitsverteilungen der Merkmalswerte in Abbildung 4.6, worin die höhere Trennschärfe der Diskriminanzfunktion zum Ausdruck kommt.

Diskriminanzwert

Die Diskriminanzwerte wurden durch Wahl von b_0 so skaliert, dass der kritische Diskriminanzwert gerade Null ist. Man sieht, dass nur ein Element von Gruppe A rechts vom kritischen Diskriminanzwert und zwei Elemente von Gruppe B links davon liegen. Insgesamt werden also nur noch drei Elemente falsch klassifiziert. Diese Elemente sind in Abbildung 4.11 mit einem Stern gekennzeichnet.

Die Diskriminanzachse lässt sich bei gegebener Diskriminanzfunktion sehr einfach konstruieren. Man braucht nur für einen beliebigen Wert z den Punkt $(b_1 \cdot z, b_2 \cdot z)$ in das Koordinatensystem einzutragen und mit dem Nullpunkt zu verbinden.

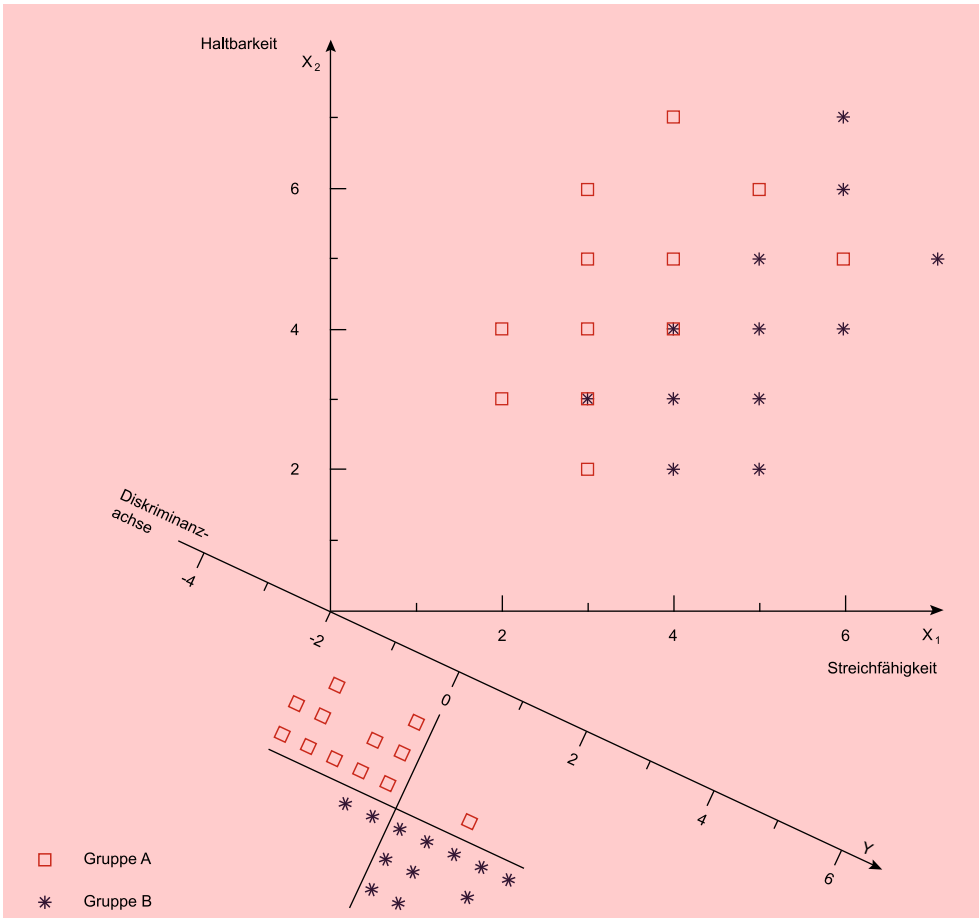


Abbildung 4.14: Darstellung der optimalen Diskriminanzachse

Die sich so ergebende Gerade bildet die Diskriminanzachse. Für $z = 4$ erhält man z. B. den Punkt

$$X_1 = \ 1,031 \cdot 4 = \ 4,12$$
$$X_2 = -0,565 \cdot 4 = -2,26$$

Diskriminanzfunktion

Dessen Koordinaten sind in Abbildung 4.15 durch Linien markiert. Die Diskriminanzfunktion (4.10) liefert für diesen Punkt den Wert $Y = 3,54$, der auf der Diskriminanzachse verzeichnet ist. Im Koordinatenursprung gilt $Y = b_0 = -1,98$. Durch diese beiden Werte ist die Skala auf der Diskriminanzachse determiniert.

Damit lassen sich die Diskriminanzwerte beliebiger Punkte durch Projektion auf die Diskriminanzachse ermitteln. Beispielweise ergibt sich für Element 7 aus Gruppe B (Person 19) ein Wert nahe Eins. Der genaue Wert, der sich aus Abbildung 4.11 entnehmen lässt, beträgt 1,013.

Durch die Diskriminanzfunktion wird die Steigung bzw. Neigung der Diskriminanzachse bestimmt. Eine Veränderung des Quotienten der Diskriminanzkoeffizienten b_2/b_1 bewirkt somit eine Rotation der Diskriminanzachse um den Koordinatenursprung (vgl. Abbildung 4.16).

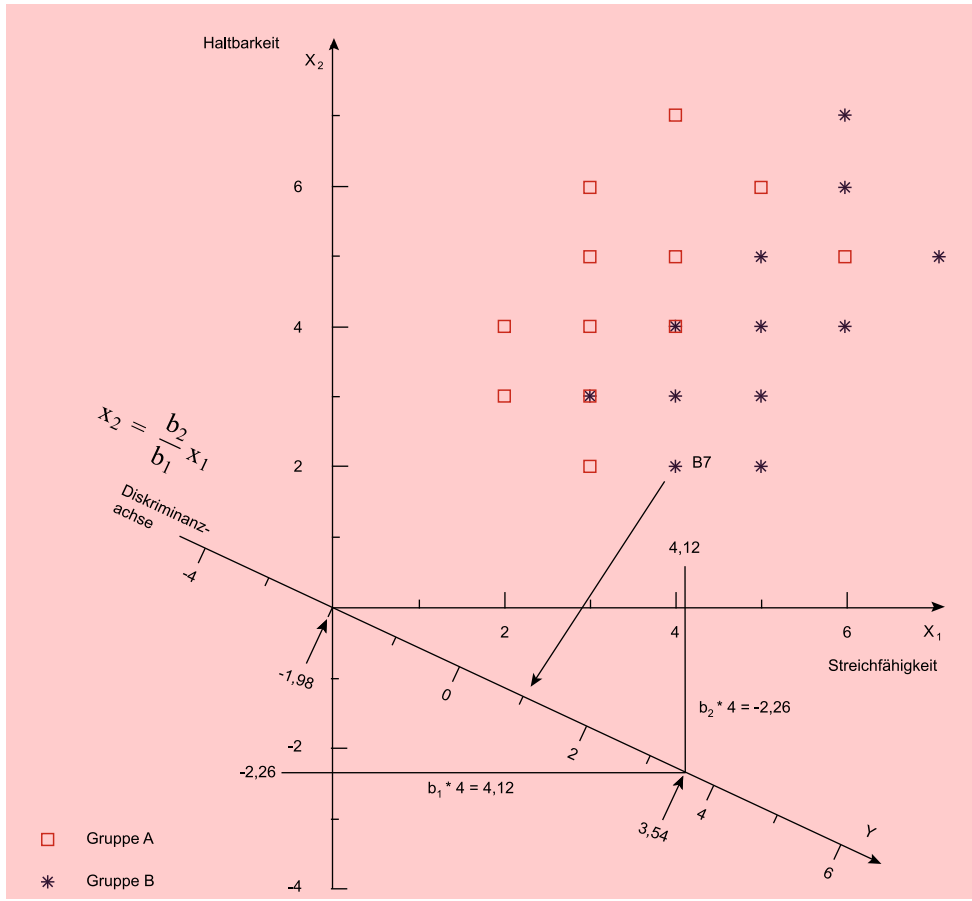


Abbildung 4.15: Konstruktion der Diskriminanzachse

Umgekehrt könnte man somit, zumindest angenähert, die optimale Diskriminanzfunktion auch geometrisch durch Rotation der Diskriminanzachse ermitteln.

In Abbildung 4.16 sind die Projektionen der Gruppencentroide auf die Diskriminanzachsen Y_2 und Y_4 eingezeichnet. An der größeren Distanz zwischen den Projektionspunkten ist zu erkennen, dass die Achse Y_2 besser diskriminiert als die Achse Y_4. Noch schlechter diskriminiert die Achse Y_3 und noch besser die Achse Y_1. Da die Diskriminanzachse Y_1 parallel zur Verbindungslinie der Gruppencentroide verläuft, wird auf ihr die Distanz der Projektionspunkte maximal. Sie bildet somit die optimale Diskriminanzachse.

Graphische Darstellung

Die geometrische Ermittlung der optimalen Diskriminanzfunktion ist allerdings nicht mehr durchführbar, wenn, was gewöhnlich der Fall ist, mehr als zwei Merkmalsvariablen vorliegen. Überdies ist die Distanz der Gruppencentroide nur dann ein geeignetes Diskriminanzkriterium, wenn nur zwei Gruppen mit annähernd gleicher Streuung betrachtet werden.

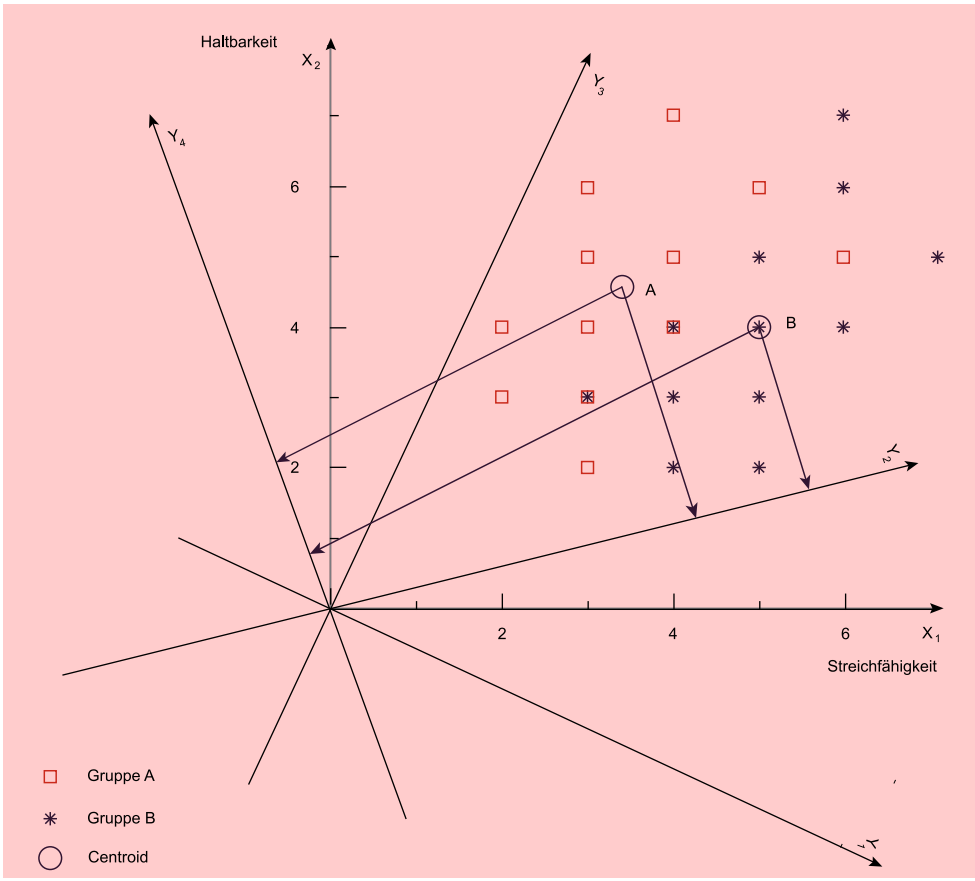


Abbildung 4.16: Rotation der Diskriminanzachse und Projektionen der Gruppencentroide

4.2.3.4 Normierung der Diskriminanzfunktion

Durch die Maximierung des Diskriminanzkriteriums wird nur das Verhältnis der Diskriminanzkoeffizienten b_2/b_1 bestimmt. Multipliziert man die Koeffizienten mit einem konstanten Faktor, so ändern sich dadurch zwar die Diskriminanzwerte und auch die Skaleneinheit auf der Diskriminanzachse, der Wert des Diskriminanzkriteriums wie auch die Lage der Diskriminanzachse aber ändern sich nicht. Die Werte der Koeffizienten sind also nicht eindeutig bestimmt. Eine Veränderung des konstanten Gliedes b_0 bewirkt ebenfalls nur eine Veränderung der Diskriminanzwerte bzw. eine Verschiebung der Skala auf der Diskriminanzachse. Der Wert von b_0 bestimmt die Entfernung des Nullpunktes der Skala vom Nullpunkt des Koordinatensystems.

Normierung

Zwecks Erzielung eindeutiger Werte für die Parameter der Diskriminanzfunktion ist daher eine *Normierung* erforderlich. Diese erfolgt mehr oder minder willkürlich nach Zweckmäßigkeitsgründen. Es existieren daher unterschiedliche Konventionen, unter denen sich die folgende durchgesetzt hat: Die Diskriminanzkoeffizienten werden so normiert, dass die *Innergruppen-Varianz* aller Diskriminanzwerte (pooled within-groups variance) Eins ergibt. Sie errechnet sich, indem man die Streuung in den

Gruppen durch die Zahl der Freiheitsgrade dividiert:

$$s^2 = \frac{SS_w}{I - G} \tag{4.12}$$

Zwei-Gruppen-Fall

Anschließend wird der Wert von b_0 so gewählt, dass der Gesamtmittelwert der Diskriminanzwerte Null wird. Dadurch erhält im Normalfall auch der kritische Diskriminanzwert Y^* für den Zwei-Gruppen-Fall den Wert Null.[12]

4.2.3.5 Vergleich mit der Regressionsanalyse

Die Diskriminanzanalyse wurde oben formal als ein Verfahren charakterisiert, mittels dessen eine nominal skalierte Variable (die Gruppierungsvariable) durch eine Mehrzahl von metrisch skalierten Variablen (den Merkmalsvariablen) erklärt oder prognostiziert werden soll. Im Unterschied dazu ist bei der Regressionsanalyse auch die abhängige Variable metrisch skaliert.

Da sich eine binäre Variable formal immer wie eine metrische Variable behandeln lässt, besteht im *Zwei-Gruppen-Fall* eine formale Übereinstimmung zwischen Diskriminanz- und Regressionsanalyse. Mit einer Gruppierungsvariablen, die für Elemente der Gruppe A den Wert 1 und für Elemente der Gruppe B den Wert 2 annimmt, erhält man die folgende *Regressionsfunktion*:

Regressionsfunktion

$$Y = 0,98 + 0,269X_1 - 0,147X_2 \qquad (R^2 = 0,477)$$

Das Bestimmtheitsmaß R^2 besagt, dass 47,7 % der Streuung der Gruppierungsvariablen durch die Regressionsfunktion erklärt werden (vgl. Kapitel 1).

Multipliziert man die Regressionskoeffizienten mit dem Faktor 3,83, so erhält man die in (4.10) angegebenen Koeffizienten der optimalen Diskriminanzfunktion. Die erhaltene Regressionsfunktion ist also lediglich anders „normiert“ als die Diskriminanzfunktion.

Modelltheoretische Unterschiede

Trotz der formalen Ähnlichkeit bestehen gravierende *modelltheoretische Unterschiede* zwischen Regressionsanalyse und Diskriminanzanalyse. Die abhängige Variable des Regressionsmodells ist eine Zufallsvariable, während die unabhängigen Variablen fix sind. Im statistischen Modell der Diskriminanzanalyse, das auf R.A. Fisher zurückgeht,[13] verhält es sich genau umgekehrt, d. h. die Gruppen sind fixiert und die Merkmale variieren zufällig (stochastisch). Bei der Durchführung statistischer Tests wird unterstellt, dass die Merkmalsvariablen multivariat normalverteilt sind, mit annähernd gleicher Kovarianzstruktur in den Gruppen.

4.2.3.6 Mehrfache Diskriminanzfunktionen

Mehr-Gruppen-Fall

Im *Mehr-Gruppen-Fall*, d. h. bei mehr als zwei Gruppen, können mehr als eine Diskriminanzfunktion ermittelt werden. Bei G Gruppen lassen sich maximal $G - 1$ Diskriminanzfunktionen, die jeweils orthogonal (rechtwinklig bzw. unkorreliert) zueinander sind, bilden. Die Anzahl der Diskriminanzfunktionen kann allerdings nicht größer sein als die Anzahl J der Merkmalsvariablen, sodass die maximale Anzahl von Diskriminanzfunktionen durch $\min\{G - 1, J\}$ gegeben ist. Gewöhnlich wird man jedoch mehr Merkmalsvariablen als Gruppen haben. Ist das nicht der Fall, so sollte die Anzahl der Gruppen vermindert werden.

[12] Dieser Konvention wird auch im Programm SPSS gefolgt.
[13] Vgl. Fisher (1936), S. 179 ff.

Auch im Mehr-Gruppen-Fall werden die Diskriminanzfunktionen durch Maximierung des Diskriminanzkriteriums

$$\Gamma = \frac{SS_b}{SS_w} = \frac{\text{erklärte Streuung}}{\text{nicht erklärte Streuung}}$$

ermittelt. Der Maximalwert des Diskriminanzkriteriums

$$\gamma = \max\{\Gamma\}$$

wird als Eigenwert bezeichnet, da er sich mathematisch durch Lösung eines sog. Eigenwertproblems auffinden lässt (vgl. mathematischer Anhang A).

Eigenwert

Zu jeder Diskriminanzfunktion gehört ein Eigenwert. Für die Folge der Eigenwerte gilt

$$\gamma_1 \geq \gamma_2 \geq \gamma_3 \geq \dots$$

Eine zweite Diskriminanzfunktion wird so ermittelt, dass sie einen maximalen Anteil derjenigen Streuung erklärt, die nach Ermittlung der ersten Diskriminanzfunktion als Rest verbleibt. Da die erste Diskriminanzfunktion so ermittelt wurde, dass ihr Eigenwert und damit ihr Erklärungsanteil maximal wird, kann der Erklärungsanteil der zweiten Diskriminanzfunktion (bezogen auf die gesamte Streuung) nicht größer sein. Entsprechend wird jede weitere Diskriminanzfunktion so ermittelt, dass sie jeweils einen maximalen Anteil der verbleibenden Reststreuung erklärt.

Als Maß für die relative Wichtigkeit einer Diskriminanzfunktion wird der *Eigenwertanteil* (erklärter Varianzanteil)

Eigenwertanteil

$$EA_k = \frac{\gamma k}{\gamma_1 + \gamma_2 + \dots + \gamma_K} \tag{4.13}$$

verwendet. Er gibt die durch die k-te Diskriminanzfunktion erklärte Streuung als Anteil der Streuung an, die insgesamt durch die Menge der K möglichen Diskriminanzfunktionen erklärt wird. Die Eigenwertanteile summieren sich zu Eins, während die Eigenwerte selbst auch größer als Eins sein können. Auf die statistische Signifikanzprüfung von Diskriminanzfunktionen wird im folgenden Abschnitt eingegangen.

Die Wichtigkeit (diskriminatorische Bedeutung) der sukzessiv ermittelten Diskriminanzfunktionen nimmt in der Regel sehr schnell ab. Empirische Erfahrungen zeigen, dass man auch bei großer Anzahl von Gruppen und Merkmalsvariablen meist mit zwei Diskriminanzfunktionen auskommt.[14] Dies hat unter anderem den Vorteil, dass sich die Ergebnisse leichter interpretieren und auch graphisch darstellen lassen.

Vorteil

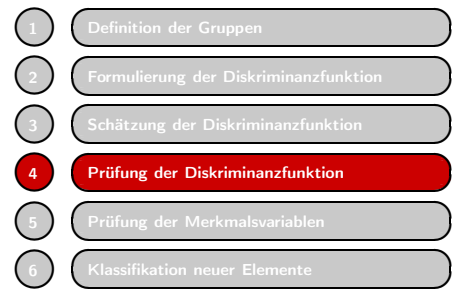
Bei zwei Diskriminanzfunktionen lässt sich (analog der Diskriminanzachse bei einer Diskriminanzfunktion) eine *Diskriminanzebene* bilden. Die Elemente der Gruppen, die geometrisch gesehen Punkte im J-dimensionalen Raum der Merkmalsvariablen bilden, lassen sich in der Diskriminanzebene graphisch darstellen. Desgleichen lassen sich auch die Merkmalsvariablen in der Diskriminanzebene als Vektoren darstellen. Die Diskriminanzanalyse kann somit auch, alternativ zur Faktorenanalyse oder zur Multidimensionalen Skalierung, für Positionierungsanalysen Verwendung finden.[15]

Diskriminanzebene

[14] Vgl. Cooley/Lohnes (1971), S. 244.

[15] Auf der Diskriminanzanalyse basiert z.B. das Programm „Adaptive Perceptual Mapping“ (APM), das von der amerikanischen Firma Sawtooth Software (Ketchum, ID) kommerziell angeboten wird. Vgl. dazu Johnson (1987), S. 143-158.

4.2.4 Prüfung der Diskriminanzfunktion



Güte der Diskriminanzfunktion

Die Güte (Trennkraft) einer Diskriminanzfunktion lässt sich durch die Unterschiedlichkeit der Gruppen, wie sie sich in den Diskriminanzwerten widerspiegelt, messen. Zwecks Prüfung der Diskriminanzfunktion kann man daher auf das oben abgeleitete Diskriminanzkriterium zurückgreifen.

Eine zweite Möglichkeit zur Prüfung der Diskriminanzfunktion besteht darin, die durch die Diskriminanzfunktion bewirkte Klassifizierung der Untersuchungsobjekte mit deren tatsächlicher Gruppenzugehörigkeit zu vergleichen. Beide Möglichkeiten sind inhaltlich eng miteinander verknüpft und müssen somit zu ähnlichen Ergebnissen führen. Die zweite Möglichkeit soll hier zunächst behandelt werden.

4.2.4.1 Prüfung der Klassifikation

In Abbildung 4.11 wurden die Diskriminanzwerte aller 24 Käufer sowie die Mittelwerte und Standardabweichungen in den beiden Gruppen zusammengestellt. Die Mittelwerte kennzeichnen die Lage der Gruppenmittel (Centroide) auf der Diskriminanzachse (vgl. Abbildung 4.14). Für das Gesamtmittel und damit für den kritischen Diskriminanzwert ergibt sich gemäß der durchgeführten Normierung der Wert Null.

Trefferquote

Die korrekt klassifizierten Elemente der Gruppe A müssen negative und die der Gruppe B positive Diskriminanzwerte haben. Aus Abbildung 4.11, wie auch aus Abbildung 4.14, ist ersichtlich, dass ein Element von Gruppe A und zwei Elemente von Gruppe B falsch zugeordnet werden. Insgesamt werden somit 21 von 24 Beurteilungen korrekt klassifiziert und die „*Trefferquote*“ beträgt 87,5 %.

Klassifikationsmatrix

Die Häufigkeiten der korrekt und falsch klassifizierten Elemente für die verschiedenen Gruppen lassen sich übersichtlich in einer sog. *Klassifikationsmatrix* (auch Confusion-Matrix genannt) zusammenfassen. Abbildung 4.17 zeigt die Klassifikationsmatrix für das Beispiel. In der Hauptdiagonale stehen die Fallzahlen der korrekt klassifizierten Elemente jeder Gruppe und in den übrigen Feldern die der falsch klassifizierten Elemente. In Klammern sind jeweils die relativen Häufigkeiten angegeben. Die Klassifikationsmatrix lässt sich analog auch für mehr als zwei Gruppen erstellen.

Tatsächliche	Prognostizierte Gruppenzugehörigkeit	
Gruppenzugehörigkeit	Marke A	Marke B
Marke A	11 (91,7 %)	1 (8,3 %)
Marke B	2 (16,7 %)	10 (83,3 %)

Abbildung 4.17: Klassifikationsmatrix

Um die Klassifikationsfähigkeit einer Diskriminanzfunktion richtig beurteilen zu können, muss man deren Trefferquote mit derjenigen Trefferquote vergleichen, die man bei einer rein *zufälligen Zuordnung* der Elemente, z. B. durch Werfen einer Münze oder durch Würfeln, erreichen würde. Im vorliegenden Fall bei zwei Gruppen mit gleicher Größe wäre bei zufälliger Zuordnung bereits eine Trefferquote von 50 % zu erwarten.

Die Trefferquote, die man durch zufällige Zuordnung erreichen kann, liegt noch höher bei ungleicher Größe der Gruppen. So kann bei einem Größenverhältnis von 80 zu 20 und rein zufälliger Zuordnung (mit den Wahrscheinlichkeiten 0,8 und 0,2) eine Trefferquote von 80 % erwartet werden.[16] Eine Diskriminanzfunktion kann nur dann von Nutzen sein, wenn sie eine höhere Trefferquote erzielt, als nach dem Zufallsprinzip zu erwarten ist.

Weiterhin ist zu berücksichtigen, dass die Trefferquote immer überhöht ist, wenn sie, wie allgemein üblich, auf Basis derselben Stichprobe berechnet wird, die auch für die Schätzung der Diskriminanzfunktion verwendet wurde. Da die Diskriminanzfunktion immer so ermittelt wird, dass die Trefferquote in der verwendeten Stichprobe maximal wird, ist bei Anwendung auf eine andere Stichprobe mit einer niedrigeren Trefferquote zu rechnen. Dieser *Stichprobeneffekt* vermindert sich allerdings mit zunehmendem Umfang der Stichprobe.

Stichprobeneffekt

Eine *bereinigte Trefferquote* lässt sich gewinnen, indem man die verfügbare Stichprobe zufällig in zwei Unterstichproben aufteilt, eine Lernstichprobe und eine Kontrollstichprobe. Die Lernstichprobe wird zur Schätzung der Diskriminanzfunktion verwendet. Mit Hilfe dieser Diskriminanzfunktion werden sodann die Elemente der Kontrollstichprobe klassifiziert und hierfür die Trefferquote berechnet. Diese Vorgehensweise ist allerdings nur dann zweckmäßig, wenn eine hinreichend große Stichprobe zur Verfügung steht, da mit abnehmender Größe der Lernstichprobe die Zuverlässigkeit der geschätzten Diskriminanzkoeffizienten abnimmt. Außerdem wird die vorhandene Information nur unvollständig genutzt.[17]

bereinigte Trefferqoute

4.2.4.2 Prüfung des Diskriminanzkriteriums

Der Eigenwert (Maximalwert des Diskriminanzkriteriums)

$$\gamma = \frac{SS_b}{SS_w} = \frac{\text{erklärte Streuung}}{\text{nicht erklärte Streuung}}$$

bildet ein Maß für die Güte (Trennkraft) der Diskriminanzfunktion. Er besitzt jedoch den Nachteil, dass er nicht auf Werte zwischen Null und Eins normiert ist. Da SS_b und SS_w beliebige positive Werte annehmen können, kann der Eigenwert auch größer als Eins sein.

Nachteil

[16] Die gleiche Trefferquote wird erzielt, wenn man nach der simplen Regel verfährt: Ordne alle Elemente der größten Gruppe zu.

[17] Bessere Möglichkeiten zur Erzielung von unverzerrten Trefferquoten bieten Methoden der Kreuz-Validierung (Cross-Validation). Einen einfachen Spezialfall bildet die Leave-one-out-Methode (LOO). Ein jedes Element wird klassifiziert, indem man es zuvor aus der Stichprobe vom Umfang N herausnimmt und dann die Diskriminanzfunktion auf Basis der übrigen $N - 1$ Elemente schätzt. Auf diese Art lässt sich unter vollständiger Nutzung der vorhandenen Information eine verbesserte Klassifizierungstabelle und Schätzung der Trefferquote erzielen. Vgl. dazu auch die Ausführungen im folgenden Kapitel in Abschnitt 4.2.4.2. Für die Diskriminanzanalyse bietet SPSS keine entsprechende Option an.

Im Gegensatz dazu sind die folgenden Quotienten auf Werte von Null bis Eins normiert:

$$\frac{\gamma}{1+\gamma} = \frac{SS_b}{SS_b + SS_w} = \frac{\text{erklärte Streuung}}{\text{Gesamtstreuung}} \tag{4.14}$$

$$\frac{1}{1+\gamma} = \frac{SS_w}{SS_b + SS_w} = \frac{\text{nicht erklärte Streuung}}{\text{Gesamtstreuung}} \tag{4.15}$$

Im Zwei-Gruppen-Fall, in dem sich, wie oben dargelegt, auch die Regressionsanalyse anwenden lässt, entspricht (4.14) dem Bestimmtheitsmaß $R^2 = 0,477$, das als Gütemaß bei der Regressionsanalyse üblich ist. In der Diskriminanzanalyse wird üblicherweise die Wurzel von (4.14) als Gütemaß verwendet. Sie wird als kanonischer Korrelationskoeffizient bezeichnet.[18]

Kanonischer Korrelationskoeffizient

Kanonischer Korrelationskoeffizient

$$c = \sqrt{\frac{\gamma}{1+\gamma}} = \sqrt{\frac{\text{erklärte Streuung}}{\text{Gesamtstreuung}}} \tag{4.16}$$

Im Zwei-Gruppen-Fall ist die kanonische Korrelation identisch mit der (einfachen) Korrelation zwischen den geschätzten Diskriminanzwerten und der Gruppierungsvariable. Im Beispiel erhält man für den kanonischen Korrelationskoeffizienten den Wert

$$c = \sqrt{\frac{0,912}{1+0,912}} = 0,691$$

Wilks' Lambda

Das gebräuchlichste Kriterium zur Prüfung der Diskriminanz bildet Wilks' Lambda (auch als U-Statistik bezeichnet). Es entspricht dem Ausdruck in (4.15).

Wilks' Lambda

$$\Lambda = \frac{1}{1+\gamma} = \frac{\text{nicht erklärte Streuung}}{\text{Gesamtstreuung}} \tag{4.17}$$

Wilks' Lambda ist ein „inverses" Gütemaß, d. h. kleinere Werte bedeuten höhere Trennkraft der Diskriminanzfunktion und umgekehrt.

Im Beispiel erhält man für Wilks' Lambda den Wert

$$\Lambda = \frac{1}{1+0,912} = 0,523$$

[18] Der Begriff stammt aus der kanononischen Korrelationsanalyse. Mit diesen Verfahren lässt sich die Beziehung zwischen zwei Mengen von jeweils metrisch skalierten Variablen untersuchen. Fasst man jede Menge mittels einer Linearkombination zu einer kanonischen Variablen zusammen, so ist der kanonische Korrelationskoeffizient der einfache Korrelationskoeffizient (nach Bravais/Pearson) zwischen den beiden kanonischen Variablen. Die Linearkombinationen werden bei der kanonischen Analyse so ermittelt, dass der kanonische Korrelationskoeffizient maximal wird.
Die Diskriminanzanalyse lässt sich als Spezialfall einer kanonischen Analyse interpretieren. Jede nominal skalierte Variable mit G Stufen, und somit auch die Gruppierungsvariable einer Diskriminanzanalyse, lässt sich äquivalent durch $G-1$ binäre Variablen ersetzen. Die Diskriminanzanalyse bildet somit eine kanonische Analyse zwischen einer Menge von binären Variablen und einer Menge metrisch skalierter Merkmalsvariablen. Vgl. hierzu Tatsuoka (1988), S. 235 ff.

Zwischen c und Λ besteht die folgende Beziehung

$$c^2 + \Lambda = 1$$

Die Bedeutung von Wilks' Lambda liegt darin, dass es sich in eine probabilistische Variable transformieren lässt und damit Wahrscheinlichkeitsaussagen über die Unterschiedlichkeit von Gruppen erlaubt. Dadurch wird eine statistische *Signifikanzprüfung der Diskriminanzfunktion* möglich. Die Transformation

Signifikanzprüfung

$$\chi^2_{emp} = -\left[N - \frac{J+G}{2} - 1\right] \ln \Lambda \tag{4.18}$$

mit

N : Anzahl der Fälle
J : Anzahl der Variablen
G : Anzahl der Gruppen
Λ : Wilks' Lambda
ln : natürlicher Logarithmus

liefert eine Variable, die angenähert wie χ^2 (Chi-quadrat) verteilt ist mit $J \cdot (G-1)$ Freiheitsgraden (degrees of freedom).[19] Der χ^2-Wert wird mit kleinerem Λ größer. Höhere Werte bedeuten daher auch größere Unterschiedlichkeit der Gruppen. Für das Beispiel erhält man:

$$\chi^2_{emp} = -\left[24 - \frac{2+2}{2} - 1\right] \ln 0,523 = 13,6$$

Für 2 Freiheitsgrade lässt sich damit aus der χ^2-Tabelle im Anhang A.4 dieses Buches ein p-Wert (empirisches Signifikanzniveau) von annähernd 0,001 entnehmen.[20] Die ermittelte Diskriminanzfunktion ist also hoch signifikant.

Die Signifikanzprüfung beinhaltet einen Test der Nullhypothese H_0 gegen die Alternativhypothese H_1:

H_0 : Die beiden Gruppen unterscheiden sich nicht.
H_1 : Die beiden Gruppen unterscheiden sich.

Angewendet auf das Beispiel besagt die Nullhypothese, dass die beiden Gruppen von Stammkäufern sich hinsichtlich ihrer Einstellungen nicht unterscheiden. Unter dieser Hypothese ist hier für χ^2_{emp} der Wert 2 (= Zahl der Freiheitsgrade) zu erwarten. Tatsächlich aber ergibt sich χ^2_{emp} = 13,6. Die Wahrscheinlichkeit, dass sich bei Richtigkeit von H_0 (und somit also rein zufallsbedingt) ein so großer oder größerer Wert für χ^2_{emp} ergibt, beträgt nur 0,1 %. Damit ist es höchst unwahrscheinlich, dass H_0 richtig ist. H_0 ist folglich abzulehnen und damit H_1 anzunehmen. Mit der Irrtumswahrscheinlichkeit (Signifikanzniveau) von 0,1 % lässt sich also sagen, dass die beiden Gruppen sich unterscheiden.

Nullhypothese

[19] Die χ^2-Verteilung ergibt sich als Verteilung der Summe von quadrierten unabhängigen normalverteilten Variablen. Sie konvergiert, allerdings recht langsam, mit wachsender Zahl von Freiheitsgraden gegen die Normalverteilung.

[20] In Excel lässt sich der p-Wert von Chi-quadrat durch die Funktion CHIVERT(x; df) mit $x = \chi^2_{emp}$ berechnen. Es gilt damit hier p = CHIVERT(13,614; 2) = 0,0011.
Für die Ablehnung der Nullhypothese mit Signifikanzniveau $\alpha = 0,001$ müsste genau genommen $p < \alpha = 0,001$ gelten.

In Abbildung 4.18 sind die Werte verschiedener Gütemaße zusammengestellt. Im *Mehr-Gruppen-Fall*, wenn sich K Diskriminanzfunktionen bilden lassen, können diese einzeln mit Hilfe der obigen Maße beurteilt und miteinander verglichen werden. Um die *Unterschiedlichkeit der Gruppen* zu prüfen, müssen dagegen alle Diskriminanzfunktionen bzw. deren Eigenwerte gemeinsam berücksichtigt werden. Ein geeignetes Maß hierfür ist das multivariate Wilks' Lambda. Man erhält es durch Multiplikation der univariaten Lambdas.

Multivariates Wilks' Lambda

Variable	Diskriminanz (Eigenwert γ)	Wilks' Lambda Λ	Chi-quadrat χ^2	Signifikanz (p-Wert)
Y	0,912	0,523	13,6	0,001

Abbildung 4.18: Gütemaße der Diskriminanzfunktion

Multivariates Wilks' Lambda

$$\Lambda = \prod_{k=1}^{K} \frac{1}{1+\gamma_k} \tag{4.19}$$

mit

γ_k = Eigenwert der k-ten Diskriminanzfunktion

Zwecks Signifikanzprüfung der Unterschiedlichkeit der Gruppen bzw. der Gesamtheit der Diskriminanzfunktionen kann wiederum mittels der Transformation (4.18) eine χ^2-Variable gebildet werden.

Um zu entscheiden, ob nach Ermittlung der ersten k Diskriminanzfunktionen die restlichen $K-k$ Diskriminanzfunktionen noch signifikant zur Unterscheidung der Gruppen beitragen können, ist es von Nutzen, Wilks' Lambda in folgender Form zu berechnen:

Wilks' Lambda für residuelle Diskriminanz

(nach Ermittlung von k Diskriminanzfunktionen)

$$\Lambda_k = \prod_{q=k+1}^{K} \frac{1}{1+\gamma_q} \qquad (k = 0, 1, \ldots, K-1) \tag{4.20}$$

mit

γ_q = Eigenwert der q-ten Diskriminanzfunktion

Die zugehörige χ^2-Variable, die man durch Einsetzen von Λ_k in (4.17) erhält, besitzt $(J-k)\cdot(G-k-1)$ Freiheitsgrade. Für $k=0$ ist Formel (4.20) identisch mit (4.19).

Residuelle Diskriminanz

Wird die residuelle Diskriminanz insignifikant, so kann man die Ermittlung weiterer Diskriminanzfunktionen abbrechen, da diese nicht signifikant zur Trennung der Gruppen beitragen können. Diese Vorgehensweise bietet allerdings keine Gewähr dafür, dass die bereits ermittelten k Diskriminanzfunktionen alle signifikant sind (ausgenommen bei $k=1$), sondern stellt lediglich sicher, dass diese *in ihrer Gesamtheit*

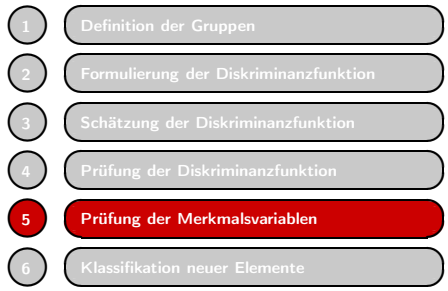
signifikant trennen. Ist die residuelle Diskriminanz bereits für $k = 0$ insignifikant, so bedeutet dies, dass die Nullhypothese nicht widerlegt werden kann. Es besteht dann kein empirischer Befund für einen systematischen Unterschied zwischen den Gruppen. Die Bildung von Diskriminanzfunktionen erscheint somit nutzlos.

Die statistische Signifikanz einer Diskriminanzfunktion besagt andererseits noch nicht, dass diese auch wirklich gut trennt, sondern lediglich, dass sich die Gruppen bezüglich dieser Diskriminanzfunktion signifikant unterscheiden. Wie bei allen statistischen Tests gilt auch hier, dass ein signifikanter Unterschied nicht auch „relevant" sein muss. Wenn nur der Stichprobenumfang hinreichend groß ist, so wird auch ein sehr kleiner Unterschied statistisch signifikant. Es sind daher auch die Unterschiede der Mittelwerte (vgl. Abbildung 4.7) sowie die Größe des kanonischen Korrelationskoeffizienten oder von Wilks' Lambda zu beachten.

Interpretierbarkeit

Aus Gründen der Interpretierbarkeit und graphischen Darstellbarkeit kann es bei einer Mehrzahl von signifikanten Diskriminanzfunktionen sinnvoll sein, nicht alle signifikanten Diskriminanzfunktionen zu berücksichtigen, sondern sich mit nur zwei oder maximal drei Diskriminanzfunktionen zu begnügen.

4.2.5 Prüfung der Merkmalsvariablen



Merkmalswichtigkeit

Es ist aus zweierlei Gründen von Interesse, die Wichtigkeit der Merkmalsvariablen in der Diskriminanzfunktion beurteilen zu können. Zum einen, um die Unterschiedlichkeit der Gruppen zu *erklären*, und zum anderen, um unwichtige Variablen aus der Diskriminanzfunktion zu *entfernen*.

Die diskriminatorische Bedeutung der Merkmalsvariablen hatten wir bereits isoliert (univariat) betrachtet. Sie zeigt sich in der Unterschiedlichkeit ihrer Mittelwerte zwischen den Gruppen (vgl. Abbildung 4.7) oder besser noch am Wert des Diskriminanzkriteriums bei Anwendung auf die Merkmalsvariablen (vgl. Abbildung 4.12, Zeile 1 und 2). Diese Werte lassen sich vor Ermittlung der Diskriminanzfunktion berechnen.

Ebenfalls lässt sich auch mit Hilfe von Wilks' Lambda vor Durchführung einer Diskriminanzanalyse für jede Merkmalsvariable isoliert deren Trennfähigkeit überprüfen. Die Berechnung erfolgt in diesem Fall durch Streuungszerlegung gemäß Formel (4.8). Zur Signifikanzprüfung kann der allgemein übliche F-Test anstelle des χ^2-Tests verwendet werden. Es gilt:

Streuungszerlegung

$$F_{emp} = \frac{SS_b/(G-1)}{SS_w/(N-G)} = \Lambda \frac{N-G}{G-1} = \Lambda \frac{22}{1} \tag{4.21}$$

Das Ergebnis entspricht dem einer einfachen Varianzanalyse zwischen Gruppierungs- und Merkmalsvariable (vgl. Kapitel 3). Die Ergebnisse sind in Abbildung 4.19 wiedergegeben.[21]

[21] In Excel erhält man den p-Wert für $x = F_{emp}$ durch die Funktion FVERT(x;df1;df2). Es gilt damit hier FVERT(10,24;1;22) = 0,004 und FVERT(0,673;1;22) = 0,421.

Variable	Diskriminanz	Wilks' Lambda	F-Wert	Signifikanz
X_1	0,466	0,682	10,24	0,004
X_2	0,031	0,970	0,67	0,421

Abbildung 4.19: Univariate Diskriminanzprüfung der Merkmalsvariablen

Für die Diskriminanz von Variable 1 (Streichfähigkeit) ergibt sich ein empirisches Signifikanzniveau von 0,4 %, für Variable 2 (Haltbarkeit) dagegen von 42,1 %. Schon Abbildung 4.6 hatte erkennen lassen, dass die Streichfähigkeit eine bessere diskriminatorische Eignung besitzt.

Interdependenz

Infolge möglicher Interdependenz zwischen den Merkmalsvariablen ist eine univariate Prüfung der Diskriminanz nicht ausreichend. Obgleich Variable 2 allein nur eine minimale Diskriminanz besitzt, trägt sie doch in Kombination mit Variable 1 erheblich zur Erhöhung der Diskriminanz bei, wie sich durch einen Vergleich der Abbildung 4.18 und Abbildung 4.19 erkennen lässt (der Diskriminanzwert in Abbildung 4.18 ist bedeutend größer als die Summe der beiden Diskriminanzwerte in Abbildung 4.19).

Die Basis für die *multivariate Beurteilung der diskriminatorischen* Bedeutung einer Merkmalsvariablen, also ihre Bedeutung im Rahmen der Diskriminanzfunktion, bilden die Diskriminanzkoeffizienten. Diese repräsentieren den Einfluss einer Merkmalsvariablen auf die Diskriminanzvariable. Im Beispiel ergab sich:

$$b_1 = 1,031$$
$$b_2 = -0,565$$

Modifikation

Diese Werte sind allerdings noch zu modifizieren, da die Größe eines Diskriminanzkoeffizienten auch von eventuell willkürlichen Skalierungseffekten beeinflusst wird. Hat man z. B. eine Merkmalsvariable „Preis“ und ändert deren Maßeinheit von [€] auf [Cent], so würde sich der zugehörige Diskriminanzkoeffizient um den Faktor 100 vergrößern. Auf die diskriminatorische Bedeutung hat die Skalentransformation keinen Einfluss.

Um derartige Effekte auszuschalten, muss man die Diskriminanzkoeffizienten *standardisieren*, indem man sie mit der Standardabweichung der betreffenden Merkmalsvariablen multipliziert.[22]

Standardisierter Diskriminanzkoeffizient

$$b_j^* = b_j \cdot s_j \tag{4.22}$$

mit

b_j = Diskriminanzkoeffizient von Merkmalsvariable j
s_j = Standardabweichung von Merkmalsvariable j

[22] Die standardisierten Diskriminanzkoeffizienten entsprechen den Beta-Werten der Regressionsanalyse. Sie stimmen dann mit den normierten Diskriminanzkoeffizienten überein, wenn die Merkmalsvariablen vor Durchführung der Diskriminanzanalyse so standardisiert werden, dass ihre Mittelwerte Null und ihre gepoolten Innergruppen-Standardabweichungen Eins ergeben.

Standardisierung

Zweckmäßigerweise wird für die Standardisierung auf die Innergruppen-Streuung zurückgegriffen. Für die *Innergruppen-Varianz* der Merkmalsvariablen (pooled within-groups variance) gilt analog zu (4.12):

$$s_j = \sqrt{\frac{W_{jj}}{I - G}}$$

Mit den Werten aus Abbildung 4.8 ergibt sich:

$$s_1 = \sqrt{\frac{29}{24 - 2}} = 1,148$$

$$s_2 = \sqrt{\frac{49}{24 - 2}} = 1,492$$

Man erhält damit die standardisierten Diskriminanzkoeffizienten:

$$b_1^* = b_1 \cdot s_1 = \ \ 1,031 \cdot 1,148 = \ \ 1,184$$
$$b_2^* = b_2 \cdot s_2 = -0,565 \cdot 1,492 = -0,843$$

Für die Beurteilung der diskriminatorischen Bedeutung spielt das Vorzeichen der Koeffizienten keine Rolle. Wie schon zuvor gesehen, besitzt Variable 2 (Haltbarkeit) eine geringere Bedeutung als Variable 1 (Streichfähigkeit). Die Bedeutung von Variable 2 ist aber weit größer, als eine isolierte Betrachtung erkennen lässt.

Zur Unterscheidung von den standardisierten Diskriminanzkoeffizienten werden die (normierten) Koeffizienten in der Diskriminanzfunktion auch als *nicht-standardisierte Diskriminanzkoeffizienten* bezeichnet. Zur Berechnung von Diskriminanzwerten müssen immer die nicht-standardisierten Diskriminanzkoeffizienten verwendet werden.

Im Falle von mehrfachen Diskriminanzfunktionen existieren für jede Merkmalsvariable mehrere Diskriminanzkoeffizienten. Um die diskriminatorische Bedeutung einer Merkmalsvariablen bezüglich aller Diskriminanzfunktionen zu beurteilen, sind die mit den Eigenwertanteilen gemäß Formel (4.13) gewichteten absoluten Werte der Koeffizienten einer Merkmalsvariablen zu addieren. Man erhält auf diese Weise die *mittleren Diskriminanzkoeffizienten*:

Mittlerer Diskriminanzkoeffizient

$$\bar{b}_j = \sum_{k=1}^{K} \left| b_{jk}^* \right| \cdot EA_k \tag{4.23}$$

mit

b_{jk}^* = Standardisierter Diskriminanzkoeffizient für Merkmalsvariable j bezüglich Diskriminanzfunktion k

EA_k = Eigenwertanteil der Diskriminanzfunktion k

Clusteranalyse

Bei Vorschaltung einer Clusteranalyse für die Gruppenbildung können bei der nachfolgenden Diskriminanzanalyse dieselben oder andere Variablen verwendet werden. Im ersten Fall will man die Eignung der Variablen für die Clusterbildung überprüfen, im zweiten Fall die durch die Clusteranalyse erzeugte Gruppierung erklären. Dabei bezeichnet man die für die Clusteranalyse verwendeten Variablen auch als „*aktive*“ und die für die Diskriminanzanalyse verwendeten Variablen als „*passive*“ Variablen. Beispiel: Gruppierung (Segmentierung) von Personen nach ihrem Kaufverhalten (aktiv) durch Clusteranalyse und Erklärung der Unterschiede im Kaufverhalten durch psychographische Variablen (passiv) mittels Diskriminanzanalyse.

4.2.6 Klassifizierung neuer Elemente

Klassifikations-konzepte

Für die Klassifizierung von neuen Elementen lassen sich das *Distanzkonzept*, das *Wahrscheinlichkeitskonzept* und die *Klassifizierungsfunktionen* unterscheiden.

Distanzkonzept

Das Distanzkonzept wurde oben bereits angesprochen. Danach wird ein Element i in diejenige Gruppe g eingeordnet, der es auf der Diskriminanzachse am nächsten liegt, d. h. bezüglich derer die Distanz zwischen Element und Gruppenmittel (Centroid) minimal wird. Dies ist äquivalent damit, ob das Element links oder rechts vom kritischen Diskriminanzwert liegt. Bei mehreren Diskriminanzfunktionen aber wird die Anwendung etwas schwieriger.

Wahrscheinlichkeits-konzept

Auf dem Distanzkonzept basiert auch das Wahrscheinlichkeitskonzept, welches die Behandlung der Klassifizierung als ein statistisches Entscheidungsproblem ermöglicht. Es besitzt daher unter allen Konzepten die größte Flexibilität, ist aber, besonders für einen Nicht-Statistiker, auch schwerer verständlich. Wir behandeln es daher an letzter Stelle.

4.2.6.1 Klassifizierungsfunktionen

Die von R.A. Fisher entwickelten Klassifizierungsfunktionen bilden ein bequemes Hilfsmittel, um die Klassifizierung direkt auf Basis der Merkmalswerte (ohne Verwendung von Diskriminanzfunktionen) durchzuführen. Die Klassifizierungsfunktionen sind allerdings nur dann anwendbar, wenn gleiche Streuung in den Gruppen unterstellt werden kann, d. h. wenn die Kovarianzmatrizen der Gruppen annähernd gleich sind (s.u.). Da die Klassifizierungsfunktionen auch als (lineare) Diskriminanzfunktionen bezeichnet werden, können sich leicht Verwechslungen mit den (kanonischen) Diskriminanzfunktionen (vgl. Abschnitt 4.2.2) ergeben.

Klassifizierungs-funktionen

Für jede Gruppe g ist eine gesonderte Klassifizierungsfunktion zu bestimmen. Man erhält damit G Funktionen folgender Form:

Fisher's Klassifizierungsfunktionen

$$\begin{aligned} F_1 &= b_{01} + b_{11}X_1 + b_{21}X_2 + \ldots + b_{J1}X_J \\ F_2 &= b_{02} + b_{12}X_1 + b_{22}X_2 + \ldots + b_{J2}X_J \\ &\vdots \\ F_G &= b_{0G} + b_{1G}X_1 + b_{2G}X_2 + \ldots + b_{JG}X_J \end{aligned} \tag{4.24}$$

Zur Durchführung der Klassifizierung eines Elementes ist mit dessen Merkmalswerten für jede Gruppe g ein Funktionswert F_g zu berechnen. Das Element ist derjenigen Gruppe g zuzuordnen, für die der Funktionswert F_g maximal ist. Die Funktionswerte selbst haben keinen interpretatorischen Gehalt.

Für das *Beispiel* erhält man die folgenden zwei Klassifizierungsfunktionen (vgl. mathematischer Anhang D):

$$F_A = -6,597 + 1,729 X_1 + 1,280 X_2$$
$$F_B = -10,22 + 3,614 X_1 + 0,247 X_2$$

Für die Merkmalswerte

$$X_1 = 6 \quad \text{und} \quad X_2 = 7$$

erhält man durch Einsetzen in die Klassifizierungsfunktionen die Funktionswerte:

$$F_A = 12,7$$
$$F_B = 13,2$$

Das Element ist also in die Gruppe B einzuordnen. Aus dieser Gruppe stammt auch Person 21 (vgl. Abbildung 4.5), die identische Merkmalswerte besitzt.

A-priori-Wahrscheinlichkeiten

Die Klassifizierungsfunktionen ermöglichen auch die Einbeziehung von *A-priori-Wahrscheinlichkeiten*. Damit sind Wahrscheinlichkeiten gemeint, die a priori, d. h. vor Durchführung einer Diskriminanzanalyse hinsichtlich der Gruppenzugehörigkeit, gegeben sind oder geschätzt werden können.

Mittels der A-priori-Wahrscheinlichkeiten lässt sich gegebenenfalls berücksichtigen, dass die betrachteten Gruppen mit unterschiedlicher Häufigkeit in der Realität vorkommen. A priori ist z. B. von einer Person eher zu erwarten, dass sie Wähler einer großen Partei oder Käufer einer Marke mit großem Marktanteil ist, als Wähler einer kleinen Partei oder Käufer einer kleinen Marke. Entsprechend den relativen Größen der Gruppen, soweit diese bekannt sind, können daher A-priori-Wahrscheinlichkeiten gebildet werden. Der Untersucher kann aber auch durch subjektive Schätzung der A-priori-Wahrscheinlichkeiten seine persönliche Meinung, die er unabhängig von den in die Diskriminanzfunktion eingehenden Informationen gebildet hat, in die Rechnung einbringen.

Die A-priori-Wahrscheinlichkeiten müssen sich über die Gruppen zu Eins addieren:

$$\sum_{g=1}^{G} P(g) = 1$$

Zur Berücksichtigung der A-priori-Wahrscheinlichkeit $P(g)$ sind die Klassifizierungsfunktionen wie folgt zu modifizieren:

$$F_g := F_g + \ln P(g) \qquad (g = 1, \ldots, G) \tag{4.25}$$

Bei der Durchführung einer Klassifizierung lassen sich auch individuelle A-priori-Wahrscheinlichkeiten $P_i(g)$ berücksichtigen. Werden nur gruppenspezifische A-priori-Wahrscheinlichkeiten $P(g)$ berücksichtigt, so lassen sich diese in die Berechnung des konstanten Gliedes b_{0g} einer Funktion F_g einbeziehen:

$$b_{0g} = a_g + \ln P(g)$$

Sind keine A-priori-Wahrscheinlichkeiten bekannt, so kann immer $P(g) = 1/G$ gesetzt werden. So wurden auch die konstanten Glieder in den obigen Funktionen wie folgt berechnet (vgl. auch mathematischer Anhang D):

$$b_{0A} = -5,904 + \ln 0,5 = -6,597$$
$$b_{0B} = -9,529 + \ln 0,5 = -10,222$$

In dieser Form erhält man auch bei Anwendung von SPSS die Klassifizierungsfunktionen, wenn keine A-priori-Wahrscheinlichkeiten angegeben werden. Wenn die A-priori-Wahrscheinlichkeiten gleich sind, dann hat ihre Einbeziehung natürlich keinen Effekt auf das Ergebnis der Klassifizierung.

4.2.6.2 Das Distanzkonzept

Distanzkonzept

Gemäß dem Distanzkonzept wird ein Element i in diejenige Gruppe g eingeordnet, der es am nächsten liegt, d. h. bezüglich derer die Distanz zwischen Element und Gruppenmittel (Centroid) minimal wird. Üblicherweise werden die *quadrierten Distanzen*

$$D_{ig}^2 = \left(Y_i - \bar{Y}_g\right)^2 \qquad (g = 1, \ldots, G) \tag{4.26}$$

verwendet. Bei einer Mehrzahl von K Diskriminanzfunktionen wird analog die quadrierte euklidische Distanz im K-dimensionalen Diskriminanzraum zwischen dem Element i und dem Centroid der Gruppe g herangezogen.

Quadrierte euklidische Distanz

$$D_{ig}^2 = \sum_{k=1}^{K} \left(Y_{ki} - \bar{Y}_{kg}\right)^2 \qquad (g = 1, \ldots, G) \tag{4.27}$$

mit

Y_{ki} = Diskriminanzwert von Element i bezüglich Diskriminanzfunktion k
$\bar{Y}_{kg}$ = Centroid von Gruppe g bezüglich Diskriminanzfunktion k

Die Anwendbarkeit der euklidischen Distanz ist zulässig infolge von Orthogonalität und Normierung der Diskriminanzfunktionen. Alternativ lassen sich auch Distanzen im J-dimensionalen Raum der Merkmalsvariablen berechnen. Es müssen dabei jedoch die unterschiedlichen Maßeinheiten (Standardabweichungen) der Variablen wie auch die Korrelationen zwischen den Variablen berücksichtigt werden. Ein verallgemeinertes Distanzmaß, bei dem dies der Fall ist, ist die *Mahalanobis-Distanz*. Bei nur zwei Variablen errechnet sich die quadrierte Mahalanobis-Distanz wie folgt:

Euklidische Distanz

Mahalanobis-Distanz

$$D_{ig}^2 = \frac{\left(X_{1i} - \bar{X}_{1g}\right)^2 s_2^2 + \left(X_{2i} - \bar{X}_{2g}\right)^2 s_1^2 - 2\left(X_{1i} - \bar{X}_{1g}\right)\left(X_{2i} - \bar{X}_{2g}\right) s_{12}}{s_1^2 s_2^2 - s_{12}^2}$$

Dabei sind durch s_1^2 bzw. s_2^2 die empirischen Varianzen und durch s_{12} die empirische Kovarianz der beiden Variablen bezeichnet. Die Mahalanobis-Distanz nimmt zu, wenn die Korrelation zwischen den Variablen (und damit s_{12}) abnimmt. Da die Standardabweichungen der Diskriminanzvariablen immer Eins und deren Korrelationen Null sind, sind folglich die euklidischen Distanzen im Diskriminanzraum zugleich auch Mahalanobis-Distanzen (vgl. hierzu auch die Ausführungen im Anhang B dieses Kapitels).

Die Klassifizierung nach euklidischen Distanzen im Raum der Diskriminanzvariablen ist der Klassifizierung nach Mahalanobis-Distanzen im Raum der Merkmalsvariablen äquivalent, wenn alle K möglichen Diskriminanzfunktionen berücksichtigt werden.[23] Liegen die Diskriminanzfunktionen vor, so bedeutet es eine erhebliche Erleichterung, wenn die Distanzen im Diskriminanzraum gebildet werden.

[23] Vgl. dazu Tatsuoka (1988), S. 232 ff.

Es ist für die Durchführung der Klassifizierung nicht zwingend, alle mathematisch möglichen Diskriminanzfunktionen zu berücksichtigen. Vielmehr reicht es aus, sich auf die wichtigen oder die signifikanten Diskriminanzfunktionen zu beschränken, da sich dadurch bei nur unbedeutendem Informationsverlust die Berechnung wesentlich vereinfacht. Die Beschränkung auf die signifikanten Diskriminanzfunktionen kann überdies den Vorteil haben, dass Zufallsfehler in den Merkmalsvariablen herausgefiltert werden.

Modifizierte Distanzen

Die obigen Ausführungen unterstellen, dass die Streuungen in den Gruppen annähernd gleich sind. Wenn diese Annahme nicht aufrechterhalten werden kann, müssen modifizierte Distanzen verwendet werden, deren Berechnung im Anhang B dieses Kapitels gezeigt wird. Bei Verwendung von SPSS kann die Annahme gleicher Streuungen (Kovarianzmatrizen der Merkmalsvariablen) durch Berechnung von *Box's M* überprüft werden.[24] Mittels eines F-Tests lässt sich die Signifikanz dieser Annahme prüfen. Niedrige Signifikanzwerte deuten auf ungleiche Streuungen hin.

Box's M

Die Klassifikation auf Basis des Distanzkonzeptes führt zum gleichen Ergebnis wie die Klassifikation mit Hilfe der Klassifizierungsfunktionen, wenn alle Diskriminanzfunktionen berücksichtigt und wenn gleiche Streuungen in den Gruppen unterstellt werden.

4.2.6.3 Das Wahrscheinlichkeitskonzept

Wahrscheinlichkeitskonzept

Das Wahrscheinlichkeitskonzept, das auf dem Distanzkonzept aufbaut, ist das flexibelste Konzept zur Klassifizierung von Elementen. Insbesondere ermöglicht es, wie schon die Klassifizierungsfunktionen, die Berücksichtigung von (ungleichen) *A-priori-Wahrscheinlichkeiten* (vgl. 4.2.6.1). Zusätzlich ermöglicht es auch die Berücksichtigung von (ungleichen) *„Kosten" der Fehlklassifikation*. Ohne diese Erweiterungen führt es zu den gleichen Ergebnissen, wie das Distanzkonzept. Abbildung 4.20 gibt einen Überblick über Möglichkeiten der drei Konzepte zur Klassifizierung.

	Klassifiz.-Funktionen	Distanz-Konzept	Wahrsch.-Konzept
Unterschiedliche A-priori-Wahrscheinlichkeiten	ja	nein	ja
Unterschiedliche Kosten der Fehlklassifikation	nein	nein	ja
Berücksichtigung ungleicher Streuungen in den Gruppen	nein	ja	ja
Unterdrückung irrelevanter Diskriminanzfunktionen	nein	ja	ja

Abbildung 4.20: Vergleich der Konzepte zur Klassifizierung

[24] Vgl. dazu Cooley/Lohnes (1971), S. 229.

Klassifizierungsregel

Im Wahrscheinlichkeitskonzept kommt die nachfolgende *Klassifizierungsregel* zur Anwendung:

> Ordne ein Element i derjenigen Gruppe g zu, für die
> die Wahrscheinlichkeit $P(g|Y_i)$ maximal ist.

Dabei bezeichnet $P(g|Y_i)$ die Wahrscheinlichkeit für die Zugehörigkeit von Element i mit Diskriminanzwert Y_i zu Gruppe g $(g = 1, ..., G)$.

In der Terminologie der statistischen Entscheidungstheorie werden die Klassifizierungswahrscheinlichkeiten als *A-posteriori-Wahrscheinlichkeiten* bezeichnet. Zu ihrer Berechnung wird das Bayes-Theorem angewendet.

Bayes-Theorem

$$P(g|Y_i) = \frac{P(Y_i|g)P_i(g)}{\sum_{g=1}^{G} P(Y_i|g)P_i(g)} \qquad (g = 1, \ldots, G) \tag{4.28}$$

mit

$$\begin{aligned} P(g|Y_i) &= \text{A-posteriori-Wahrscheinlichkeit} \\ P(Y_i|g) &= \text{Bedingte Wahrscheinlichkeit} \\ P_i(g) &= \text{A-priori-Wahrscheinlichkeit} \end{aligned}$$

Bayes-Theorem

Im Bayes-Theorem werden die a priori gegebenen Wahrscheinlichkeiten mit bedingten Wahrscheinlichkeiten, in denen die in den Merkmalsvariablen enthaltene Information zum Ausdruck kommt, verknüpft. Die bedingte Wahrscheinlichkeit $P(Y_i|g)$ gibt an, wie wahrscheinlich ein Diskriminanzwert Y_i für das Element i wäre, wenn dieses zu Gruppe g gehören würde. Sie lässt sich durch Transformation der Distanz D ermitteln.

Kosten der Fehlklassifikation

Bei Durchführung einer Klassifizierung im Rahmen von konkreten Problemstellungen (Entscheidungsproblemen) ist es häufig der Fall, dass die Konsequenzen oder *„Kosten" der Fehlklassifikation* zwischen den Gruppen differieren. So ist z. B. in der medizinischen Diagnostik der Schaden, der dadurch entsteht, dass eine bösartige Krankheit nicht rechtzeitig erkannt wird, sicherlich größer, als die irrtümliche Diagnose einer bösartigen Krankheit. Das Beispiel macht gleichzeitig deutlich, dass die Bewertung der „Kosten" sehr schwierig sein kann. Eine ungenaue Bewertung aber ist i. d. R. besser als keine Bewertung und damit keine Berücksichtigung der unterschiedlichen Konsequenzen.

Die Berücksichtigung von ungleichen Kosten der Fehlklassifikation kann durch Anwendung der *Bayes'schen Entscheidungsregel* erfolgen, die auf dem Konzept des statistischen Erwartungswertes basiert.[25] Es ist dabei gleichgültig, ob der Erwartungswert eines Kosten- bzw. Verlustkriteriums minimiert oder der eines Nutzen- bzw. Gewinnkriteriums maximiert wird.

[25] Vgl. dazu z. B. Schneeweiss (1967); Bamberg/Coenenberg (1992).

Klassifizierung durch Anwendung der Bayes-Regel

Bayes-Regel

Ordne ein Element i derjenigen Gruppe g zu, für die der Erwartungswert der Kosten

$$E_g(K) = \sum_{h=1}^{G} K_{gh} P(h|Y_i) \qquad (g = 1, \ldots, G) \tag{4.29}$$

minimal ist. Dabei bezeichnet

$P(h|Y_i)$ = Wahrscheinlichkeit für die Zugehörigkeit von Element i mit Diskriminanzwert Y_i zu Gruppe h $(h = 1, \ldots, G)$

K_{gh} = Kosten der Einstufung in Gruppe g, wenn das Element zu Gruppe h gehört

Beispiel

Die Anwendung der Bayes-Regel soll an einem kleinen *Beispiel* verdeutlicht werden. Ein Bankkunde i möchte einen Kredit in Höhe von 1.000€ für ein Jahr zu einem Zinssatz von 10 % aufnehmen. Für die Bank stellt sich das Problem, den möglichen Zinsgewinn gegen das Risiko eines Kreditausfalls abzuwägen. Für den Kunden wurden folgende *Klassifizierungswahrscheinlichkeiten* ermittelt:

$$P(1|Y_i) = 0,8 \qquad \text{(Kreditrückzahlung)}$$
$$P(2|Y_i) = 0,2 \qquad \text{(Kreditausfall)}$$

Kosten der Fehlklassifikation

Wenn die Einordnung in Gruppe 1 mit einer Vergabe des Kredites und die Einordnung in Gruppe 2 mit einer Ablehnung gekoppelt ist, so lassen sich die folgenden *Kosten einer Fehlklassifikation* angeben (Abbildung 4.21):

Einordnung in Gruppe g	tatsächliche Gruppenzugehörigkeit	
	Rückzahlung	Ausfall
	1	2
1: Vergabe	-100	1000
2: Ablehnung	100	0

Abbildung 4.21: Kosten einer Fehlkalkulation (Beispiel)

Vergibt die Bank den Kredit, so erlangt sie bei ordnungsgemäßer Tilgung einen Gewinn (negative Kosten) in Höhe von 100€, während ihr bei Zahlungsunfähigkeit des Kunden ein Verlust in Höhe von 1000€ entsteht. Vergibt die Bank dagegen den Kredit nicht, so entstehen ihr eventuell Opportunitätskosten (durch entgangenen Gewinn) in Höhe von 100€.

Die *Erwartungswerte* der Kosten für die beiden Handlungsalternativen errechnen sich mit den obigen Wahrscheinlichkeiten wie folgt:

$$\begin{aligned} Vergabe: \quad & E_1(K) = -100 \cdot 0,8 + 1000 \cdot 0,2 = 120 \\ Ablehnung: \quad & E_2(K) = \quad 100 \cdot 0,8 + \quad 0 \cdot 0,2 = 80 \end{aligned}$$

Die erwarteten Kosten der zweiten Alternative sind niedriger. Folglich ist der Kreditantrag bei Anwendung der Bayes-Regel abzulehnen, obgleich die Wahrscheinlichkeit einer Kreditrückzahlung weit höher ist als die eines Kreditausfalls.

4.2.6.4 Berechnung der Klassifizierungswahrscheinlichkeiten

Die Klassifizierungswahrscheinlichkeiten lassen sich aus den Distanzen unter Anwendung des Bayes-Theorems wie folgt berechnen (vgl. mathematischer Anhang C):

$$P(g|Y_i) = \frac{\exp\left(-D_{ig}^2/2\right) P_i(g)}{\sum_{g=1}^{G} \exp\left(-D_{ig}^2/2\right) P_i(g)} \qquad (g = 1, \ldots, G) \tag{4.30}$$

mit

D_{ig} = Distanz zwischen Element i und dem Centroid von Gruppe g
$P_i(g)$ = A-priori-Wahrscheinlichkeit für die Zugehörigkeit von Element i zu Gruppe g

Beispiel: Für ein Element i mit den Merkmalswerten

$$X_{1i} = 6 \quad \text{und} \quad X_{2i} = 7$$

erhält man durch Anwendung der oben ermittelten Diskriminanzfunktion den folgenden Diskriminanzwert:

$$Y_i = -1,98 + 1,031 \cdot 6 - 0,565 \cdot 7 = 0,252$$

Zuordnungsbeispiel

Bezüglich der beiden Gruppen A und B erhält man gemäß (4.25) und mit den Mittelwerten aus Abbildung 4.11 die *quadrierten Distanzen*:

$$D_{ig}^2 = \left(Y_i - \bar{Y}_g\right)^2$$
$$D_{iA}^2 = (0,252 - (-0,914))^2 = 1,360$$
$$D_{iB}^2 = (0,252 - 0,914)^2 = 0,438$$

Die Transformation der Distanzen liefert die Werte (Dichten):

$$f(Y_i|g) = \exp\left(-D_{ig}^2/2\right)$$
$$f(Y_i|A) = 0,507$$
$$f(Y_i|B) = 0,803$$

Damit erhält man durch (4.30) unter Vernachlässigung von A-priori-Wahrscheinlichkeiten die gesuchten Klassifizierungswahrscheinlichkeiten:

$$P(g|Y_i) = \frac{f(Y_i|g)}{f(Y_i|A) + f(Y_i|B)}$$
$$P(A|Y_i) = \frac{0,507}{0,507 + 0,803} = 0,39$$
$$P(B|Y_i) = \frac{0,803}{0,507 + 0,803} = 0,61$$

Das Element i ist folglich in die Gruppe B einzuordnen. Dasselbe Ergebnis liefert auch das Distanzkonzept. Unterschiedliche Ergebnisse können sich nur bei Einbeziehung von unterschiedlichen A-priori-Wahrscheinlichkeiten ergeben.

Sind die *A-priori-Wahrscheinlichkeiten*

$$P_i(A) = 0,4 \quad \text{und} \quad P_i(B) = 0,6$$

gegeben und sollen diese in die Schätzung einbezogen werden, so erhält man stattdessen die folgenden Klassifizierungswahrscheinlichkeiten:

$$P(g|Y_1) = \frac{f(Y_i|g)P_i(g)}{f(Y_i|A)P_i(A) + f(Y_i|B)P_i(B)}$$
$$P(A|Y_i) = \frac{0,507 \cdot 0,4}{0,507 \cdot 0,4 + 0,803 \cdot 0,6} = 0,30$$
$$P(B|Y_i) = \frac{0,803 \cdot 0,6}{0,507 \cdot 0,4 + 0,803 \cdot 0,6} = 0,70$$

Da sich hier die in den Merkmalswerten enthaltene Information und die A-priori-Information gegenseitig bestärken, erhöht sich die relative Sicherheit für die Einordnung von Element i in Gruppe B.

Annahme gleicher Streuungen

Die obigen Berechnungen basieren auf der *Annahme gleicher Streuungen* (Kovarianzmatrizen) in den Gruppen. Die Überprüfung dieser Annahme mit Hilfe von Box's M liefert für das Beispiel ein Signifikanzniveau von über 95 %. Die Nullhypothese, dass die Kovarianzmatrizen gleich sind, kann damit nicht abgelehnt werden und die Annahme gleicher Kovarianzmatrizen erscheint damit gerechtfertigt. Bei Anwendung von SPSS kann die Klassifizierung wahlweise unter der Annahme gleicher Streuungen (Voreinstellung) wie auch unter Berücksichtigung der individuellen Streuungen in den Gruppen durchgeführt werden.

4.2.6.5 Überprüfung der Klassifizierung

Die Summe der Klassifizierungswahrscheinlichkeiten, die man durch Anwendung des Bayes-Theorems erhält, ergibt immer Eins. Die Anwendung des Bayes-Theorems schließt also aus, dass ein zu klassifizierendes Element eventuell keiner der vorgegebenen Gruppen angehört. Die Klassifizierungswahrscheinlichkeiten erlauben deshalb auch keine Aussage darüber, ob und wie wahrscheinlich es ist, dass ein klassifiziertes Element überhaupt einer der betrachteten Gruppen angehört.

Kontrolle der Klassifizierung

Aus diesem Grunde ist es zur Kontrolle der Klassifizierung zweckmäßig, für die gewählte Gruppe g (mit der höchsten Klassifizierungswahrscheinlichkeit) die bedingte Wahrscheinlichkeit $P(Y_i|g)$ zu überprüfen. In Formel (4.30) wurde die explizite Berechnung der bedingten Wahrscheinlichkeiten umgangen. Sie müssen daher bei Bedarf gesondert ermittelt werden.

Die bedingte Wahrscheinlichkeit ist in Abbildung 4.22 dargestellt. Je größer die Distanz D_{ig} wird, desto unwahrscheinlicher wird es, dass für ein Element von Gruppe g eine gleich große oder gar größere Distanz beobachtet wird, und desto geringer wird damit die Wahrscheinlichkeit der Hypothese „Element i gehört zu Gruppe g“. Die bedingte Wahrscheinlichkeit $P(Y_i|g)$ ist die Wahrscheinlichkeit bzw. das Signifikanzniveau dieser Hypothese.

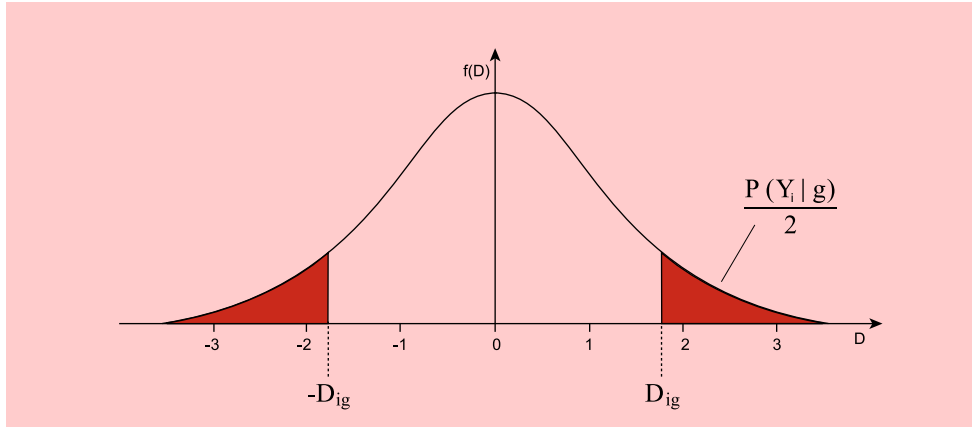


Abbildung 4.22: Darstellung der bedingten Wahrscheinlichkeit (rote Fläche) unter der Dichtefunktion der standardisierten Normalverteilung

Im Gegensatz zu den A-priori- und A-posteriori-Wahrscheinlichkeiten müssen sich die bedingten Wahrscheinlichkeiten über die Gruppen nicht zu Eins addieren. Die bedingten Wahrscheinlichkeiten eines Elementes können daher bezüglich aller Gruppen beliebig klein sein. Da die bedingte Wahrscheinlichkeit für die Gruppe mit der höchsten Klassifizierungswahrscheinlichkeit am größten ist, braucht sie nur für diese Gruppe überprüft zu werden. Etwas anderes kann gelten, wenn A-priori-Wahrscheinlichkeiten berücksichtigt wurden.

Bedingte Wahrscheinlichkeit

Die bedingte Wahrscheinlichkeit lässt sich mit Hilfe einer Tabelle der standardisierten Normalverteilung bestimmen.[26] Für das oben betrachtete Element mit dem Diskriminanzwert $Y_i = 0,252$ und der minimalen Distanz

$$D_{iB} = 0,914 - 0,252 = 0,662$$

erhält man die bedingte Wahrscheinlichkeit

$$P(Y_i|B) = 0,51$$

Gut die Hälfte aller Elemente der Gruppe B ist also weiter entfernt vom Centroid, als das Element i. Das Element i fällt daher nicht durch ungewöhnliche Merkmalsausprägungen auf.

Im Vergleich dazu sei ein Element r mit den Merkmalswerten

$$X_{1r} = 1 \quad \text{und} \quad X_{2r} = 6$$

betrachtet. Für dieses Element erhält man den Diskriminanzwert $Y_r = -4,339$ und die Klassifizierungswahrscheinlichkeiten

$$P(A|Y_r) = 0,9996$$
$$P(B|Y_r) = 0,0004$$

[26] Eine genauere Berechnung ermöglicht Excel: p = (1-STANDNORMVERT(z))*2 ergibt für z = 0.662 den Wert p = 0,508.

Das Element wäre also der Gruppe A zuzuordnen. Die Distanz zum Centroid von Gruppe A beträgt $D_{rA} = 3,42$. Damit ergibt sich bezüglich Gruppe A die bedingte Wahrscheinlichkeit

$$P(Y_r|A) = 0,0006$$

Die Wahrscheinlichkeit dafür, dass ein Element der Gruppe A eine so große Distanz aufweist, wie das Element r, ist also außerordentlich gering. Bezüglich Gruppe B wäre die bedingte Wahrscheinlichkeit natürlich noch geringer. Man muss sich daher fragen, ob dieses Element überhaupt einer der beiden Gruppen angehört.

4.3 Fallbeispiel

4.3.1 Problemstellung

Beispiel

Nachfolgend soll die Diskriminanzanalyse an einem Fallbeispiel unter Anwendung des Computer-Programms SPSS durchgeführt werden. Nachdem unser Margarinehersteller sich mit der Frage befasst hat, welche Eigenschaften bei einer Margarine von Wichtigkeit sind, möchte er jetzt herausfinden, wie die Margarinemarken selbst wahrgenommen werden, d. h.

Fragestellung

- ob signifikante Unterschiede in der Wahrnehmung verschiedener Marken bestehen und
- welche Eigenschaften für die unterschiedliche Wahrnehmung der Marken relevant sind.

Zu diesem Zweck wurde eine Befragung von 18 Personen durchgeführt, wobei diese veranlasst wurden, 11 Butter- und Margarinemarken jeweils bezüglich 10 verschiedener Variablen auf einer siebenstufigen Rating-Skala zu beurteilen (vgl. Abbildung 4.23). Da nicht alle Personen alle Marken beurteilen konnten, umfasst der Datensatz nur 127 Markenbeurteilungen anstelle der vollständigen Anzahl von 198 Markenbeurteilungen (18 Personen x 11 Marken). Eine Markenbeurteilung umfasst dabei die Skalenwerte der 10 Merkmalsvariablen.

Markenbeurteilung

Von den 127 Markenbeurteilungen sind nur 92 vollständig, während die restlichen 35 Beurteilungen fehlende Werte, sog. Missing Values, enthalten. Missing Values bilden ein unvermeidliches Problem bei der Durchführung von Befragungen (z. B. weil Personen nicht antworten können oder wollen oder als Folge von Interviewerfehlern). Die unvollständigen Beurteilungen sollen zunächst in der Diskriminanzanalyse nicht berücksichtigt werden, sodass sich die Fallzahl auf 92 verringert. In SPSS existieren verschiedene Optionen zur Behandlung von Missing Values.

Missing Values

Um die Zahl der Gruppen zu vermindern, wurden die 11 Marken zu drei Gruppen (Marktsegmenten) zusammengefasst. Die Gruppenbildung wurde durch Anwendung einer Clusteranalyse vorgenommen (vgl. Kapitel 8). In Abbildung 4.24 ist die Zusammensetzung der Gruppen angegeben. Mittels Diskriminanzanalyse soll jetzt untersucht werden, ob und wie sich diese Gruppen unterscheiden.[27]

[27] Da die Beurteilungen der verschiedenen Marken von denselben Personen vorgenommen werden, lässt sich gegen die Anwendung der Diskriminanzanalyse einwenden, dass die Stichproben der Gruppen (Markencluster) nicht unabhängig sind. Dieses inhaltliche Problem soll hier zugunsten der Demonstration von SPSS an einem größeren Datensatz zurückgestellt werden.

Emulsionsfette (Butter und Margarine)		Merkmalsvariablen (subjektive Beurteilungen)	
1	Sanella	1	Streichfähigkeit
2	Homa	2	Preis
3	SB	3	Haltbarkeit
4	Delicado	4	Anteil ungesättigter Fettsäuren
5	Holländische Markenbutter	5	Back- und Brateignung
6	Weihnachtsbutter	6	Geschmack
7	Du darfst	7	Kaloriengehalt
8	Becel	8	Anteil tierischer Fette
9	Botteram	9	Vitamingehalt
10	Flora Soft	10	Natürlichkeit
11	Rama		

Abbildung 4.23: Untersuchte Marken und Variablen im Fallbeispiel

Eine weitergehende Problemstellung, der hier allerdings nicht nachgegangen werden soll, könnte in der Kontrolle der Marktpositionierung eines neuen Produktes bestehen. Mittels der oben behandelten Techniken der Klassifizierung ließe sich überprüfen, ob das Produkt sich bezüglich seiner Wahrnehmung durch die Konsumenten in das angestrebte Marktsegment einordnet.

Marktsegmente (Gruppen)	Marken im Segment
A	Du darfst, Becel
B	Sanella, Homa, SB, Botteram, Flora Soft, Rama
C	Delicado, Holländische Markenbutter, Weihnachtsbutter

Abbildung 4.24: Definition der Gruppen

4.3.2 Ergebnisse

Im Folgenden wird einerseits gezeigt, wie mit SPSS die Diskriminanzanalyse durchgeführt wird. Andererseits werden die wichtigsten Ergebnisse des Programmausdrucks von SPSS wiedergegeben und kommentiert. Abbildung 4.25 zeigt zunächst, wie das relevante Analyseverfahren Diskriminanzanalyse aus dem Menüpunkt „Analysieren“ als eine der klassifizierenden Prozeduren aufgerufen wird.

Auswahl des Verfahrens

Nachdem die Diskriminanzanalyse als Verfahren ausgewählt wurde, wird das in Abbildung 4.26 wiedergegebene Dialogfenster geöffnet. In diesem Beispiel zur Diskriminanzanalyse soll untersucht werden, ob und wie sich die drei aus einer vorherigen Clusteranalyse resultierenden Gruppen unterscheiden. In der Variable „Segment“ ist für jede Marke die Gruppenzugehörigkeit enthalten. Diese ist somit aus der linken Variablenliste auszuwählen und in das Feld „Gruppenvariable“ zu verschieben. Die Festlegung der für die Analyse relevanten Gruppen erfolgt über die Schaltfläche „Bereich definieren“ unterhalb der Gruppenvariable.

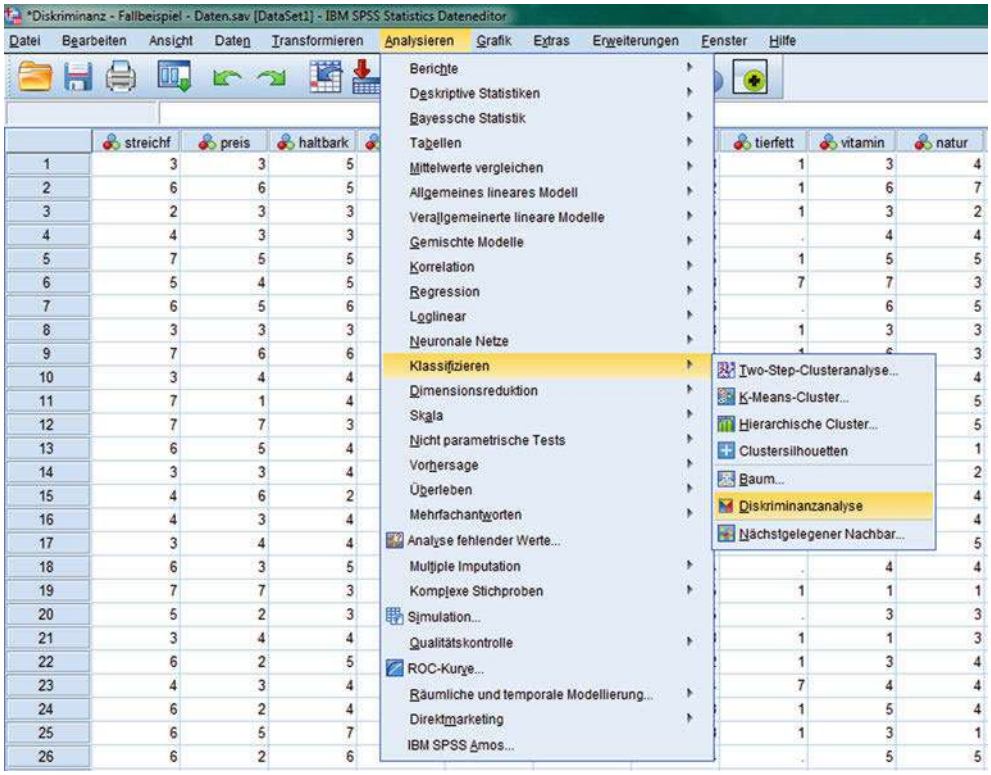


Abbildung 4.25: Daten-Editor mit Auswahl des Analyseverfahrens „Diskriminanzanalyse“

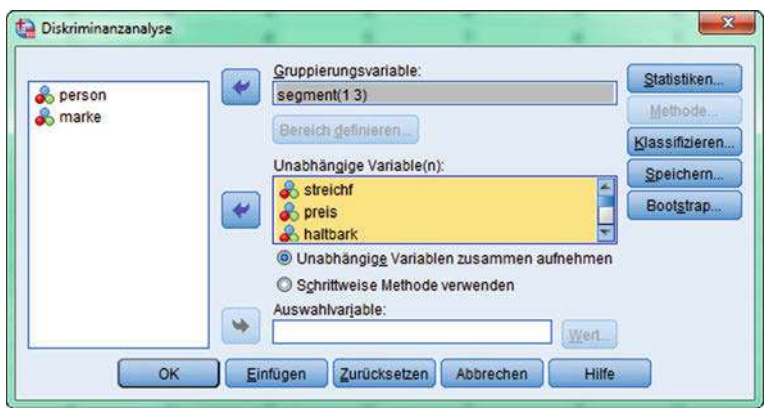


Abbildung 4.26: Dialogfenster „Diskriminanzanalyse“

Für die Durchführung der Diskriminanzanalyse existieren, wie auch bei der Regressionsanalyse, zwei Methoden, eine blockweise (direkte) Methode und eine schrittweise Methode. Bei der blockweisen Methode werden alle Merkmalsvariablen, die der Untersucher auswählt, simultan in die Diskriminanzfunktion aufgenommen, bei der schrittweisen Methode dagegen erfolgt die Aufnahme sukzessiv und wird durch einen Algorithmus gesteuert. Voreingestellt ist die blockweise Methode (Option: „Unab-

hängige Variablen zusammen aufnehmen“), mit der wir hier beginnen. Zur Auswahl der Merkmalsvariablen sind diese in das Fenster „Unabhängige Variable(n)“ zu verschieben. Mittels der Schaltfläche „Statistiken“ lassen sich über den Standardoutput hinaus zusätzliche Analysen und Ergebnisse anfordern. Nach Anklicken öffnet sich das Dialogfenster „Statistik“ in Abbildung 4.27.

Abbildung 4.27: Dialogfenster „Statistik“

Univariate Prüfung der Diskriminanz

Zunächst sollen die unabhängigen Variablen separat auf ihre diskriminatorische Eignung hin untersucht werden, d.h. es soll untersucht werden, wie gut die 10 Merkmalsvariablen jeweils isoliert zwischen den drei Gruppen trennen (vgl. dazu Abschnitt 4.2.5). Dazu ist im Dialogfenster „Statistik“ die Option „Univariate ANOVA“ auszuwählen. Das Ergebnis zeigt Abbildung 4.28. Mit Ausnahme der Variablen „Haltbark“, „Ungefett“ und „Backeign“ trennen alle Variablen signifikant mit einer Irrtumswahrscheinlichkeit unter 5 %. Am besten trennt die Variable „Natur“.

Gleichheitstest der Gruppenmittelwerte

	Wilks-Lambda	F	df1	df2	Signifikanz
streichf	,798	11,246	2	89	,000
preis	,916	4,074	2	89	,020
haltbark	,952	2,264	2	89	,110
ungefett	,993	,321	2	89	,726
backeign	,944	2,619	2	89	,078
geschmac	,795	11,484	2	89	,000
kalorien	,836	8,703	2	89	,000
tierfett	,712	17,980	2	89	,000
vitamin	,885	5,806	2	89	,004
natur	,703	18,813	2	89	,000

Abbildung 4.28: Univariate Trennfähigkeit der Merkmalsvariablen

Ermittlung und Beurteilung der Diskriminanzfunktionen

Diskriminanzfunktionen

Bei drei Gruppen lassen sich zwei *Diskriminanzfunktionen* bilden. Um die kanonischen Diskriminanzkoeffizienten (nicht-standardisierte Diskriminanzkoeffizienten) im Output zu erhalten, ist im Dialogfenster „Statistik“ die Option „Nicht standardisiert“ auszuwählen. Standardmäßig werden nur die standardisierten Diskriminanzkoeffizienten ausgegeben. In Abbildung 4.29 sind die geschätzten Diskriminanzkoeffizienten der beiden Diskriminanzfunktionen wiedergegeben. Damit lassen sich die Diskriminanzwerte und die Mittelpunkte (Centroide) der drei Gruppen bezüglich der beiden Diskriminanzfunktionen berechnen. Sie sind im unteren Teil von Abbildung 4.29 angegeben. Abbildung 4.30 (Standardoutput) enthält die in Abschnitt 4.2.4.2 behandelten *Gütemaße* zur Beurteilung der Diskriminanzfunktionen. Die Fußnote a in der zweiten Spalte des oberen Abschnitts der abgebildeten Tabelle zeigt an, dass beide Diskriminanzfunktionen bei der Klassifizierung berücksichtigt werden.

Kanonische Diskriminanzfunktionskoeffizienten

	Funktion 1	Funktion 2
streichf	-,140	,408
preis	,223	-,127
haltbark	-,336	-,276
ungefett	-,091	-,126
backeign	-,020	,131
geschmac	,190	,372
kalorien	,268	-,102
tierfett	,189	,166
vitamin	-,180	,429
natur	,486	-,332
(Konstant)	-2,164	-2,322

Nicht-standardisierte Koeffizienten

Funktionen bei den Gruppen-Zentroiden

segment	Funktion 1	Funktion 2
Segment A	-,773	,885
Segment B	-,613	-,349
Segment C	2,088	,045

Nicht-standardisierte kanonische Diskriminanzfunktionen, die bezüglich des Gruppen-Mittelwertes bewertet werden

Abbildung 4.29: Koeffizienten der beiden Diskriminanzfunktionen und Werte der Centroide

Eigenwerte

Funktion	Eigenwert	% der Varianz	Kumulierte %	Kanonische Korrelation
1	1,420[a]	85,7	85,7	,766
2	,238[a]	14,3	100,0	,438

a. Die ersten 2 kanonischen Diskriminanzfunktionen werden in dieser Analyse verwendet.

Wilks-Lambda

Test der Funktion(en)	Wilks-Lambda	Chi-Quadrat	df	Signifikanz
1 bis 2	,334	92,718	20	,000
2	,808	18,029	9	,035

Abbildung 4.30: Gütemaße der Diskriminanzfunktionen

Eigenwertanteil

Aus Spalte 2 und 3 im oberen Teil ist ersichtlich, dass die relative Wichtigkeit der zweiten Diskriminanzfunktion mit 14,3 % Eigenwertanteil (Varianzanteil) wesentlich geringer ist als die der ersten Diskriminanzfunktion mit 85,7 % Eigenwertanteil (vgl. Abschnitt 4.2.3.6). Die kumulativen Eigenwertanteile in Spalte 4 erhält man durch Summierung der Eigenwertanteile. Für den letzten Wert (hier den zweiten) muss sich daher immer 100 ergeben. Die folgende Spalte enthält die kanonischen Korrelationskoeffizienten gemäß Formel (4.16).

Interpolation des Outputs

Im unteren Teil der Abbildung findet man die Werte für das residuelle Wilks' Lambda (nach Bildung von 0 und 1 Diskriminanzfunktionen) gemäß Formel (4.20).

Daneben sind die zugehörigen χ^2-Werte nebst Freiheitgraden und Signifikanzniveau angegeben. Sie zeigen, dass auch die zweite Diskriminanzfunktion noch signifikant (mit Irrtumswahrscheinlichkeit = 3,5 %) zur Trennung der Gruppen beiträgt.

Abbildung 4.31 zeigt die *standardisierten Diskriminanzkoeffizienten*, aus denen sich die Wichtigkeit der Merkmalsvariablen innerhalb der beiden Diskriminanzfunktionen erkennen lässt. Die größte diskriminatorische Bedeutung besitzt die Variable „Natur" für die Diskriminanzfunktion 1 und die Variable „Streichf" für die Diskriminanzfunktion 2. Mit Hilfe der Option „Bootstrap" (siehe Abbildung 4.26) lassen sich auch Standardfehler und Konfidenzintervalle für die geschätzten standardisierten Diskriminanzkoeffizienten ermitteln.

Um die diskriminatorische Bedeutung einer Merkmalsvariablen bezüglich aller Diskriminanzfunktionen zu beurteilen, sind gemäß Formel (4.22) durch Gewichtung der absoluten Werte der Koeffizienten mit dem Eigenwertanteil der betreffenden Diskriminanzfunktion die *mittleren Diskriminanzkoeffizienten* zu berechnen.

Mittlerer Diskriminanzkoeffizient

Es ergibt sich hier mit den Eigenwertanteilen aus Abbildung 4.30 für die Variable 5 = „Backeign" der niedrigste und für die Variable 10 = „Natur" der höchste Wert für den mittleren Diskriminanzkoeffizienten:

$$\bar{b}_5 = 0,032 \cdot 0,857 + 0,214 \cdot 0,143 = 0,058$$
$$\bar{b}_{10} = 0,620 \cdot 0,857 + 0,423 \cdot 0,143 = 0,592$$

Die Variable „Backeign" besitzt somit die geringste und die Variable „Natur" die größte diskriminatorische Bedeutung.

Standardisierte kanonische Diskriminanzfunktionskoeffizienten

	Funktion 1	2
streichf	-,206	,598
preis	,359	-,204
haltbark	-,396	-,325
ungefett	-,130	-,180
backeign	-,032	,214
geschmac	,243	,475
kalorien	,390	-,148
tierfett	,443	,389
vitamin	-,238	,566
natur	,620	-,423

Abbildung 4.31: Standardisierte Diskriminanzkoeffizienten

Klassifizierung

Abbildung 4.32 zeigt die geschätzten Koeffizienten der *Klassifizierungsfunktionen* nach R.A. Fisher (vgl. Abschnitt 4.2.6.1). Um sie zu erhalten, ist im Dialogfenster „Statistik“ (Abbildung 4.27) die Option „Fisher“ auszuwählen.

Klassifizierungsfunktion

Klassifizierungsfunktionskoeffizienten

	segment Segment A	Segment B	Segment C
streichf	2,516	1,990	1,772
preis	,576	,768	1,320
haltbark	1,580	1,866	,850
ungefett	1,714	1,855	1,559
backeign	,159	-,005	-,007
geschmac	,351	-,077	,584
kalorien	,855	1,025	1,709
tierfett	1,122	,948	1,523
vitamin	-,083	-,641	-,958
natur	1,516	2,004	3,187
(Konstant)	-23,073	-20,111	-28,805

Lineare Diskriminanzfunktionen nach Fisher

Abbildung 4.32: Koeffizienten der Klassifizierungsfunktionen

Abbildung 4.33: Dialogfenster „Klassifizieren“

Dialogfenster „Klassifizieren“

Über das Dialogfenster „Klassifizieren“ (vgl. Abbildung 4.33) ist es möglich, die A-priori-Wahrscheinlichkeiten sowie die Kovarianzen zur Klassifizierung der Gruppen festzulegen. Des Weiteren können hier zusätzlich Auswertungen und Diagramme angefordert werden. Bezüglich der A-priori-Wahrscheinlichkeit und der Kovarianzmatrix wird hier die jeweilige Voreinstellung beibehalten. Es soll jedoch zusätzlich eine „Zusammenfassungstabelle“ ausgegeben werden, die Abbildung 4.34 zeigt.

In der Zusammenfassungstabelle sind zunächst die Anzahl der verarbeiteten Fälle und die Anzahl der wegen fehlender Werte ausgeschlossenen Fälle angegeben. Sodann sind nach den Gruppen gegliedert die verwendeten Fälle angegeben. Insgesamt wurden 92 vollständige Beurteilungen (ohne Missing Values) abgegeben, die für die Klassifizierung verwendet wurden.

Die A-priori-Wahrscheinlichkeiten werden, wenn vom Benutzer nichts anderes verlangt wird, gleichmäßig auf die Gruppen verteilt und haben damit keinen Einfluss auf die Klassifizierungsergebnisse.

Im unteren Teil von Abbildung 4.34 ist die *Klassifikationsmatrix* (vgl. Abschnitt 4.2.4.1) angegeben. Sie zeigt zum einen für jede Gruppe die Häufigkeiten der tatsächlichen und der geschätzten Gruppenzugehörigkeit und zum anderen die Trefferquoten. Insgesamt wurden 69 der 92 Fälle richtig zugeordnet. Das ergibt eine Trefferquote von 75 %. Bei zufälliger Einordnung der Elemente in die drei Gruppen wäre dagegen (unter Vernachlässigung der unterschiedlichen Gruppengrößen) eine Trefferquote von 33,3 % zu erwarten. Auch für die kleinen Gruppen (Segment A und C) ergeben sich passable Trefferquoten.

Individuelle Klassifizierungsergebnisse

Im Dialogfenster „Klassifizieren“ (Abbildung 4.33) wurde die Option „Fallweise Ergebnisse“ ausgewählt. Die Ausgabe wurde dabei auf die ersten 15 Fälle beschränkt. Diese Werte sind in Abbildung 4.35 wiedergegeben. Für jedes Element lassen sich die folgenden Angaben entnehmen (Vgl. hierzu den Abschnitt 4.2.6):

- Tatsächliche Gruppenzugehörigkeit eines Elements.
- Geschätzte (vorhergesagte) Gruppenzugehörigkeit. Falsche Zuordnungen sind durch Sternchen gekennzeichnet.

Klassifizierungsergebnisse[a]

			Vorhergesagte Gruppenzugehörigkeit			
		segment	Segment A	Segment B	Segment C	Gesamt
Original	Anzahl	Segment A	12	7	0	19
		Segment B	9	38	4	51
		Segment C	3	0	19	22
	%	Segment A	63,2	36,8	,0	100,0
		Segment B	17,6	74,5	7,8	100,0
		Segment C	13,6	,0	86,4	100,0

a. 75,0% der ursprünglich gruppierten Fälle wurden korrekt klassifiziert.

Zusammenfassung der Verarbeitung von Klassifizierungen

Verarbeitet		127
Ausgeschlossen	Fehlende oder außerhalb des Bereichs liegende Gruppencodes	0
	Wenigstens eine Diskriminanzvariable fehlt	35
In der Ausgabe verwendet		92

Abbildung 4.34: Zusammenfassungstabelle mit Klassifikationsmatrix

- Bedingte Wahrscheinlichkeit P(D>d|G=g) dafür, dass ein Element der Gruppe g eine Distanz > d zum Centroid von Gruppe g aufweist.[28]
- Klassifizierungswahrscheinlichkeit P(G=g|D=d), also die A-posteriori-Wahrscheinlichkeit dafür, dass ein Element mit Distanz d zur Gruppe g gehört. Sie errechnet sich über das Bayes-Theorem (4.27) aus den A-priori-Wahrscheinlichkeiten und den bedingten Wahrscheinlichkeiten. Aus ihnen ist ersichtlich, wie viel Vertrauen man in die Zuordnung eines Elements zu einer bestimmten Gruppe haben kann. So wird z.B. das erste Element der Gruppe 2 mit einer Wahrscheinlichkeiten von 97 % zugeordnet. Man kann daher recht sicher sein, dass diese Zuordnung korrekt ist.
- Mahalanobis-Distanz zum Centroid der prognostizierten Gruppe.
- Es werden außerdem die Wahrscheinlichkeiten und Distanzen für die Gruppe mit der jeweils zweithöchsten Klassifizierungswahrscheinlichkeit angegeben. Für das erste Element beträgt die Wahrscheinlichkeit für die Zuordnung zu Gruppe 1 nur 2,8 % gegenüber 97 % für Gruppe 2.
- In der letzten Spalte finden sich die Diskriminanzwerte bezüglich der beiden Diskriminanzfunktionen.

[28] Wir hatten oben eine etwas andere Formulierung verwendet, indem wir für die einfache Diskriminanzanalyse mit nur einer Diskriminanzfunktion Wahrscheinlichkeiten als Funktion der Diskriminanzwerte betrachtet hatten und nicht, wie hier, als Funktion der Distanzen. Durch den Diskriminanzwert aber ist die Distanz eindeutig bestimmt.

Abbildung 4.35: Individuelle Klassifizierungsergebnisse

Fallweise Statistiken

				Höchste Gruppe				Zweithöchste Gruppe			Diskriminanzwerte	
	Fallnummer	Tatsächliche Gruppe	Vorhergesagte Gruppe	P(D>d | G=g) p	df	P(G=g | D=d)	Quadrierter Mahalanobis-Abstand zum Zentroid	Gruppe	P(G=g | D=d)	Quadrierter Mahalanobis-Abstand zum Zentroid	Funktion 1	Funktion 2
Original	1	2	2	,057	2	,970	5,728	1	,028	12,818	-1,202	-2,669
	2	2	2	,533	2	,471	1,259	1	,306	2,120	,430	,065
	3	2	2	,553	2	,879	1,185	1	,105	5,436	-,564	-1,437
	5	2	1**	,757	2	,498	,556	2	,457	,728	-,206	,401
	6	2	1**	,237	2	,945	2,877	2	,053	8,637	-,850	2,580
	8	2	2	,455	2	,892	1,575	1	,103	5,885	-1,055	-1,524
	9	2	1**	,482	2	,904	1,461	2	,093	6,011	-,826	2,093
	10	2	2	,533	2	,733	1,260	3	,144	4,521	,265	-1,049
	11	2	2	,562	2	,550	1,151	1	,449	1,560	-1,634	-,020
	12	2	2	,936	2	,756	,132	1	,223	2,578	-,513	-,699
	17	2	2	,386	2	,425	1,903	3	,422	1,916	,767	-,367
	19	2	2	,899	2	,780	,212	1	,203	2,908	-,580	-,809
	21	2	2	,235	2	,930	2,895	1	,051	8,681	-,428	-2,041
	22	2	2	,575	2	,534	1,108	1	,465	1,384	-1,591	,040
	23	2	3**	,291	2	,403	2,466	1	,389	2,536	,817	,969

**. Falsch klassifizierter Fall

Zwecks weiterer Verarbeitung können die Individualdaten auch in die Datendatei abgespeichert werden. Sie werden dann als neue Variablen an den bestehenden Datensatz im Dateneditor angehängt. Hierzu ist im Dialogfenster der Diskriminanzanalyse die Option „Speichern“ zu wählen, worauf sich das in Abbildung 4.36 gezeigte Dialogfenster öffnet.

Abbildung 4.36: Dialogfenster „Speichern“

Grafische Darstellungen

Jedes Element umfasst hier die Ausprägungen von 10 Merkmalsvariablen. Es lässt sich damit als Punkt in einem 10-dimensionalen Merkmalsraum vorstellen, was allerdings nicht darstellbar ist. Nach Durchführung der Diskriminanzanalyse aber lassen sich die Elemente als Punkte in einem zweidimensionalen Raum darstellen, der durch die beiden Diskriminanzfunktionen aufgespannt wird. Die Diskriminanzwerte aus Abbildung 4.35 können dabei als Koordinaten der Punkte verwendet werden. Insbesondere bildet der Diskriminanzraum hier eine (Diskriminanzebene), die der Diskriminanzachse im Zwei-Gruppen-Fall (bei nur einer Diskriminanzfunktion) entspricht.

Um eine derartige Grafik zu erhalten, ist im Dialogfenster „Klassifizieren“ (Abbilung 4.37) die Option „Kombinierte Gruppen“ zu wählen. Das Ergebnis zeigt Abbildung 4.38. Die Gruppencentroide sind ebenfalls in dieser Abbildung markiert. Mittels der Option „Gruppenspezifisch“ lässt sich auch für jede Gruppe separat eine derartiges Streudiagramm erstellen.

Abbildung 4.37: Dialogfenster „Klassifizieren“

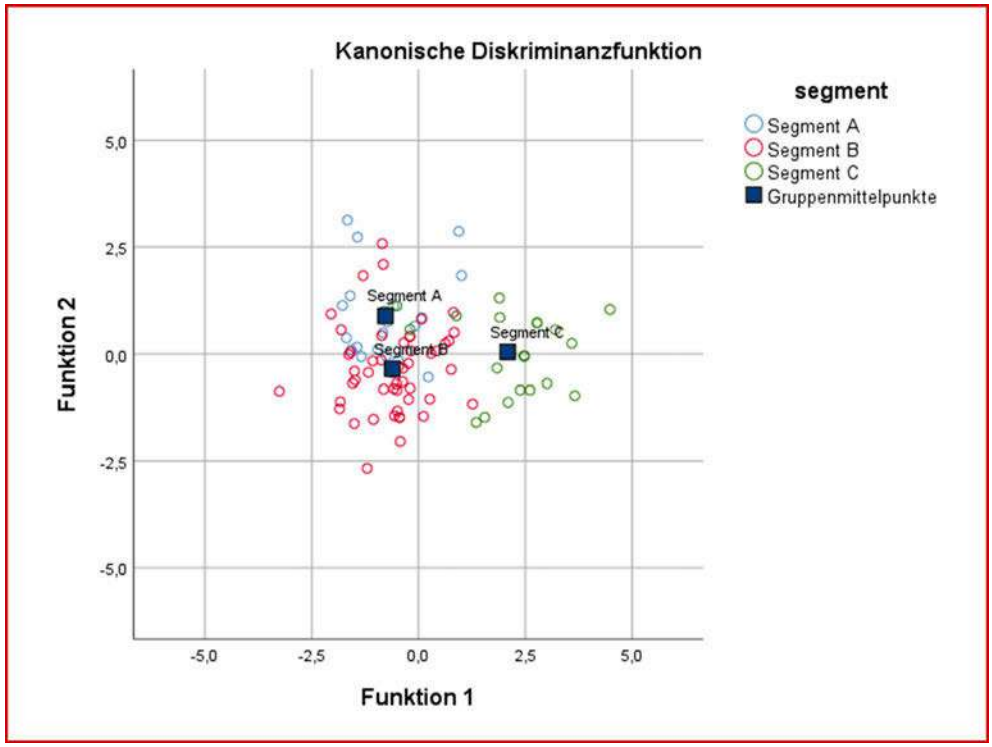


Abbildung 4.38: Darstellung der Gruppen im Diskriminanzraum

Eine weitere Darstellung ist das in Abbildung 4.39 dargestellte *Klassifizierungsdiagramm*. Man erhält es mittels der Option „Territorien“ im Dialogfenster „Klassifizieren“. Das Diagramm zeigt die Aufteilung der Diskriminanzebene in Gebiete (Territorien), die den Zugehörigkeitsbereich der Gruppen markieren. Innerhalb der Gebietsgrenzen ist die Klassifizierungswahrscheinlichkeit für die betreffende Gruppe größer als für die übrigen Gruppen. Auf den Gebietsgrenzen sind die Klassifizierungswahrscheinlichkeiten für die angrenzenden Gruppen identisch. Sie entsprechen dem kritischen Diskriminanzwert auf der Diskriminanzachse. Für die Lineare Diskriminanzanalyse sind die Gebietsgrenzen immer linear.

Prüfung der Annahmen

Box's M

Die obigen Klassifizierungsergebnisse wie auch die Klassifizierungsfunktionen basieren auf der Annahme gleicher Streuungen der Merkmalsvariablen in den Gruppen. Durch Auswahl der Option „Box' M“ im Dialogfenster „Statistik“ (vgl. Abbildung 4.40) ist es möglich, einen Test auf Gleichheit Kovarianz-Matrizen durchzuführen. Das Ergebnis dieses Tests ist in Abbildung 4.41 zu sehen. Als Maß der Streuung einer Gruppe wird die logarithmierte Determinante der Kovarianzmatrix der 10 Merkmalsvariablen verwendet. Man ersieht daraus, dass die Streuung der zweiten Gruppe (die auch die meisten Fälle enthält) bedeutend größer ist, als die der beiden anderen Gruppen. Auf diesen Werten basiert die Berechnung von Box's M, das sich mit einem F-Test auf Signifikanz prüfen lässt. Der F-Wert ist hier so groß und der zugehörige p-Wert (Signifikanz) mit $p = 0{,}004$ so klein, dass die Annahme gleicher Streuungen nicht aufrechterhalten werden kann.

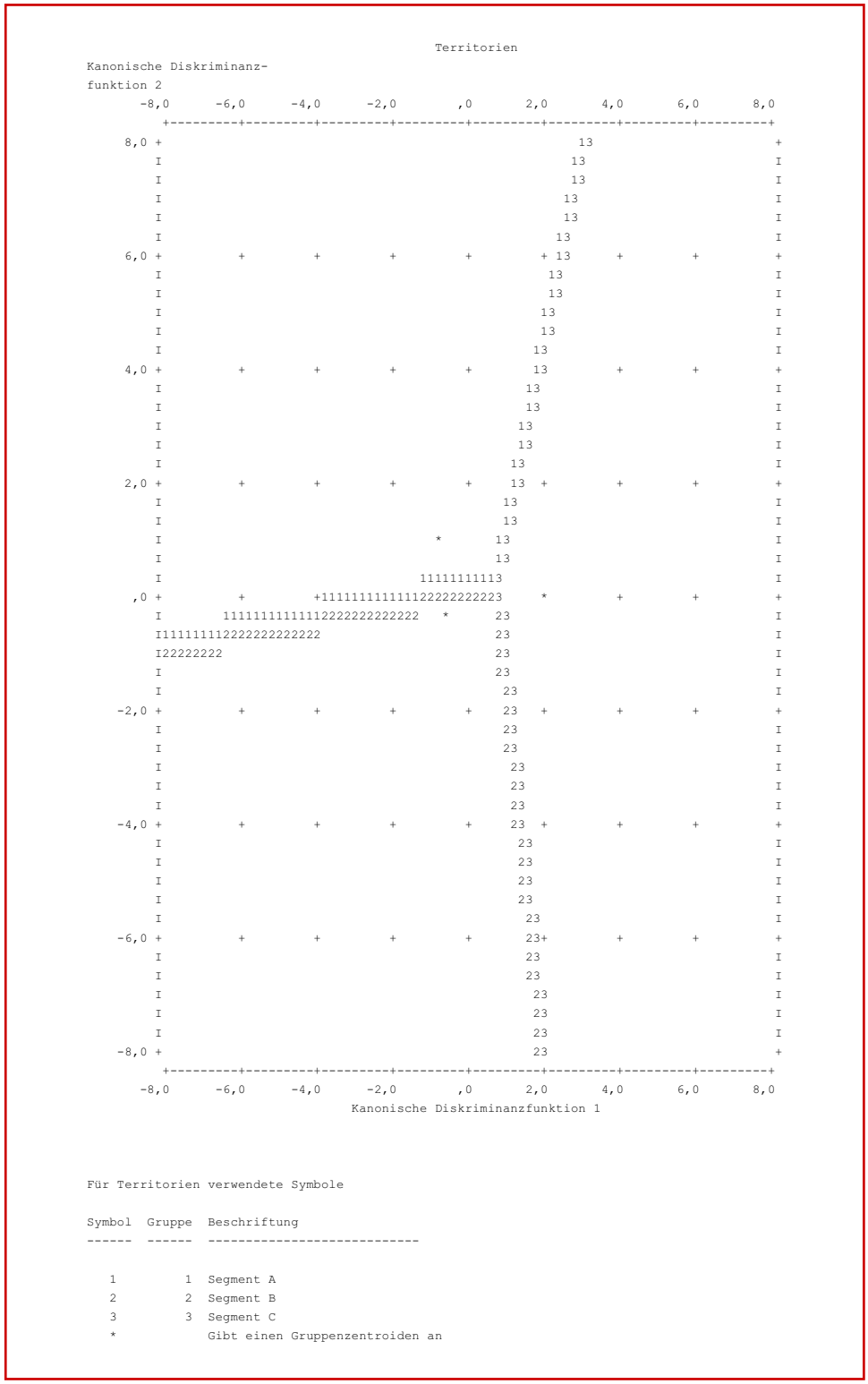


Abbildung 4.39: Klassifizierungsdiagramm (Gebietskarte der Gruppen)

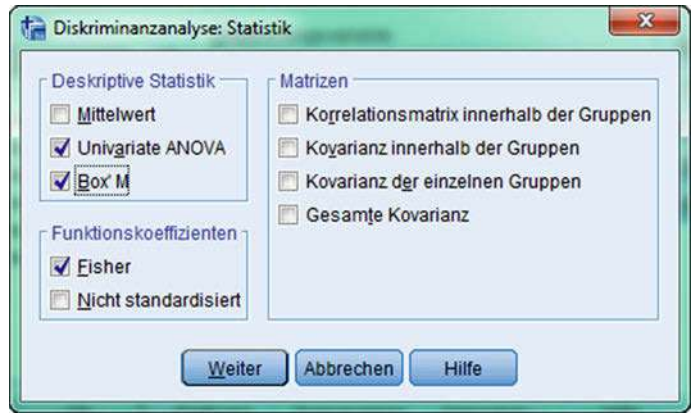


Abbildung 4.40: Dialogfenster „Statistik“

Log-Determinanten

segment	Rang	Log-Determinante
Segment A	10	2,051
Segment B	10	4,600
Segment C	10	2,771
Gemeinsam innerhalb der Gruppen	10	5,822

Die Ränge und natürlichen Logarithmen der ausgegebenen Determinanten sind die der Gruppen-Kovarianz-Matrizen.

Textergebnisse

Box-M		192,994
F	Näherungswert	1,391
	df1	110
	df2	8657,069
	Signifikanz	,004

Testet die Null-Hypothese der Kovarianz-Matrizen gleicher Grundgesamtheit.

Abbildung 4.41: Box-M-Test auf Gleichheit der Kovarianzmatrizen

Falls ungleiche Kovarianz-Matrizen vorliegen, kann dies durch Anwendung einer Quadratischen Diskriminanzanalyse (QDA) berücksichtigt werden.[29] SPSS enthält zwar keine QDA, ermöglicht aber dennoch die Berücksichtigung unterschiedlicher Kovarianz-Matrizen. Siehe dazu die Ausführungen im Anhang C. Zur Durchführung einer Diskriminanzanalyse unter Berücksichtigung ungleicher Streuungen ist im Dialogfenster „Klassifizieren“ (Abbildung 4.33 bzw. 4.37) die Option „Gruppen-spezifisch“

[29] Zur Quadratischen Diskriminanzanalyse siehe z.B. Schlittgen (2009), S. 359; Hastie/Tibshirani/Friedman (2009), S. 110.

Klassifizierungsergebnisse[a]

			Vorhergesagte Gruppenzugehörigkeit			
		segment	Segment A	Segment B	Segment C	Gesamt
Original	Anzahl	Segment A	13	5	1	19
		Segment B	7	38	6	51
		Segment C	3	0	19	22
	%	Segment A	68,4	26,3	5,3	100,0
		Segment B	13,7	74,5	11,8	100,0
		Segment C	13,6	,0	86,4	100,0

a. 76,1% der ursprünglich gruppierten Fälle wurden korrekt klassifiziert.

Abbildung 4.42: Klassifikationsmatrix bei Verwendung gruppenspezifischer Kovarianzmatrizen

zu wählen. Die sich ergebende Klassifikationsmatrix zeigt Abbildung 4.42 und das Klassifizierungsdiagramm zeigt Abbildung 4.43.

Ein der Vergleich der Klassifikationsmatrix mit Abbildung 4.34 zeigt, dass jetzt 70 der 92 Fälle richtig zugeordnet werden (also ein Fall mehr) und sich die Trefferquote auf 76,1 % erhöht. Die Klassifizierungsdiagramme sehen recht ähnlich aus, aber die Gebietsgrenzen sind jetzt nicht mehr linear. Insgesamt sind die Änderungen recht gering und zeigen, dass die LDA, wie schon eingangs bemerkt wurde, relativ unempfindlich gegenüber Verletzungen der Annahmen ist.

Man beachte, dass die Klassifizierungsfunktionen immer auf Basis der vereinten (gepoolten) Kovarianzmatrizen der Merkmalsvariablen berechnet werden (vgl. mathematischer Anhang D) und sich somit, im Gegensatz zu den Klassifizierungswahrscheinlichkeiten, nicht verändern. Die Klassifizierungsfunktionen erlauben daher keine Berücksichtigung ungleicher Streuungen der Gruppen.

4.3.3 Schrittweise Diskriminanzanalyse

Wir hatten oben unterstellt, dass zunächst alle vorhandenen Merkmalsvariablen in die Diskriminanzfunktion einbezogen werden und hatten gezeigt, wie sich durch Berechnung der standardisierten Diskriminanzkoeffizienten unwichtige Merkmalsvariablen erkennen lassen, die sodann aus der Diskriminanzfunktion eliminiert werden können. Insbesondere bei mehrfachen Diskriminanzfunktionen kann diese Vorgehensweise mühevoll sein.

Schrittweise Diskriminanzanalyse

Eine alternative und weit bequemere Vorgehensweise bietet die schrittweise Diskriminanzanalyse, bei der die Merkmalsvariablen einzeln nacheinander in die Diskriminanzfunktion einbezogen werden. Dabei wird jeweils diejenige Variable ausgewählt, die ein bestimmtes Gütemaß maximiert. Es wird also zunächst eine Diskriminanzanalyse mit einer Merkmalsvariablen, dann mit zwei Merkmalsvariablen und so fort durchgeführt.

Wird Wilks' Lambda als Gütemaß verwendet, so wird dieses, da es sich um ein inverses Gütemaß handelt, minimiert. Bei mehr als zwei Gruppen kommt hier das multivariate Wilks' Lambda gemäß Formel (4.19) zur Anwendung.

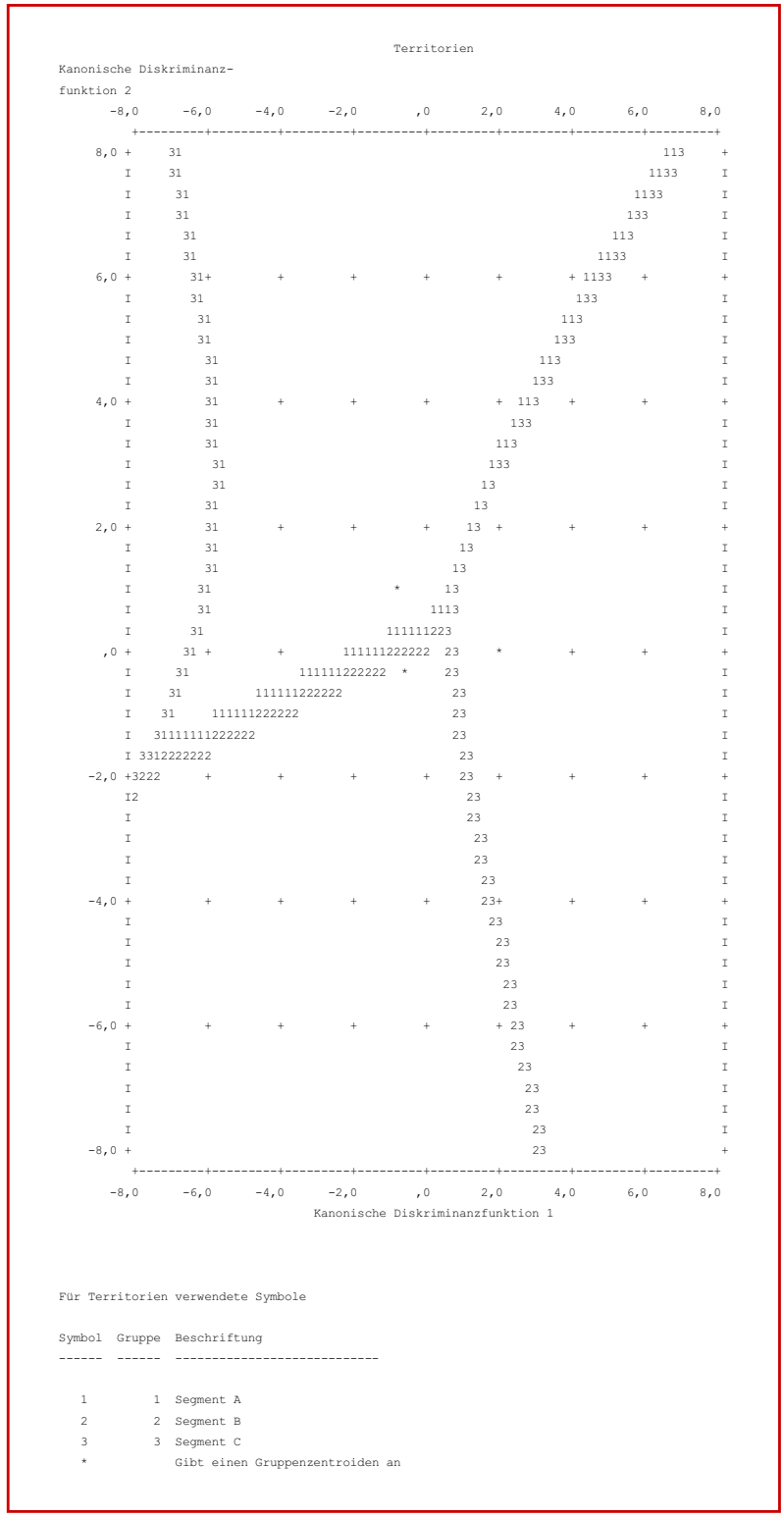


Abbildung 4.43: Klassifizierungsdiagramm bei Verwendung gruppenspezifischer Kovarianzmatrizen

Bei Anwendung einer schrittweisen Diskriminanzanalyse werden nur Merkmalsvariablen in die Diskriminanzfunktion aufgenommen, die signifikant zur Verbesserung der Diskriminanz beitragen, wobei das Signifikanzniveau durch den Anwender vorgegeben werden kann. Der Algorithmus wählt dann automatisch aus der Menge der Merkmalsvariablen die wichtigsten aus. Aus der Rangfolge, mit der die Variablen in die Diskriminanzfunktion(en) aufgenommen werden, lässt sich deren relative Wichtigkeit erkennen.

Die prinzipielle Vorgehensweise der schrittweisen Diskriminanzanalyse ist identisch mit der schrittweisen Regressionsanalyse. Dort wurden auch Vorbehalte gegen die unkritische Anwendung dieser Methode geäußert.

4.3.4 SPSS-Kommandos

Abbildung 4.44 zeigt abschließend die Syntaxdatei mit den SPSS-Kommandos für das Fallbeispiel. Syntaxdatei

```
* MVA: Fallbeispiel Diskriminanzanalyse.
* DATENDEFINITION.
DATA LIST FIXED
  /Streichf 8 Preis 10  Haltbark 12  Ungefett 14
   Backeign 16  Geschmac 18 Kalorien  20  Tierfett 22
   Vitamin 24  Natur 26  Person 27-29 Marke 30-32.

* DATENMODIFIKATION
* Definition der Segmente (Gruppen):
*   A: Du darfst, Becel
*   B: Sanella, Homa, SB, Botteram, Flora Soft, Rama
*   C: Delicado, Hollaendische Butter, Weihnachtsbutter.

COMPUTE Segment = Marke.
RECODE  SEGMENT (7,8=1) (1,2,3,9,10,11=2) (4,5,6=3).
VALUE LABELS Segment 1 "Segment A"
                    2 "Segment B"
                    3 "Segment C".

BEGIN DATA
  1   3 3 5 4 1 2 3 1 3 4  1  1
  2   6 6 5 2 2 5 2 1 6 7  3  1
  3   2 3 3 3 2 3 5 1 3 2  4  1
.....................................
 127   5 4 4 1 4 4 1 1 1 4 18 11
END DATA.

* PROZEDUR.
* Diskriminanzanalyse für den Margarinemarkt.
DISCRIMINANT
  /GROUPS = Segment (1,3)
  /VARIABLES  = Streichf TO Natur
  /ANALYSIS   =  ALL
  /METHOD     = DIRECT
  /PRIORS     = EQUAL
  /SAVE       CLASS = predict SCORES = score PROBS = prob
  /CLASSIFY   = NONMISSING
  /STATISTICS = MEAN STDDEV UNIVF BOXM RAW COEFF TABLE
  /PLOT       = COMBINED MAP.

* Anlayse mit gruppensezifischen Kovarianzmatrizen.
DISCRIMINANT
  /GROUPS=Segment(1 3)
  /VARIABLES=Streichf Preis TO Natur
  /ANALYSIS ALL
  /CLASSIFY = SEPARATE
  /STATISTICS = BOXM TABLE
  /PLOT       = COMBINED MAP.
```

Abbildung 4.44: SPSS-Kommandodatei für das Fallbeispiel

4.4 Anwendungsempfehlungen

Empfehlungen

Nachfolgend seien einige Empfehlungen für die Durchführung einer Diskriminanzanalyse zusammengestellt. Dabei werden auch Hinweise für die Handhabung von SPSS gegeben.

Erhebung der Daten und Formulierung der Diskriminanzfunktion

- Die Stichprobe darf keine Elemente enthalten, die gleichzeitig zu mehr als nur einer Gruppe gehören (z. B. Person mit zwei Berufen).
- Der Umfang der Stichprobe sollte wenigstens doppelt so groß sein wie die Anzahl der Merkmalsvariablen.
- Die Anzahl der Merkmalsvariablen sollte größer sein als die Anzahl der Gruppen.

Schätzung der Diskriminanzfunktion mit SPSS

- Zunächst sollte die Schätzung (Optimierung) nach dem Kriterium WILKS erfolgen, entweder blockweise (METHOD = DIRECT) oder schrittweise (METHOD = WILKS).
- Wenn Unsicherheit bezüglich der auszuwählenden Merkmalsvariablen besteht, sollte das Kriterium RAO angewendet werden.
- Soll insbesondere eine Unterscheidung der am schlechtesten trennbaren Gruppen erreicht werden, so sind die Kriterien MAHAL, MAXMINF oder MINRESID anzuwenden.
- Graphische Darstellungen erleichtern die Interpretation und können somit vor Fehlurteilen schützen. Eine Beschränkung auf zwei Diskriminanzfunktionen ist daher im Mehr-Gruppen-Fall von Vorteil.

Klassifizierung

- Die Gleichheit der Gruppenstreuungen ist zu prüfen. Gegebenenfalls sind die individuellen Gruppenstreuungen zu berücksichtigen. Es entfällt damit die Anwendbarkeit von Klassifizierungsfunktionen.
- Im Mehr-Gruppen-Fall sollten nicht alle mathematisch möglichen, sondern nur die signifikanten bzw. wichtigsten Diskriminanzfunktionen für die Klassifizierung verwendet werden.
- Bei ungleichen Kosten einer Fehlklassifikation muss die Klassifizierung auf Basis des Wahrscheinlichkeitskonzeptes vorgenommen werden.

4.5 Mathematischer Anhang

A. Schätzung der Diskriminanzfunktion

Ergänzend zum Text wird nachfolgend die Methode zur Schätzung der Diskriminanzfunktion näher erläutert.

Anstelle der gesuchten normierten Diskriminanzfunktion (4.1) wird zunächst eine *nicht-normierte Diskriminanzfunktion* der Form

$$Y = v_1 X_1 + v_2 X_2 + \ldots + v_J X_J \tag{A1}$$

ermittelt. Die Koeffizienten v_j seien proportional zu den Koeffizienten b_j und damit ebenfalls optimal im Sinne des Diskriminanzkriteriums. Nach Einsetzen von (A1) in das Diskriminanzkriterium gemäß Formel (4.6)

$$\Gamma = \frac{\sum_{g=1}^{G} I_g \left(\bar{Y}_g - \bar{Y}\right)^2}{\sum_{g=1}^{G} \sum_{i=1}^{I_g} \left(Y_{gi} - \bar{Y}_g\right)^2}$$

erhält man in Matrizenschreibweise folgenden Ausdruck:

$$\Gamma = \frac{\mathbf{v'Bv}}{\mathbf{v'Wv}} \tag{A2}$$

mit

$\mathbf{v}$ = Spaltenvektor der nicht-normierten Diskriminanzkoeffizienten v_j $(j = 1, \ldots, J)$
$\mathbf{B}$ = (JxJ)-Matrix für die Streuung der J Merkmalsvariablen *zwischen den Gruppen*
$\mathbf{W}$ = (JxJ)-Matrix für die Streuung der J Merkmalsvariablen *in den Gruppen*

Die Matrixelemente von $\mathbf{B}$ und $\mathbf{W}$ lauten:

$$B_{jr} = \sum_{g=1}^{G} I_g \left(\bar{X}_{jg} - \bar{X}_j\right)\left(\bar{X}_{rg} - \bar{X}_r\right) \tag{A3}$$

$$W_{jr} = \sum_{g=1}^{G} \sum_{i=1}^{I_g} \left(X_{jgi} - \bar{X}_{jg}\right)\left(X_{rgi} - \bar{X}_{rg}\right) \tag{A4}$$

mit

X_{jgi} = Merkmalsausprägung von Element i in Gruppe g bezüglich Merkmalsvariable j $(j, r = 1, \ldots, J)$
$\bar{X}_{jg}$ = Mittelwert von Variable j in Gruppe g
I_g = Fallzahl in Gruppe g
G = Anzahl der Gruppen

Die Maximierung von Γ mittels vektorieller Differentiation nach v liefert für den Maximalwert γ von Γ die folgende Bedingung:

$$\frac{\delta\Gamma}{\delta\nu} = \frac{2\left[(Bv)(v'Wv) - (v'Bv)(Wv)\right]}{(v'Wv)^2} = 0 \tag{A5}$$

Dabei ist durch $\mathbf{0}$ ein Null-Vektor bezeichnet. Nach Division von Zähler und Nenner durch ($\mathbf{v'Wv}$) und unter Verwendung der Definition (A2) für γ erhält man

$$\frac{2\left[\mathbf{Bv} - \gamma\mathbf{Wv}\right]}{\mathbf{v'Wv}} = \mathbf{0} \tag{A6}$$

Dieser Ausdruck lässt sich umformen in

$$(\mathbf{B} - \gamma\mathbf{W})\,\mathbf{v} = \mathbf{0} \tag{A7}$$

Falls $\mathbf{W}$ regulär ist (Rang J besitzt) und sich somit invertieren lässt, kann man (A7) weiter umformen in

$$(\mathbf{A} - \gamma\mathbf{E})\,\mathbf{v} = \mathbf{0} \qquad \text{mit} \quad \mathbf{A} = \mathbf{W}^{-1}\mathbf{B} \tag{A8}$$

wobei durch $\mathbf{E}$ die Einheitsmatrix bezeichnet ist. Die Lösung von (A8) bildet ein klassisches *Eigenwertproblem*. Zu finden ist der größte Eigenwert γ der Matrix A. Der gesuchte Vektor $\mathbf{v}$ ist somit ein zugehöriger Eigenvektor.

Die gesuchten Diskriminanzkoeffizienten sollen die *Normierungsbedingung*

$$\frac{1}{I-G}\mathbf{b}'\mathbf{W}\mathbf{b} = 1 \qquad \text{mit} \quad I = I_1 + I_2 + \ldots + I_G \tag{A9}$$

Eigenwertproblem

erfüllen, d. h. die „gepoolte“ (vereinte) Innergruppen-Varianz der Diskriminanzwerte (pooled within-groups variance) in der Stichprobe vom Umfang I soll den Wert Eins erhalten. Die *normierten Diskriminanzkoeffizienten* erhält man somit durch folgende Transformation:

$$\mathbf{b} = \mathbf{v}\frac{1}{s} \qquad \text{mit} \quad s^2 = \frac{1}{I-G}\mathbf{v}'\mathbf{W}\mathbf{v} \tag{A10}$$

Dabei ist s die gepoolte Innergruppen-Standardabweichung der Diskriminanzwerte, die man mit den nicht-normierten Diskriminanzkoeffizienten v erhalten würde. Mit Hilfe der normierten Diskriminanzkoeffizienten wird sodann das konstante Glied der Diskriminanzfunktion wie folgt berechnet:

$$b_0 = -\sum_{j=1}^{J} b_j \bar{X}_j \tag{A11}$$

Diskriminanz-funktionen

Weitere Diskriminanzfunktionen lassen sich in analoger Weise ermitteln, indem man den jeweils nächstgrößten Eigenwert aufsucht. Jede so ermittelte Diskriminanzfunktion ist orthogonal zu den vorher ermittelten Funktionen und erklärt einen Teil der jeweils verbleibenden Reststreuung in den Gruppen. Das Rechenverfahren der Diskriminanzanalyse beinhaltet somit eine Hauptkomponentenanalyse der Matrix $\mathbf{A}$. Die Anzahl der positiven Eigenwerte und damit der möglichen Diskriminanzfunktionen kann nicht größer sein als $\min\{G-1, J\}$.

Beispiel: Als Beispiel dienen die Daten in Abbildung 4.5 für zwei Gruppen und zwei Variablen. Bei zwei Merkmalsvariablen umfassen die Matrizen $\mathbf{B}$ und $\mathbf{W}$ in (A2) nur jeweils vier Elemente. Mit den Werten aus den Abbildungen 4.8 und 4.10 erhält man

$$\mathbf{B} = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} = \begin{bmatrix} 13,5 & -4,5 \\ -4,5 & 1,5 \end{bmatrix}$$

und

$$\mathbf{W} = \begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix} = \begin{bmatrix} 29 & 21 \\ 21 & 49 \end{bmatrix}$$

Die Inversion von $\mathbf{W}$ ergibt:

$$\mathbf{W^{-1}} = \begin{bmatrix} 0,05 & -0,02143 \\ -0,02143 & 0,02959 \end{bmatrix}$$

und die Multiplikation der Inversen mit $\mathbf{B}$ liefert die Matrix

$$\mathbf{A} = \mathbf{W^{-1}B} = \begin{bmatrix} 0,77143 & -0,25714 \\ -0,42245 & 0,14082 \end{bmatrix}$$

Durch Nullsetzen der Determinante

$$\begin{vmatrix} 0,77143 - \gamma & -0,25714 \\ -0,42245 & 0,14082 - \gamma \end{vmatrix} \det$$

erhält man schließlich die quadratische Gleichung

$$\gamma^2 - \gamma \cdot 0,91225 + 0 = 0$$

deren Nullstelle $\gamma = 0,91225$ der gesuchte Eigenwert der Matrix $\mathbf{A}$ ist (im Zwei-Gruppen-Fall existiert nur eine von 0 verschiedene Nullstelle).

Nach Subtraktion des Eigenwertes von den Diagonalelementen in $\mathbf{A}$ ergibt sich die reduzierte Matrix

$$\mathbf{R} = \mathbf{A} - \gamma\mathbf{E} = \begin{bmatrix} -0,14082 & -0,25714 \\ -0,42245 & -0,77143 \end{bmatrix}$$

Der zugehörige Eigenvektor v lässt sich durch Lösung des Gleichungssystems

$$\mathbf{Rv} = \mathbf{0}$$

finden. Da die Zeilen der Matrix $\mathbf{R}$ proportional zueinander sind (sonst wäre das Gleichungssystem nicht lösbar), lässt sich unschwer erkennen, dass die beiden folgenden Vektoren Lösungsvektoren sind:

$$v = \begin{bmatrix} 0,77143 \\ -0,42245 \end{bmatrix} \quad \text{oder} \quad \begin{bmatrix} -0,25714 \\ 0,14082 \end{bmatrix}$$

Man erhält sie, indem man die Diagonalelemente von $\mathbf{R}$ vertauscht und ihre Vorzeichen ändert. Natürlich ist auch jede proportionale Transformation dieser Vektoren ein zulässiger Lösungsvektor.

Nicht-normierte Diskriminanzfunktion

Wählt man die Elemente des ersten Vektors als Diskriminanzkoeffizienten, so erhält man damit die *nicht-normierte Diskriminanzfunktion*

$$Y = 0,77143X_1 - 0,42245X_2$$

Unter Anwendung von (A10) erhält man den *Normierungsfaktor*

$$\frac{1}{s} = 1,33656$$

und nach Multiplikation mit $\mathbf{v}$ erhält man den Vektor der *normierten Diskriminanzkoeffizienten*

$$\mathbf{b} = \begin{bmatrix} 1,03106 \\ -0,56463 \end{bmatrix}$$

Formel (A11) liefert damit für das konstante Glied

$$b_0 = -(1,03106 \cdot 4,25 - 0,56463 \cdot 4,25) = -1,9823$$

Normierte Diskriminanzfunktion

Die *normierte Diskriminanzfunktion* lautet somit:

$$Y = -1,9823 + 1,03106X_1 - 0,56463X_2$$

Mit ihr lassen sich die Diskriminanzwerte in Abbildung 4.11 berechnen.

Die Koeffizienten der normierten Diskriminanzfunktion werden zur Unterscheidung von den standardisierten Diskriminanzkoeffizienten gemäß (4.21) auch als nichtstandardisierte Diskriminanzkoeffizienten bezeichnet (z.B. in SPSS). Eine standardisierte Diskriminanzfunktion existiert dagegen i. d. R. nicht, es sei denn, dass jede Merkmalsvariable bereits so standardisiert wäre, dass ihr Gesamtmittel Null und ihre gepoolte Innergruppen-Varianz Eins ist. In diesem Falle wären die Koeffizienten der normierten Diskriminanzfunktion gleichzeitig standardisierte Diskriminanzkoeffizienten.

B. Berechnung von Distanzen

Centroid

Auf Basis von J *Merkmalsvariablen* X_j lässt sich die Mahalanobis-Distanz (verallgemeinerte Distanz) zwischen einem Element i und dem Centroid der Gruppe g wie folgt berechnen:

$$D_{ig}^2 = \left(\mathbf{X_i} - \bar{\mathbf{X}}_\mathbf{g}\right)' \mathbf{C^{-1}} \left(\mathbf{X_i} - \bar{\mathbf{X}}_\mathbf{g}\right) \tag{B1}$$

mit

$$\mathbf{X}_i' = [X_{1i}, X_{2i}, \ldots, X_{Ji}]$$
$$\bar{\mathbf{X}}'g = \left[\bar{X}_{1g}, \bar{X}_{2g}, \ldots, \bar{X}_{Jg}\right]$$

und

$$\mathbf{C} = \frac{\mathbf{W}}{\mathbf{I} - \mathbf{G}} \quad \text{(Kovarianzmatrix)} \tag{B2}$$

$\mathbf{C}$ ist die gepoolte Innergruppen-Kovarianzmatrix der Merkmalsvariablen, die man aus der Streuungsmatrix $\mathbf{W}$ gemäß (A4) nach Division durch die Anzahl der Freiheitsgrade erhält.

Die Kovarianzmatrix der Diskriminanzvariablen bildet unter der Annahme gleicher Streuungen eine Einheitsmatrix $\mathbf{E}$. Die Berechnung der Mahalanobis-Distanz auf Basis von K Diskriminanzvariablen Y_k vereinfacht sich daher wie folgt:

$$D_{ig}^2 = \left(\mathbf{Y_i} - \bar{\mathbf{Y}}_\mathbf{g}\right)' \mathbf{E} \left(\mathbf{Y_i} - \bar{\mathbf{Y}}_\mathbf{g}\right) = \sum_{k=1}^{K} \left(Y_{ki} - \bar{Y}_{kg}\right)^2 \tag{B3}$$

Bei Berücksichtigung ungleicher Streuungen in den Gruppen ist das folgende modifizierte Distanzmaß zu berechnen:

$$Q_{ig}^2 = \left(\mathbf{Y_i} - \bar{\mathbf{Y}}_\mathbf{g}\right)' \mathbf{C_g^{-1}} \left(\mathbf{Y_i} - \bar{\mathbf{Y}}_\mathbf{g}\right) + \ln|\mathbf{C_g}| \tag{B4}$$

mit

$\mathbf{C_g}$ = Kovarianzmatrix der Diskriminanzvariablen in Gruppe g
$|\mathbf{C_g}|$ = Determinante der Kovarianzmatrix

Diese Distanzen können entweder direkt zur Klassifizierung (nach minimaler Distanz) oder zur Berechnung von Klassifizierungswahrscheinlichkeiten verwendet werden (vgl. hierzu Tatsuoka (1988), S. 350 ff.).

C. Berechnung von Klassifizierungswahrscheinlichkeiten

Unter Bezugnahme auf den zentralen Grenzwertsatz der Statistik lässt sich unterstellen, dass die Diskriminanzwerte und damit die Distanzen der Elemente einer Gruppe g vom Centroid dieser Gruppe normalverteilt sind. Damit lässt sich für ein Element i mit Diskriminanzwert Y_i unter der der Hypothese „Element i gehört zu Gruppe g“ die folgende *Dichte* angeben:

$$f(Y_i|g) = \frac{1}{\sqrt{2\pi}s_g} e^{-\left(D_{ig}^2\right)/2s_g^2} \tag{C1}$$

Zentraler Grenzwertsatz

mit

$$D_{ig}^2 = \left(Y_i - \bar{Y}_g\right)^2$$

Besitzen alle Gruppen gleiche Streuung, so gilt infolge der Normierung der Diskriminanzfunktion für deren Standardabweichungen:

Gleiche Gruppenstreuung

$$s_g = 1 \qquad (g = 1, \ldots, G)$$

Die obige Dichtefunktion vereinfacht sich damit zu:

$$f(Y_i|g) = \frac{1}{\sqrt{2\pi}} e^{-(D_{ig})^2/2} \tag{C2}$$

Die Verwendung einer stetigen Verteilung der Diskriminanzwerte erfordert, dass die übliche diskrete Formulierung des Bayes-Theorems gemäß (4.28) zwecks Berechnung von Klassifizierungswahrscheinlichkeiten modifiziert wird (vgl. hierzu Tatsuoka (1988), S. 358 ff.). Setzt man anstelle der bedingten Wahrscheinlichkeiten $P(Y_i|g)$ die Dichten $f(Y_i|g)$ gemäß (C2) unter Weglassung des konstanten Terms

Bayes-Theorem

$$1/\sqrt{2\pi}$$

in die Bayes-Formel ein, so erhält man anstelle von (4.28) die folgende Formel zur Berechnung der *Klassifizierungswahrscheinlichkeiten*:

$$P(g|Y_i) = \frac{\exp\left(-D_{ig}^2/2\right) P_i(g)}{\sum_{g=1}^{G} \exp\left(-D_{ig}^2/2\right) P_i(g)} \qquad (g = 1, \ldots, G) \tag{C3}$$

Für die Anwendung dieser Formel macht es keinen Unterschied, ob die Klassifizierung auf Basis einer oder mehrerer Diskriminanzfunktionen erfolgen soll. Im zweiten Fall bilden die Diskriminanzwerte und Centroide jeweils Vektoren und die Distanzen sind gemäß (4.27) bzw. (B3) zu berechnen.

Ungleiche Gruppenstreuung

Bei wesentlich *unterschiedlicher Streuung* in den Gruppen kann die vereinfachte Dichtefunktion gemäß (C2) nicht länger verwendet werden, sondern es muss auf die Formel (C1) zurückgegriffen werden. Zwecks Vereinfachung der Berechnung lässt sich (C1) umformen in

$$f(Y_i|g) = \frac{1}{\sqrt{2\pi}} e^{-Q_{ig}^2/2} \tag{C4}$$

mit

$$Q_{ig}^2 = \frac{\left(Y_i - \bar{Y}_g\right)^2}{s_g^2} + \ln s_g \tag{C5}$$

Es sind also unter Berücksichtigung der individuellen Streuung der Gruppen *modifizierte Distanzen* zu berechnen.

Zur Berechnung der Klassifizierungswahrscheinlichkeiten ist damit die folgende Formel anzuwenden:

$$P(g|Y_i) = \frac{f\left(Y_{ig}|g\right) P_i(g)}{\sum_{g=1}^{G} f\left(Y_{ig}|g\right) P_i(g)} \qquad (g = 1, \ldots, G) \tag{C6}$$

Bei mehreren Diskriminanzfunktionen ist anstelle von (C5) die Formel (B4) anzuwenden.

Modifizierte Distanzen

Für das *Beispiel* sind in Abbildung 4.11 die folgenden empirischen Varianzen der Diskriminanzwerte in den beiden Gruppen angegeben:

$$s_A = 1,079 \qquad \text{und} \qquad s_B = 0,915$$

Man erhält damit die folgenden Klassifizierungswahrscheinlichkeiten

$$P(A|Y_i) = 0,381$$
$$P(B|Y_i) = 0,619$$

Diese unterscheiden sich hier nur geringfügig von den in Abschnitt 4.2.6.4 unter der Annahme gleicher Streuungen berechneten Klassifizierungswahrscheinlichkeiten. In kritischen Fällen aber sollte stets untersucht werden, ob sich durch Berücksichtigung der individuellen Streuungen das Ergebnis der Klassifizierung verändert.

D. Berechnung von Klassifizierungsfunktionen

Die Koeffizienten der Klassifizierungsfunktionen (4.23) werden auf Basis der Merkmalsvariablen wie folgt berechnet:

$$b_{jg} = (I - G) \sum_{r=1}^{J} W_{jr}^{-1} \bar{X}_{rg} \qquad (j = 1, \ldots, J; g = 1, \ldots, G) \tag{D1}$$

wobei durch W_{jr} die Streuungsmaße der Merkmalsvariablen gemäß (A4) bezeichnet sind. Das konstante Glied der Funktion F_g berechnet sich unter Berücksichtigung der

Apriori-Wahrscheinlichkeit P_g durch:

$$b_{0g} = -\frac{1}{2}\sum_{j=1}^{J} b_{jg}\bar{X}_{jg} + \ln P_g \qquad (g = 1, \ldots, G) \tag{D2}$$

(Vgl. IBM Corporation (2017)). Die zur Berechnung erforderlichen Werte können dem Beispiel im Teil A dieses Anhangs entnommen werden.

Literaturhinweise

A. Basisliteratur zur Diskriminanzanalyse

Hair, J.F./Black, W.C./Babin, B.J./Anderson, R.E. (2010), Multivariate Data Analysis, 7. Auflage, Upper Saddle River (N.J.).

Herrmann, A./Homburg, C./Klarmann, M. (Hrsg.) (2008), Handbuch Marktforschung – Methoden, Anwendungen und Praxisbeispiele, 3. Auflage, Wiesbaden.

Kendall, M. (1980), Multivariate Analysis, 2. Auflage, London.

Klecka, W. (1993), Discriminant Analysis, 15. Auflage, Beverly Hills.

Morrison, D. (1990), Multivariate Statistical Methods, 3. Auflage, New York.

Schlittgen, R. (2009), Multivariate Statistik, München.

Tatsuoka, M. M. (1988), Multivariate Analysis – Techniques for Educational and Psychological Research, 2nd edition, New York.

B. Zitierte Literatur

Backhaus, K./Erichson, B./Weiber, R. (2015), Fortgeschrittene Multivariate Analyseverfahren, 3. Auflage, Berlin/Heidelberg.

Bamberg, G./Coenenberg, A. (1992), Betriebswirtschaftliche Entscheidungslehre, München.

Breiman, L./Friedman, J./Olshen, R./Stone, C. (1984), Classification and Regression Trees, New York.

Cooley, W./Lohnes, P. (1971), Multivariate Data Analysis, New York.

Fisher, R. A. (1936), The use of multiple measurement in taxonomic problems, in: *Annals of Eugenics*, Vol. 7, Nr. 2, S. 179–188.

Green, P./Tull, D./Albaum, G. (1988), Research for Marketing Decisions, 5. Auflage, Englewood Cliffs (N.J.).

Hartung, J./Elpelt, B. (2007), Multivariate Statistik, 7. Auflage, München u. a.

Hastie, T./Tibshirani, R./Friedman, J. (2009), The Elements of Statistical Learning, 2. Auflage, New York.

Häußler, W. M. (1979), Empirische Ergebnisse zu Diskriminationsverfahren bei Kreditscoringsystemen, in: *Zeitschrift für Operations Research*, Vol. 23, Nr. 8, S. 191–210.

IBM Corporation (2017), IBM SPSS Statistics Algorithms 25, ohne Ort.

Johnson, R. (1987), Adaptive Perceptual Mapping, Proceedings of the Sawtooth Software Conference on Perceptual Mapping, Chicago, S. 143–158.

Kendall, M. (1980), Multivariate Analysis, 2. Auflage, London.

Lachenbruch, P. (1975), Discriminant Analysis, London.

Lim, T./Loh, W./Shih, Y. (2000), A Comparison of Predicting Accuracy, Complexity, and Training Time of Thirty-three Old and New Classification Algorithms, in: *Machine Learning*, Vol. 40, Nr. 3, S. 203–229.

Michie, D./Spiegelhalter, D./Taylor, C. (1994), Machine Learning, Neural and Statistical Classification, Ellis Horwood Series in Artificial Intelligence, New York.

Morrison, D. (1981), On the Interpretation of Discriminant Analysis, in: Aaker, D.A., Belmont (ed.): Multivariate Analysis in Marketing: Theory and Applications, Palo Alto.

Pohar, M./Blas, ./Turk, S. (2004), Comparison of Logistic Regression and Linear Discriminant Analysis: A Simulation Study, in: *Metodoloski zvezki*, Vol. 1, Nr. 1, S. 143–161.

Schneeweiss, H. (1967), Entscheidungskriterien bei Risiko, Berlin.

Tatsuoka, M. (1988), Multivariate Analysis – Techniques for Educational and Psychological Research, 2. Auflage, New York.

5 Logistische Regression

K. Backhaus, B. Erichson, W. Plinke, R. Weiber, *Multivariate Analysemethoden*,
https://doi.org/10.1007/978-3-662-56655-8_6

5.1 Problemstellung

Bei vielen Problemstellungen in Wissenschaft und Praxis treten immer wieder die folgenden Fragen auf:

- Welcher von zwei oder mehreren alternativen Zuständen liegt vor oder welches Ereignis wird eintreffen?
- Welche Faktoren eignen sich für die Entscheidung oder Prognose und welchen Einfluss haben sie auf das Zustandekommen eines Zustandes oder Ereignisses?

Häufig geht es dabei nur um zwei alternative Zustände oder Ereignisse, z.B. hat ein Patient eine bestimmte Krankheit oder nicht? Wird er überleben oder nicht? Wird ein Kreditnehmer seinen Kredit zurückzahlen oder nicht? Wird ein potentieller Käufer ein Produkt kaufen oder nicht? In anderen Fällen geht es um mehr als zwei Alternativen, z.B. welche Marke wird ein potentieller Käufer wählen oder welcher Partei wird ein Wähler seine Stimme geben? Zur Beantwortung derartiger Fragen kann die *logistische Regression* angewendet werden.

Logistische Regression

Die logistische Regression gehört zur Klasse der *strukturen-prüfenden Verfahren*. Sie bildet, wie schon der Name erkennen lässt, eine Variante der Regressionsanalyse mit der Besonderheit, dass es sich bei der abhängigen Variablen Y um eine kategoriale Variable handelt, deren Ausprägungen $(g = 1, ..., G)$ die Alternativen (Gruppen, Response-Kategorien) repräsentieren. Da das Eintreffen von Ereignissen meist mit Unsicherheit behaftet ist, wird Y als eine Zufallsvariable betrachtet und es werden die Wahrscheinlichkeiten für die Ausprägungen von Y prognostiziert.

Für $G = 2$ Alternativen bildet Y eine binäre (dichotome) Variable und man spricht entsprechend von *binärer logistischer Regression*. Für $G \geq 3$ spricht man von *multinomialer logistischer Regression*. Im binären Fall, mit dem wir uns zunächst befassen wollen, werden die Gruppen gewöhnlich mit 0 und 1 indiziert und entsprechend ist Y dann eine 0,1-Variable. Für die Wahrscheinlichkeiten gilt:

$$P(Y = 0) = 1 - P(Y = 1)$$

und umgekehrt. Das Modell der logistischen Regression lässt sich in sehr vager Form wie folgt ausdrücken:

$$\pi(x) = f(x_1, ..., x_J) \tag{5.1}$$

Dabei bezeichnet $\pi(x) = P(Y = 1|x)$ die bedingte Wahrscheinlichkeit für das Eintreffen von Ereignis 1 (z.B. Erfolg) für gegebene Werte der unabhängigen Variablen $x_1, ..., x_J$. Die unabhängigen Variablen werden dabei linear miteinander kombiniert. Die Linearkombination

$$z(x) = \beta_0 + \beta_1 x_1 + ... + \beta_J x_J \tag{5.2}$$

wird als *systematische Komponente* des Modells bezeichnet. Das Modell ist in dieser Hinsicht identisch mit der linearen Regressionsanalyse.

Für die Ausgestaltung des Modells der logistischen Regression wird die *logistische Funktion* verwendet, woraus der Name resultiert:

Logistische Funktion

$$p = \frac{e^z}{1 + e^z} = \frac{1}{1 + e^{-z}} \tag{5.3}$$

Die logistische Funktion, die einen S-förmigen Verlauf hat, ist in Abbildung 5.1 dargestellt.[1] Sie lässt sich interpretieren als eine Verteilungsfunktion (kumulative Wahrscheinlichkeitsverteilung), die der Verteilungsfunktion der Normalverteilung sehr nahe kommt.[2] Damit kann sie verwendet werden, um eine reellwertige Variable in eine Wahrscheinlichkeit zu transformieren, also vom Wertebereich $[-\infty, +\infty]$ in den Wertebereich $[0, 1]$.

Transformiert man die *systematische Komponente* mit der *logistischen Funktion*, erhält man die folgende *logistische Regressionsfunktion*:

Logistische Regressionsfunktion

$$\pi(x) = \frac{1}{1 + e^{-z(x)}} \tag{5.4}$$

Die systematische Komponente $z(x)$ bildet einen Prädiktor für die Wahrscheinlichkeit $\pi(x)$. Je größer $z(x)$, desto größer $\pi(x)$. Entsprechend gilt: je größer $z(x)$, desto kleiner $P(Y = 0|x)$.

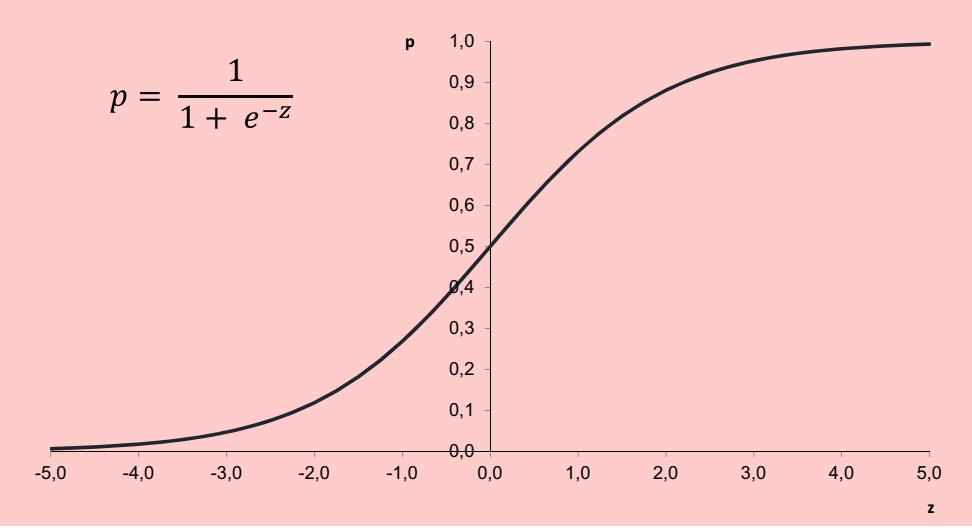


Abbildung 5.1: Logistische Funktion

In Abbildung 5.2 sind einige Anwendungsbeispiele der logistischen Regression zusammengefasst. Es sind jeweils die Anzahl der Gruppen bzw. Kategorien der abhängigen Variablen angegeben sowie die unabhängigen Variablen, die zur Schätzung der Eintrittswahrscheinlichkeiten dieser Kategorien herangezogen wurden.

Wie bei anderen Verfahren, werden auch bei der logistischen Regression in unterschiedlichem Kontext unterschiedliche Bezeichnungen für die Variablen verwendet:

- Die abhängige Variable wird auch als Y-Variable, Gruppierungsvariable oder Response-Variable bezeichnet.

[1] Die logistische Funktion wurde von dem belgischen Mathematiker Pierre-Francois Verhulst (1804–1849) zur Beschreibung und Prognose von Bevölkerungswachstum entwickelt, und zwar als verbesserte Alternative zur Exponentialfunktion. Wir hatten sie bereits im Kapitel 2 zur Zeitreihenanalyse behandelt (vgl. Abschnitt 2.3.4.2). e = 2,71828 ist die Eulersche Zahl, die auch als Basis des natürlichen Logarithmus dient.

[2] Hieraus entstammt eine breite Verwendung und Bedeutung der logistischen Funktion, da sich die Verteilungsfunktion der Normalverteilung nur als Integral ausdrücken und damit schwer berechnen lässt.

- Die unabhängigen Variablen werden auch als X-Variablen, Einflussgrößen, erklärende Variablen, Prädiktoren oder Kovariaten bezeichnet.

Teilweise wird bei den unabhängigen Variablen nur von Kovariaten gesprochen, wenn es sich um metrische Variablen handelt, und von Faktoren, wenn es sich um kategoriale Variablen handelt.[3]

Problemstellung	Abhängige Variable	Unabhängige Variable
Anbieterwechsel im Mobilfunkbereich[4]	2 Gruppen: Verbleib beim Anbieter vs. Wechsel zur Konkurrenz	4 Variablen: Nettonutzendifferenz, Amortisation spezifischer Investitionen, direkte Wechselkosten, Unsicherheitsdifferenz
Wahl der Absatzform[5]	2 Gruppen: Vertreter- vs. Handelsreisendeneinsatz	19 Variablen, u. a.: Kundenzahl je Mitarbeiter, Substituierbarkeit der Produkte, Anzahl Hotelübernachtungen, Anzahl Besuche bis Abschluss, Produktspezifische Kenntnisse
Ausbildungsadäquate Beschäftigung von Berufsanfängern mit Hochschulabschluss[6]	2 Gruppen: rund 1/2 Jahr nach Abschluss ausbildungsadäquat beschäftigt vs. inadäquat beschäftigt oder arbeitslos	15 Variablen u.a.: Geschlecht, Ausbildungsdauer (kurz/lang), Wohnstatus (Eltern: ja/nein), Fachrichtung, Berufsausbildung (ja/nein), Nebenerwerbstätigkeit (ja/nein)
Wahlverhalten von Bürgern[7]	3 Gruppen: CDU-Wähler vs. SPD-Wähler vs. Wähler anderer Parteien	Politische Einstellung, Demokratie-Zufriedenheit, Gewerkschaftsmitgliedschaft, Konfession etc.
Welche Faktoren haben Einfluss auf die Sterbewahrscheinlichkeit auf Intensivstationen?[8]	2 Gruppen: lebendig vs. verstorben	21 Variablen, u.a.: Alter, Geschlecht, Rasse, Krebserkrankung (ja/nein), chronische Nierenerkrankung (ja/nein), Blutdruck (mm HG), Pulsschlag (Schläge/min)
Einflussfaktoren auf das Geburtsgewicht von Babys[9]	2 Gruppen: normalgewichtige vs. untergewichtige Babys	Alter, Gewicht der Mutter bei der letzten Menstruation, Rasse, Anzahl der Arztbesuche in den ersten 3 Monaten der Schwangerschaft
Automatischer Filter für E-Mail-Spam (Junk Mail)[10]	2 Gruppen: E-Mail vs. Spam	57 Variablen: Häufigkeit der Verwendung bestimmter Wörter, Vorkommen bestimmter Zeichenketten

Abbildung 5.2: Anwendungsbeispiele der logistischen Regression

[3] Kategoriale unabhängige Variablen müssen, wie bei der linearen Regressionsanalyse, in binäre Variablen zerlegt werden.
[4] Vgl. Weiber/Adler (2003), S. 88 ff.
[5] Vgl. Krafft, M., 1997, S. 625 ff.
[6] Vgl. Büchel/Matiaske (1996), S. 53 ff.
[7] Vgl. Urban (1993), S. 75 ff.
[8] Vgl. Hosmer/Lemeshow/Sturdivant (2013), S. 22 ff.
[9] Vgl. ebenda, S. 25 ff.
[10] Vgl. Hastie/Tibshirani/Friedman (2009), S. 300 ff.

Vergleich mit der Diskriminanzanalyse

Die logistische Regression (LR) ähnelt hinsichtlich der Problemstellung der *Diskriminanzanalyse* (DA), die im vorhergehenden Kapitel behandelt wurde. Beide Verfahren basieren auf einem linearen Modell und sie werden daher in der Tat auch alternativ verwendet. Anstelle von Zuständen oder Ereignissen hatten wir bei der DA von separaten Gruppen gesprochen, was historisch durch die ursprünglichen Anwendungsbereiche bedingt ist. Auch die LR kann für die *Klassifizierung*, d.h. die Einordnung von Elementen in vorgegebene Gruppen, verwendet werden.

Der für den Anwender wesentliche Unterschied zwischen den beiden Verfahren besteht darin, dass die LR direkt Wahrscheinlichkeiten für das Eintreffen der alternativen Zustände oder der Zugehörigkeiten zu den einzelnen Gruppen liefert. Im Unterschied dazu liefert die DA Diskriminanzwerte, aus denen dann in einem gesonderten Schritt die Wahrscheinlichkeiten für die Gruppenzugehörigkeit berechnet werden können.

Die LR ist rechnerisch etwas aufwendiger, da die Schätzung der Parameter die Anwendung der ML-Methode (Maximum Likelihood) und somit eines iterativen Verfahrens erforderlich macht. Bezüglich der statistischen Eigenschaften der Methoden gilt als ein Vorteil der LR, dass sie auf weniger Annahmen bezüglich der verwendeten Daten basiert als die DA, und dass sie deshalb robuster in Bezug auf das Datenmaterial ist und insbesondere auch unempfindlicher gegenüber groben Ausreißern reagiert.[11] Die Erfahrung zeigt allerdings, dass beide Verfahren, insbesondere bei großen Stichproben ähnlich gute Ergebnisse liefern, auch in Fällen, wenn die Annahmen der DA nicht erfüllt sind.[12]

Anwendungsbeispiel

Nachfolgend soll die Methodik der logistischen Regression zunächst an einem kleinen Anwendungsbeispiel für den binären Fall erläutert werden.

Der Produktmanager eines Nahrungsmittelherstellers möchte untersuchen, ob und wie die Nachfrage für eine neuartige Gourmet-Butter, die preislich im Top-Segment positioniert werden soll, vom Einkommen der Konsumenten abhängt und ob sie eher von Frauen oder Männern präferiert wird. Er führt zu diesem Zweck einen Produkttest durch, in dem die Testpersonen nach Vorstellung und Verköstigung des Produktes gefragt werden, ob sie die neue Buttersorte kaufen werden. Die Testpersonen konnten dabei zwischen folgenden Antwortkategorien wählen: „ja“, „vielleicht“, „eher nicht“ und „nein“. Der Einfachheit halber fassen wir die drei letzten Antworten zu einer Kategorie zusammen und bezeichnen die alternativen Ergebnisse als „Kauf“ und „Nicht-Kauf“. Abbildung 5.3 zeigt die demografischen Merkmale von 30 befragten Personen und deren Antworten.[13]

[11] Bei der DA wird unterstellt, dass die unabhängigen Variablen multivariat normalverteilt sind, während bei der LR angenommen wird, dass für die abhängige Variable eine Binomial- oder Multinomialverteilung gilt.

[12] Vgl. dazu Michie/Spiegelhalter/Taylor (1994), S. 214; Hastie/Tibshirani/Friedman (2009), S. 128; Lim/Loh/Shih (2000), S. 216.

[13] Um die Berechnungen übersichtlicher zu gestalten wurde hier das Einkommen in Einheiten von 1000 Euro angegeben.

Person	Einkommen [Tsd. Euro]	Geschlecht 0=w, 1=m	Kauf 1=ja, 0=nein
1	2,530	0	1
2	2,370	1	0
3	2,720	1	1
4	2,540	0	0
5	3,200	1	1
6	2,940	0	1
7	3,200	0	1
8	2,720	1	1
9	2,930	0	1
10	2,370	0	0
11	2,240	1	1
12	1,910	1	1
13	2,120	0	1
14	1,830	1	1
15	1,920	1	1
16	2,010	0	0
17	2,010	0	0
18	2,230	1	0
19	1,820	0	0
20	2,110	0	0
21	1,750	1	1
22	1,460	1	0
23	1,610	0	1
24	1,570	1	0
25	1,370	0	0
26	1,410	1	0
27	1,510	0	0
28	1,750	1	1
29	1,680	1	1
30	1,620	0	0

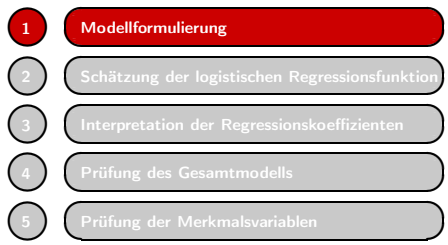
Abbildung 5.3: Daten des Anwendungsbeispiels

5.2 Vorgehensweise

Die Durchführung der logistischen Regressionsanalyse wird im Folgenden in fünf Ablaufschritte untergliedert, die in Abbildung 5.4 grafisch verdeutlicht sind. Zunächst muss aufgrund theoretischer oder sachlogischer Überlegungen die Formulierung eines Modells erfolgen. Anschließend erfolgt die Schätzung der Parameter des Modells, in diesem Fall der logistischen Regressionsfunktion. Danach erfolgt ein Abschnitt zur inhaltlichen Interpretation der gewonnenen Ergebnisse. Und schließlich erfolgt in den Schritten 4 und 5 die Güteprüfung der Schätzergebnisse, wobei wir zwischen der Güteprüfung des Gesamtmodells und der Prüfung einzelner Merkmalsvariablen (unabhängiger Variablen) differenzieren.

Abbildung 5.4: Ablaufschritte der logistischen Regressionsanalyse

5.2.1 Modellformulierung



Der Anwender muss zunächst festlegen, welche Ereignisse als mögliche Kategorien der abhängigen Variablen betrachtet werden sollen und welche Variablen *hypothetisch* als Einflussgrößen in Frage kommen und untersucht werden sollen.

Bei zahlreichen Kategorien kann es u.U. erforderlich sein, mehrere Kategorien zusammenzufassen. So hatten wir hier bereits die drei Antwortkategorien „vielleicht“, „eher nicht“ und „nein“ zu einer Kategorie „Nicht-Kauf“ zusammengefasst. Eine ähnliche Situation ergibt sich, wenn Haushalte danach eingeteilt werden, ob Kinder vorhanden sind oder nicht, ohne weiter nach der Zahl der Kinder zu differenzieren. Anders sähe es aus, wenn es um die Markenwahl zwischen Mercedes, BMW und Audi gehen würde. In diesem Fall könnte man schwerlich zwei der Kategorien zusammenfassen.

Zunächst soll hier nur das Einkommen als Einflussgröße betrachtet werden. Der Produktmanager vermutet, dass dieses einen positiven Einfluss auf das Kaufverhalten haben wird. Er formuliert daher folgendes Model:

$$Y = f(Einkommen)$$

mit $Y = 1$ für Kauf und 0 für Nicht-Kauf. Um dieses Model schätzen zu können, muss es noch näher spezifiziert werden. Dafür bestehen, wie fast immer, verschiedene Möglichkeiten.

Zu Beginn einer Analyse ist es immer zweckmäßig, sich die zu analysierenden Daten grafisch in einem Streudiagramm zu veranschaulichen, das hier in Abbildung 5.5 wiedergegeben ist. Es fällt auf, dass die Punkte auf zwei parallelen Linien angeordnet sind. Die obere Punktereihe repräsentiert die „Käufe“ und die untere die „Nicht-Käufe“. Es ist ersichtlich, dass sowohl bei niedrigem wie bei hohem Einkommen Käufe erfolgen. Allerdings sind die Käufe leicht nach rechts in Richtung „höheres Einkommen“ verschoben. Dies weist darauf hin, dass das Einkommen einen positiven Einfluss auf das Kaufverhalten hat, wie schon vom Produktmanager vermutet wurde.

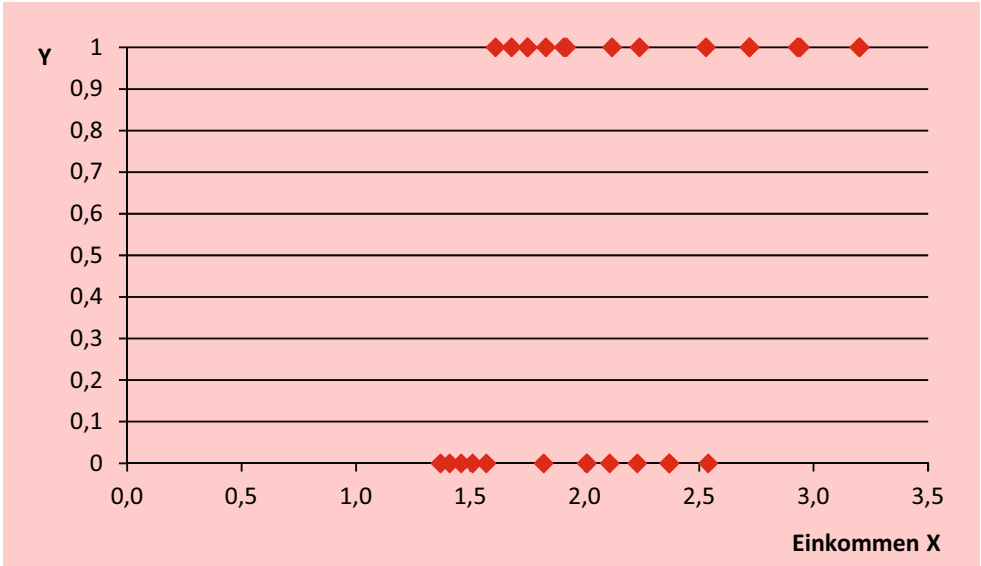


Abbildung 5.5: Streudiagramm für Kauf (Y) und Einkommen (X)

5.2.1.1 Das lineare Wahrscheinlichkeitsmodell (Modell 1)

Lineares Wahrscheinlichkeitsmodell

Das einfachste Modell zur Analyse der vorliegenden Daten ist das sog. *lineare Wahrscheinlichkeitsmodell*:[14]

$$\pi(x_k) = \alpha + \beta x_k \tag{5.5}$$

Mit $\pi(x_k)$ sind die Kaufwahrscheinlichkeiten der Testpersonen bezeichnet, die linear vom Einkommen x abhängen. Die Wahrscheinlichkeiten sind nicht beobachtbar, aber sie manifestieren sich in den Antworten auf die Kaufabfrage ($y_k = 0$ oder 1). Die Anwendung der Regressionsanalyse mit den Werten $\{y_k, x_k\}$ liefert:

$$\begin{aligned} \hat{y}_k &= a + bx_k \\ &= -0,28 + 0,39x_k \qquad\qquad R^2 = 16,6\% \end{aligned} \tag{5.6}$$

Während man für die abhängige Variable 0,1-Werte als Input eingibt, erhält man durch die Regressionsanalyse Schätzwerte $\hat{y}_k$, die jetzt metrisch skaliert sind. Mit gewissen Einschränkungen lassen sich diese Werte als Wahrscheinlichkeiten interpretieren und wir definieren daher: $p_k = \hat{y}_k$. In Abbildung 5.6 ist die geschätzte Regressionsgerade im Streudiagramm der Daten dargestellt.

Das lineare Wahrscheinlichkeitsmodell ist ein sehr grobes Modell mit strukturellen Defekten.[15] Dennoch kann es brauchbare Ergebnisse liefern. Das positive Vorzeichen des Regressionskoeffizienten b bestätigt die Vermutung des Produktmanagers, dass das Einkommen einen positiven Einfluss auf die Kaufwahrscheinlichkeit hat. Das Bestimmtheitsmaß (R-Quadrat) beträgt zwar lediglich 16,3%, was aber bei Individualdaten nicht unüblich ist.

[14] Zur Vereinfachung der Notation bezeichnen wir die Parameter mit α und β, statt mit β_0 und β_1.
[15] Da die abhängige Variable hier ein 0,1-Variable ist, kann die bei der Regressionsanalyse übliche Annahme normalverteilter Störgrößen nicht erfüllt sein.

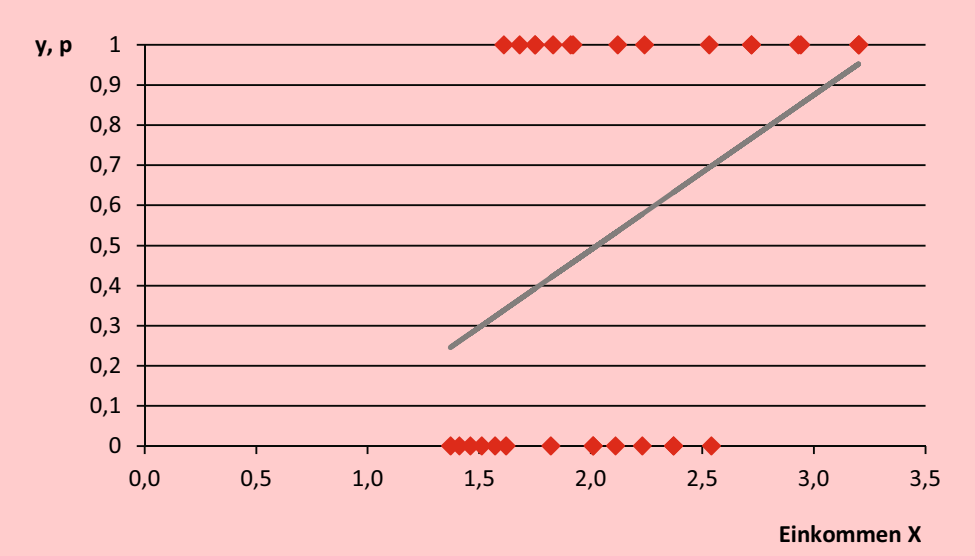


Abbildung 5.6: Geschätzte Regressionsfunktion für das lineare Wahrscheinlichkeitsmodell

Das Modell hat den Vorteil, dass es leicht zu berechnen und auch leicht zu interpretieren ist, da sich die Kaufwahrscheinlichkeiten linear mit dem Einkommen ändern. Das mittlere Einkommen liegt hier bei 2000 Euro und die erwartete Kaufwahrscheinlichkeit bei diesem Einkommen beträgt p = 0,49, wie sich mit der geschätzten Regressionsfunktion (5.6) leicht berechnen lässt. Erhöht sich das Einkommen von 2000 Euro auf 3000 Euro und damit x von 2 auf 3, so steigt die Kaufwahrscheinlichkeit um b = 0,39 auf p = 0,88 oder 88%.

Das Modell ist aber nicht logisch konsistent, denn es kann Wahrscheinlichkeiten liefern, die außerhalb des Intervalls von 0 bis 1 liegen. Für Einkommen unter 733 Euro ergeben sich negative „Wahrscheinlichkeiten“ und für Einkommen über 3.324 erhält man „Wahrscheinlichkeiten“ größer als Eins. Trotz dieser Mängel aber bietet das Modell eine recht brauchbare Approximation innerhalb des Stützbereichs, also für Einkommen von etwa 1000 bis 3000 Euro, wie wir noch zeigen werden.

5.2.1.2 Logistische Regression mit gruppierten Daten (Modell 2)

Die binäre logistische Regressionsfunktion mit einer unabhängigen Variablen lässt sich in unterschiedlicher Form ausdrücken:

$$\text{Wahrscheinlichkeit:} \qquad \pi = \frac{1}{1 + e^{-(\alpha+\beta x)}} \qquad (0 \leq \pi \leq 1) \qquad (5.7)$$

$$\text{Odds:} \qquad \frac{\pi}{1-\pi} = e^{\alpha+\beta x} \qquad (0 \leq odds \leq \infty) \qquad (5.8)$$

$$\text{Logit:} \qquad ln\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta x \qquad (-\infty \leq logit \leq \infty) \qquad (5.9)$$

Durch die Transformation der Wahrscheinlichkeit π in Odds (Chancen) wird der Wertebereich $[0, 1]$ auf den Wertebereich $[0, +\infty]$ ausgeweitet, und durch die Transformation in Logits (Log-Odds) wird er auf $[-\infty, +\infty]$ erweitert (vgl. Abbildung 5.7). In

Abschnitt 5.2.3 kommen wir auf die Odds noch zu sprechen. Momentan ist nur wichtig, dass sich mittels der Logit-Transformation von π die logistische Regressionsfunktion linearisieren lässt. Sie ist daher von zentraler Bedeutung für die logistische Regression.

$$\text{Wir definieren: } Logit(\pi) = ln\left(\frac{\pi}{1-\pi}\right) \tag{5.10}$$

Link-Funktion

Im Rahmen „*Generalisierter Linearer Modelle*“ bildet Logit(π) eine sog. *Link-Funktion*, mittels derer zwischen dem Erwartungswert einer abhängigen Variable und der systematischen Komponente des Modells eine lineare Beziehung hergestellt wird.[16]

In Abbildung 5.7 ist die Logit-Transformation grafisch dargestellt. Sie ist symmetrisch um Null und besitzt keine obere oder untere Grenze. Die Logit-Transformation ist die Umkehrfunktion der logistischen Funktion und es gilt:

$$Logit\left(\frac{1}{1+e^{-\pi}}\right) = \pi \tag{5.11}$$

Mit der Logit-Transformation lässt sich die Schätzung der logistischen Regressionsfunktion sehr vereinfachen, wenn gruppierte Daten vorliegen. Durch Gruppierung der Daten kann man für die abhängige Variable anstelle der 0,1-Daten relative Häufigkeiten berechnen, die sich als Wahrscheinlichkeiten interpretieren lassen.

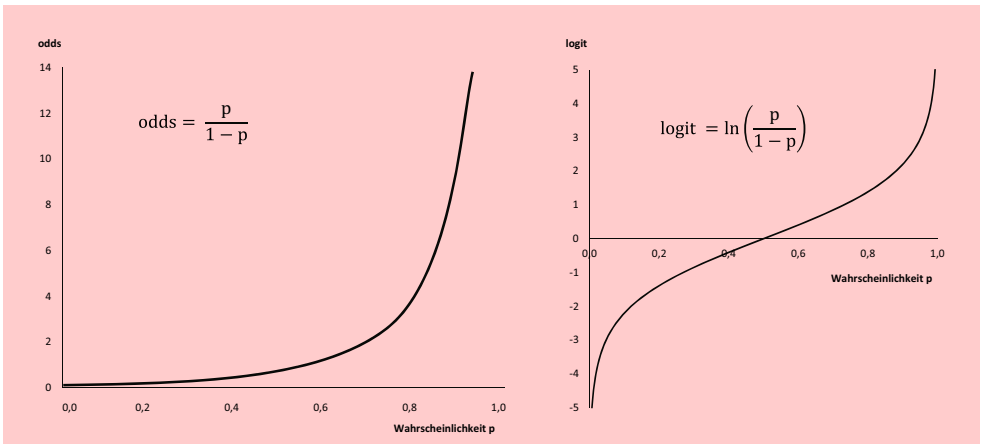


Abbildung 5.7: Odds und Logit in Abhängigkeit von der Wahrscheinlichkeit p

Im vorliegenden Fall kann man die Beobachtungen nach Einkommensklassen gruppieren. Da nur wenige Daten vorliegen, müssen wir uns hier mit nur drei Einkommensklassen begnügen, die wir als „hoch“, „mittel“ und „niedrig“ bezeichnen können. Die Daten in Abbildung 5.3 sind bereits so geordnet, dass die Personen 1-10 in die erste Einkommensklasse fallen, die Personen 11-20 in die zweite und die restlichen Personen in die dritte. Die Gruppen indizieren wir mit q. Diese Gruppen sind zu unterscheiden von den Gruppen der abhängigen Variable. Wie haben jetzt nur noch Q = 3 Beobachtungen anstatt K = 30.

[16] Dabei wird im Generalisierten Linearen Modell eine bestimmte Wahrscheinlichkeitsverteilung der abhängigen Variable unterstellt. Das Logit-Link wird insbesondere bei Vorliegen einer Binomial-Verteilung verwendet. Andere Link-Funktionen sind z.B. das Log-Link für Poisson-Verteilung, das Probit-Link für Normal-Verteilung oder das Log-Log-Link für Extremwert-Verteilung. Logit-, Probit- und Log-Log-Link haben einen sehr ähnlichen Verlauf. Vgl. dazu Agresti (2013), S. 112 ff.; Fahrmeir/Kneib/Lang (2009), S. 189 ff.; Fox (2015), S. 418 ff.

Gruppe (q)	Einkommen	Mittleres Eink. [Tsd. Euro]	Käufe	Anteil (h_q)	logit (h_q)
1	hoch	2,752	7	0,70	0,8473
2	mittel	2,020	5	0,50	0
3	niedrig	1,573	4	0,40	-0,4055

Abbildung 5.8: Nach Einkommensklassen gruppierte Daten

Abbildung 5.8 zeigt für jede Gruppe das mittlere Einkommen, die Zahl der Käufe und deren relative Häufigkeit (Anteil) in der Gruppe. Die drei sich ergebenden Datenpunkte für die relativen Häufigkeiten zeigt das Streudiagramm in Abbildung 5.9. Die logistische Regressionsfunktion laute

$$\pi(x_q) = \frac{1}{1 + e^{-(\alpha+\beta x_q)}} \tag{5.12}$$

mit q = 1 bis 3. x_q bezeichnet jetzt das mittlere Einkommen in Gruppe q. Anstelle der Wahrscheinlichkeiten π verwenden wir jetzt die relativen Häufigkeiten h. Nach Einsetzen in (5.12) und Linearisierung mit der Logit-Transformation (5.9) erhält man damit:

Logit-Transformation

$$Logit(h_q) = \alpha + \beta x_q \tag{5.13}$$

Mit y = Logit(h) und nach Einfügen eines Störterms ϵ erhalt man die folgende lineare Regressionsbeziehung:[17]

$$y_q = \alpha + \beta x_q + \epsilon \tag{5.14}$$

Deren Schätzung ergibt:

$$\hat{y}_q = -2,12 + 1,41 x_q \qquad \text{mit } R^2 = 99,6\% \tag{5.15}$$

Mit $logit(p) = \hat{y}$ erhalten wir:

$$p_k = \frac{1}{1 + e^{2,12-1,41 x_k}}$$

In die so auf aggregierter Basis geschätzte Funktion können wir individuelle Einkommenswerte einsetzen und so Schätzwerte für individuelle Wahrscheinlichkeiten erhalten. Für die erste Person mit einem Einkommen von 2.530 Euro ergibt sich:

$$p_1 = \frac{1}{1 + e^{2,12-1,41 x_q}} = \frac{1}{1 + e^{2,12-1,41 \cdot 2,53}} = 0,64$$

In Abbildung 5.9 ist die geschätzte logistische Regressionsfunktion gemeinsam mit den relativen Häufigkeiten der Käufe in den drei Einkommensgruppen dargestellt.

Da der Schätzwert b (Koeffizient von x) positiv ist, ergibt sich ein mit dem Einkommen ansteigender Verlauf, wie auch beim linearen Wahrscheinlichkeitsmodell. Je größer b, desto steiler ist der Anstieg. Im Unterschied zum linearen Wahrscheinlichkeitsmodell aber flacht der Verlauf mit der Entfernung vom mittleren Einkommen zunehmend ab. Für die Wahrscheinlichkeit können sich damit nur noch Werte zwischen 0 und 1 ergeben.

[17] Die Parameter α und β sind natürlich nicht ganz identisch in den unterschiedlichen Modellformulierungen.

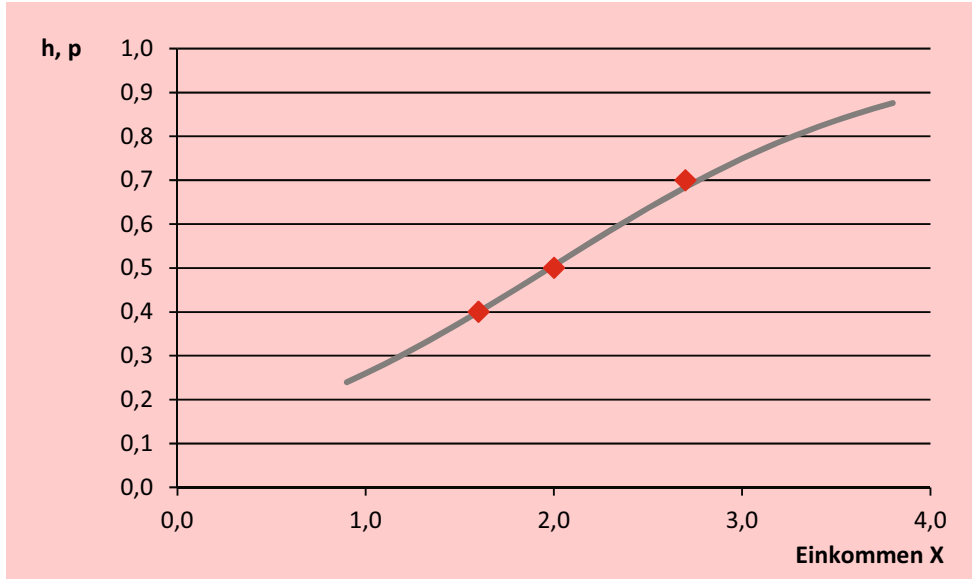


Abbildung 5.9: Logistische Regressionsfunktion für gruppierte Daten

5.2.1.3 Logistische Regression mit Individualdaten (Modell 3)

Das binäre logistische Regressionsmodell für Individualdaten (casewise data) sei analog zu (5.12) wie folgt ausgedrückt:

$$\pi(x_k) = \frac{1}{1 + e^{-(\alpha+\beta x_k)}} \tag{5.16}$$

mit $\pi(x_k) = P(Y_k = 1|x_k)$ und für $k = 1, ..., K$.

Das logistische Regressionsmodell unterstellt dabei, dass die Y_k, im Beispiel die Antworten der Testpersonen, zweiwertige Zufallsvariablen sind, die voneinander unabhängig und mit dem jeweiligen Erwartungswert $E(Y_k|x_k) = \pi(x_k)$ verteilt sind. Derartige Variablen werden auch als Bernoulli-Variablen bezeichnet und die resultierende Wahrscheinlichkeitsverteilung heißt Bernoulli-Verteilung.[18]

Für jede Person k existiert also gemäß dem logistischen Regressionsmodell eine Wahrscheinlichkeit $\pi(x_k)$ für den Butterkauf, die durch das Einkommen der betreffenden Person verändert wird. Diese Wahrscheinlichkeiten sind nicht beobachtbar, sondern nur die Realisationen y_k, die Antworten der Testpersonen.

Für Individualdaten ist keine Linearisierung der logistischen Regressionsfunktion möglich, wie es für gruppierte Daten mittels der Logit-Transformation möglich war. Es ist daher eine andere Schätzmethode anzuwenden, die *Maximum-Likelihood-Methode*, die in Abschnitt 5.2.2 beschrieben wird.

[18] Der Name geht zurück auf Jakob Bernoulli (1656–1705). Einfachstes Beispiel eines Bernoulli-Experiments ist der Münzwurf mit dem Erwartungswert $E(Y) = \pi = 0,5$ und der Varianz $V(Y) = \pi(1 - \pi)$. Im logistischen Modell ändert sich π in Abhängigkeit von x.
Die Bernoulli-Verteilung ist ein Spezialfall der Binomialverteilung für n=1 „Experimente“. Die Binomialverteilung ergibt sich aus einer Folge von Bernoulli-Experimenten. Entsprechend sind die Kaufhäufigkeiten bei Gruppierung der Daten binomialverteilt mit n = 10 (Gruppengröße). Mit wachsendem n konvergiert die Binomialverteilung gegen die Normalverteilung.

Wir nehmen hier das Ergebnis der Schätzung vorweg. Für die Parameter α und β erhalten wir die Schätzwerte

$$a = -3,67, b = 1,83$$

und damit die logistische Funktion

$$p(x_k) = \frac{1}{1 + e^{-(-3,67+1,827x_k)}} \tag{5.17}$$

die in Abbildung 5.10 dargestellt ist. Da $b > 0$ ist, ergibt sich ein mit dem Einkommen ansteigender Verlauf, wie auch bei den vorhergehenden Modellen. Der Verlauf ist entsprechend der logistischen Funktion S-förmig. Mit zunehmender Entfernung des Einkommens vom mittleren Einkommen flacht der Anstieg ab. Für die Wahrscheinlichkeit können sich damit nur Werte zwischen 0 und 1 ergeben, wie im vorherigen Modell.

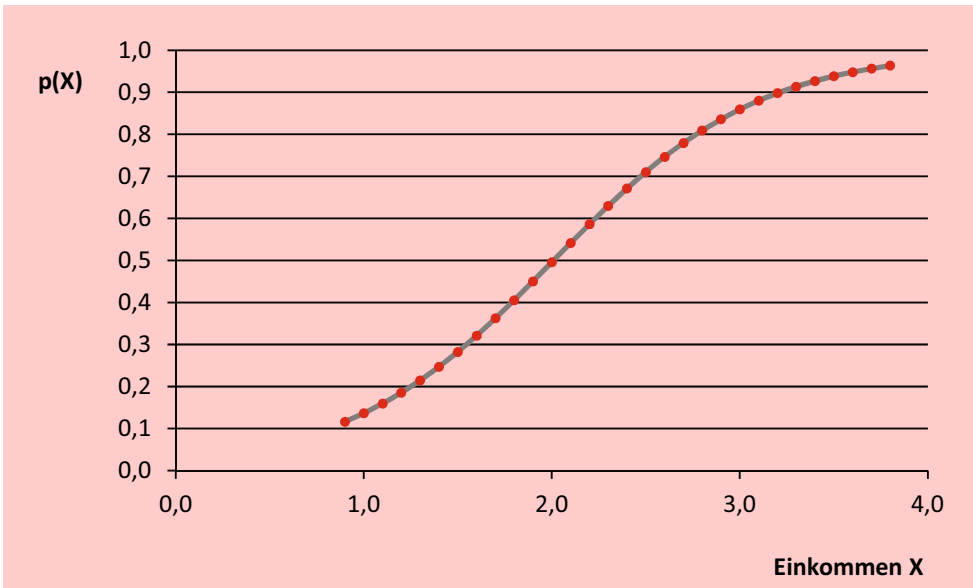


Abbildung 5.10: Geschätzte Logistische Regressionsfunktion

Vergleich der Modelle

In Abbildung 5.11 sind beispielhaft für jeweils eine Personen aus den drei Einkommensgruppen die geschätzten Wahrscheinlichkeiten der drei Modelle gegenüber gestellt. Diese liegen recht nah beieinander, besonders für die mittlere Einkommensgruppe. Mit zunehmender Entfernung des Einkommens vom Mittelwert wird sich das lineare Wahrscheinlichkeitsmodell von den anderen beiden logistischen Modellen entfernen und dann Wahrscheinlichkeiten außerhalb des Bereichs von 0 bis 1 produzieren.

In Abbildung 5.12 sind die geschätzten Funktionen der drei Modelle gemeinsam in einem Diagramm dargestellt. Der Verlauf der Kurve für das logistische Modell mit gruppierten Daten (Modell 2) ist etwas flacher als der der beiden anderen Modelle. Das liegt daran, dass durch die Gruppierung Information verloren geht. Bei einem Modell,

das keinerlei Information enthält, würde die Kurve waagerecht verlaufen, d.h. es würde für jedes Einkommen die gleiche Wahrscheinlichkeit liefern. Der Informationsverlust durch die Gruppierung macht sich hier besonders stark bemerkbar, da nur wenige Daten vorliegen und daher auch nur drei Gruppen gebildet werden konnten.

Person	Einkommen [Tsd. Euro]	Kauf	Modell 1 Lineares W.-Modell	Modell 2 LReg gruppiert	Modell 3 LReg individual
1	2,530	1	0,69	0,64	0,72
15	1,920	1	0,46	0,48	0,46
30	1,620	0	0,34	0,41	0,33

Abbildung 5.11: Vergleich von geschätzten Wahrscheinlichkeiten

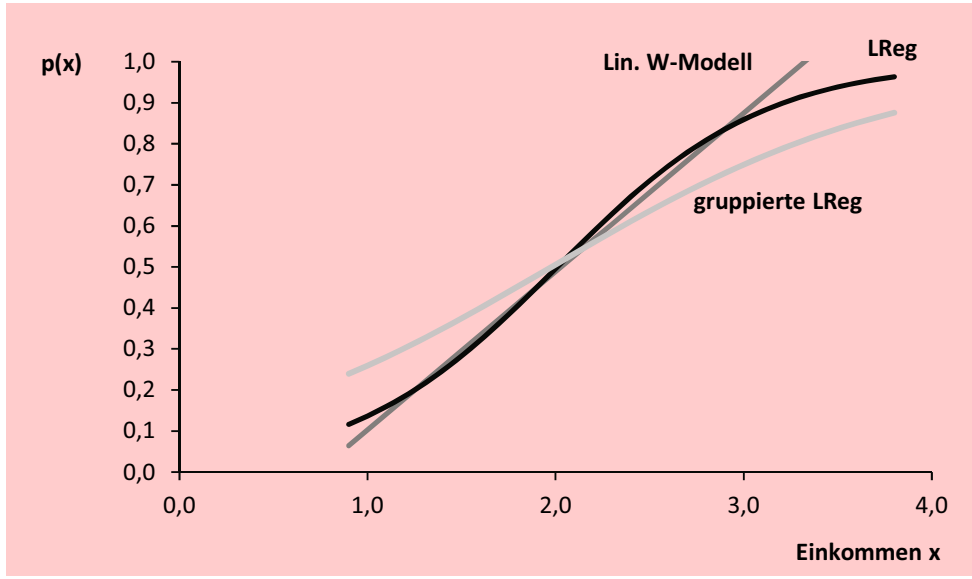


Abbildung 5.12: Vergleich der geschätzten Funktionen

Klassifizierung

Die geschätzten Wahrscheinlichkeiten können zur Prognose des Kaufverhaltens genutzt werden. Dazu ist ein *Trennwert* (cut value) p^* festzulegen. Es gelte damit für die Trennung zwischen den Alternativen:

$$y_k = \begin{cases} 1, \text{wenn } p_k > p^* \\ 0, \text{wenn } p_k \leq p^* \end{cases} \tag{5.18}$$

Trennwert

Als Trennwert wird bei zwei Alternativen gewöhnlich die Wahrscheinlichkeit $p^* = 0,5$ gewählt. Damit ist aus Abbildung 5.11 ersichtlich, dass von allen drei Modellen der Kauf bzw. Nicht-Kauf von Person 1 und 15 ex post richtig prognostiziert wird, der von Person 15 dagegen falsch. Prognosen sind üblicherweise in die Zukunft gerichtet.

Genaugenommen kann man daher von Prognosen eigentlich nur sprechen, wenn es sich um neue Käufer oder zukünftige Käufe handelt. Zur Überprüfung des Modells führen wir hier „Prognosen“ in die Vergangenheit durch. Generell unterscheidet man zwischen einer Überprüfung an der Lernstichprobe und einer Kontrollstichprobe, die aber hier nicht vorliegt.

Die Tabelle in Abbildung 5.13 zeigt die geschätzten Wahrscheinlichkeiten aller 30 Personen. Bemerkenswert ist, dass der Mittelwert der geschätzten Kaufwahrscheinlichkeiten gleich dem Anteil der Käufe (Mittel der y-Werte) ist. Das entspricht der linearen Regression (KQ-Methode), bei der ebenfalls die Mittel der beobachteten und geschätzten y-Werte immer gleich sind.

Person	Einkommen [Tsd. Euro]	Kauf	Schätzung	Prognose
1	2,530	0	0,694	1
2	2,370	1	0,767	1
3	2,720	1	0,952	1
4	2,540	0	0,852	1
5	3,200	1	0,952	1
6	2,940	0	0,767	1
7	3,200	0	0,848	1
8	2,720	1	0,582	1
9	2,930	0	0,454	1
10	2,370	0	0,535	1
11	2,240	1	0,423	1
12	1,910	1	0,458	0
13	2,120	0	0,392	1
14	1,830	1	0,338	0
15	1,920	1	0,392	0
16	2,010	0	0,365	0
17	2,010	0	0,632	0
18	2,230	1	0,697	1
19	1,820	0	0,632	0
20	2,110	0	0,493	1
21	1,750	1	0,493	0
22	1,460	1	0,578	0
23	1,610	0	0,419	0
24	1,570	1	0,531	0
25	1,370	0	0,280	0
26	1,410	1	0,323	0
27	1,510	0	0,246	0
28	1,750	1	0,261	0
29	1,680	1	0,300	0
30	1,620	0	0,342	0
Mittel	2.115	0,533	0,533	0,467

Abbildung 5.13: Geschätzte Wahrscheinlichkeiten und prognostizierte Käufe für $p^* = 0,5$

Klassifizierungstabellen

Klassifizierungstabelle

Die Prognosen für alle Personen lassen sich übersichtlich in einer *Klassifizierungstabelle* (Confusion-Matrix) zusammenstellen, wie wir sie schon bei der Diskriminanzanalyse verwendet haben und die wir hier noch etwas erweitern. Abbildung 5.14 zeigt die Klassifizierungstabelle für die logistische Regression mit Individualdaten. In der Diagonale der vier Felder (unter „Prognose“) stehen die Fallzahlen der korrekten Prognosen von Nicht-Kauf (NK) und Kauf (in fetter Schrift) und in den übrigen zwei Feldern stehen die der falschen Prognosen. In der Spalte „Summe“ stehen die Fallzahlen der beiden Gruppen und die Gesamtzahl der Fälle (hier Testpersonen). Diese Zahlen sind durch die Daten vorgegeben und müssen nicht berechnet werden. Sie müssen mit der Summe der Zellen in gleicher Zeile übereinstimmen.

	Prognose				
Gruppe	0=NK	1=Kauf	Summe	Anteil Richtige	
0 = NK	**7**	7	14	0,500	Spezifität
1 = Kauf	7	**9**	16	0,563	Sensitivität
Gesamt	14	16	30	**0,533**	Treffer

Abbildung 5.14: Klassifizierungstabelle für das logistische Modell ($p^* = 0,50$)

Rechts in der Klassifizierungstabelle stehen drei Gütemaße für die Klassifizierung, *Trefferquote*, *Sensitivität* und *Spezifität*:

Trefferquote: Anteil der richtigen Prognosen an der Zahl aller Fälle
$(7 + 9)/30 = 0,533$

Sensitivität: Anteil der richtig prognostizierten Käufe an der Zahl aller Käufe
$9/16 = 0,563$

Spezifität: Anteil der richtig prognostizierten Nicht-Käufe an der Zahl aller Nicht-Käufe
$7/14 = 0,500$

	Prognose				
Gruppe	0=NK	1=Kauf	Summe	Anteil Richtige	
0 = NK	n_{00}	n_{01}	n_0	n_{00}/n_0	Spezifität
1 = Kauf	n_{10}	n_{11}	n_1	n_{11}/n_1	Sensitivität
Gesamt			n	$(n_{00}+n_{11})/n$	Treffer

Abbildung 5.15: Berechnung der Gütemaße der Klassifizierung

Abbildung 5.15 zeigt die Berechnung in allgemeiner Form. Auf die Bedeutung der Gütemaße wird noch eingegangen.

Die erzielte Trefferquote von 53,3% ist hier sehr bescheiden und liegt nur wenig über dem, was man mit dem Werfen einer Münze erwarten würde. Zum Vergleich schauen wir uns die Klassifizierungstabelle für das lineare Wahrscheinlichkeitsmodell in Abbildung 5.16 an. Dieses Modell liefert eine Trefferquote von 0,60%, was zu der Schlussfolgerung führen kann, dass dieses Modell besser prognostiziert bzw. klassifiziert.

	Prognose				
Gruppe	0=NK	1=Kauf	Summe	Anteil Richtige	
0 = NK	**9**	5	14	0,643	Spezifität
1 = Kauf	7	**9**	16	0,563	Sensitivität
Gesamt	16	14	30	**0,600**	Treffer

Abbildung 5.16: Klassifizierungstabelle für das lineare Wahrscheinlichkeitsmodell ($p^* = 0,50$)

Das ist allerdings eine Täuschung. Die Trefferquote ist als Gütemaß nur bedingt geeignet, da sie vom mehr oder weniger willkürlich gewählten Trennwert (cut value) p^* abhängt. Wenn wir die Klassifizierung mit veränderten Trennwerten durchführen, z.B. $p^* = 0,3$ und $p^* = 0,7$, so erhalten wir mit den Daten in Abbildung 5.13 die Klassifizierungstabellen in Abbildung 5.17 und Abbildung 5.18.

	Prognose				
Gruppe	0=NK	1=Kauf	Summe	Anteil Richtige	
0 = NK	**4**	10	14	0,286	Spezifität
1 = Kauf	0	**16**	16	1,000	Sensitivität
Gesamt	4	26	30	**0,667**	Treffer

Abbildung 5.17: Klassifizierungstabelle für das lineare Wahrscheinlichkeitsmodell ($p^* = 0,30$)

	Prognose				
Gruppe	0=NK	1=Kauf	Summe	Anteil Richtige	
0 = NK	**13**	1	14	0,929	Spezifität
1 = Kauf	9	7	16	0,438	Sensitivität
Gesamt	22	8	30	**0,667**	Treffer

Abbildung 5.18: Klassifizierungstabelle für das lineare Wahrscheinlichkeitsmodell ($p^* = 0,70$)

In beiden Fällen erhalten wir jetzt eine erhöhte Trefferquote, was zeigt, welchen Einfluss der Trennwert auf die Trefferquote hat. Dass die Trefferquote hier sowohl bei Erhöhung wie auch Verringerung des Trennwertes steigt, ist allerdings unüblich.

ROC-Kurve

Ein gegenüber der Klassifizierungstabelle verallgemeinertes Konzept bildet die *ROC-Kurve* (Receiver Operating Characteristic) . Während eine Klassifizierungstabelle immer für einen bestimmten Trennwert p^* gilt, gibt die ROC-Kurve eine Zusammenfassung der Klassifizierungstabellen über die möglichen Werte von p^*.

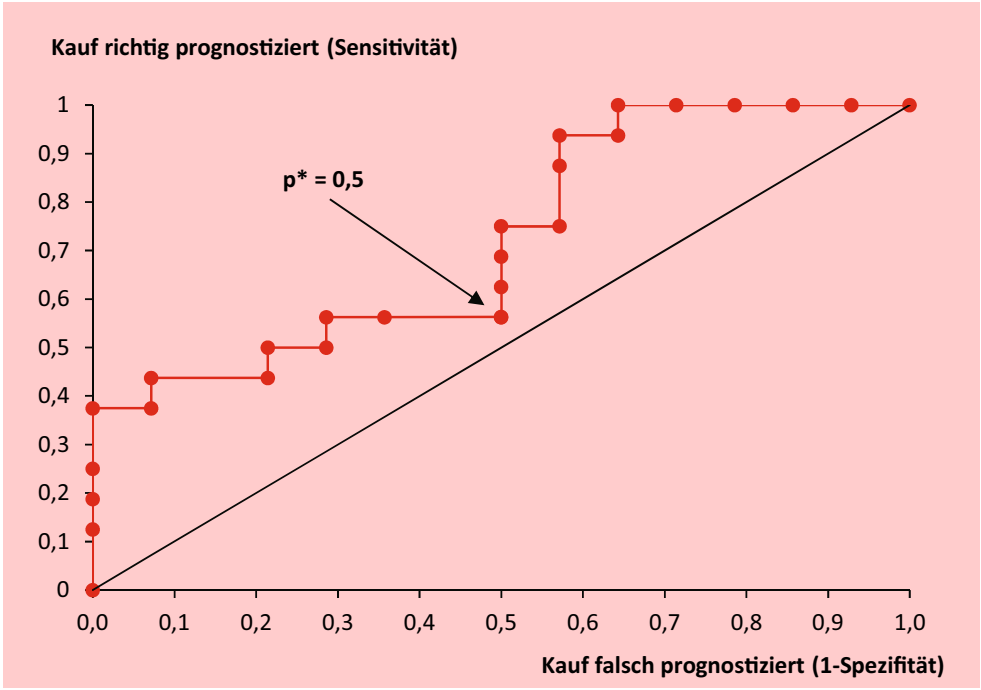


Abbildung 5.19: ROC-Kurve für die logistische Regression ($AUC = 0,723$)

Abbildung 5.19 zeigt die ROC-Kurve, die hier für das logistische Modell mit den Werten in Abbildung 5.13 erstellt wurde. Ein Punkt auf der ROC-Kurve gilt für einen bestimmten Trennwert und damit auch für eine bestimmte Klassifizierungstabelle.[19] Man erhält die ROC-Kurve, wenn man für verschiedene Trennwerte p^* die Sensitivität über 1 - Spezifität plotted. Die obige Klassifizierungstabelle in Abbildung 5.14 für $p^* = 0,5$ wird durch den markierten Punkt (0,500, 0,563) repräsentiert.

Die eingezeichnete Diagonale wäre zu erwarten, wenn die Prognose rein zufällig erfolgen würde, z.B. durch Münzwurf. Sie ermöglicht keine Diskrimination. Ein Maß für die Güte der Prognose- oder Klassifizierungsfähigkeit des Modells bildet die Fläche unter der ROC-Kurve, die als AUC (Area under Curve) bezeichnet wird. Ihr Maximum beträgt Eins. Zur Beurteilung der durch die ROC-Kurve ausgedrückte Prognosegüte gelten folgende Richtwerte:[20]

	$AUC < 0,7$:	ungenügend
$0,7 \leq$	$AUC < 0,8$:	akzeptabel
$0,8 \leq$	$AUC < 0,9$:	exzellent
	$AUC \geq 9$:	außerordentlich

[19] Das Konzept der ROC-Kurve stammt aus der Nachrichtentechnik und wurde ursprünglich im 2. Weltkrieg zur Erkennung von Radar-Signalen bzw. feindlichen Objekten entwickelt und findet heute in vielen Wissenschaftsbereichen Anwendung. Vgl. dazu z.B. Agresti (2013), S. 224 ff.; Hastie/Tibshirani/Friedman (2009), S. 313 ff.; Hosmer/Lemeshow/Sturdivant (2013), S. 173 ff. SPSS bietet eine Prozedur zur Erstellung von ROC-Kurven für gegebene Klassifizierungswahrscheinlichkeiten oder Diskriminanzwerte an. Die obige ROC-Kurve wurde mit Excel erstellt.

[20] Vgl. Hosmer/Lemeshow/Sturdivant (2013), S. 177.

Für unser Modell erhalten wir $AUC = 0,723$. Dieser Wert ergibt sich für alle drei hier verwendeten Modelle, auch wenn sie bei einzelnen Trennwerten unterschiedliche Klassifizierungstabellen liefern.[21]

Die Wahl des Trennwertes

Die Wahl des Trennwertes p^* bildet immer einen Trade-off zwischen Sensitivität und Spezifität, wie die Klassifizierungstabellen in Abbildung 5.17 und Abbildung 5.18 verdeutlichen. Die Gesamt-Trefferquote ist identisch, aber Sensitivität und Spezifität verändern sich diametral. Es sind daher auch die Konsequenzen der jeweiligen Prognosen zu berücksichtigen. Wenn es anstelle von Käufen um die Diagnose von Krankheiten geht, kann dies von eminenter Wichtigkeit sein. Ein klinischer Test für eine bestimmte Krankheit sollte positiv ausfallen, wenn der untersuchte Patient die Krankheit hat, und negativ, wenn er die Krankheit nicht hat. Die Begriffe der Sensitivität und Spezifität erhalten dann folgende Bedeutung:

Sensitivität

Fähigkeit zum richtigen Erkennen, dass der Patient von der Krankheit befallen ist, oder kurz: „Patient krank und Test positiv“ („True positiv“).

Spezifität

Fähigkeit zum richtigen Erkennen, dass der Patient nicht von der Krankheit befallen ist: „Patient gesund und Test negativ“ („True negative“).

Ist die Krankheit heilbar, falls sie schnell behandelt wird, so wäre es zweckmäßig, die Sensitivität zu erhöhen, indem man den Trennwert absenkt. In Abbildung 5.17 wird durch Verringerung des Trennwertes auf $p^* = 0,3$ die Sensitivität auf 100% erhöht.

Ist die Krankheit dagegen nicht heilbar und würde eine fälschliche Mitteilung einer Erkrankung den Patienten möglicherweise in schwere Angst und Depression versetzen, so wäre es zweckmäßig, die Spezifität zu erhöhen, indem man den Trennwert anhebt. In Abbildung 5.18 wird durch Erhöhung des Trennwertes auf $p^* = 0,7$ die Spezifität auf 92,9% erhöht.

Ähnlich ist es bei der Konstruktion eines Spam-Filters für den E-Mail- Empfang. Sei 1 = Spam und 0 = Kein-Spam (entsprechend Kauf und Nicht-Kauf im Beispiel), so misst die Sensitivität die Fähigkeit zum richtigen Erkennen von Spam. Bei hoher Sensitivität sinkt die Spezifität, also die Wahrscheinlichkeit, dass eine seriöse E-Mail (Kein-Spam) auch durchkommt und nicht fälschlich im Spam-Filter hängen bleibt. Da es unangenehm wäre, wenn eine wichtige E-Mail auf diese Art verloren geht, wird man folglich den Trennwert eher hoch ansetzen. Das hat dann zur Folge, dass die Sensitivität herabgesetzt wird und wir weiterhin viel Spam erhalten.

5.2.1.4 Multiple logistische Regression (Modell 4)

Im Folgenden befassen wir uns nur noch mit der logistischen Regression für Individualdaten. Das ist auch der Fall, der in SPSS unter „logistischer Regression“ behandelt

[21] Und man erhält auch denselben Wert für AUC, wenn man die Diskriminanzanalyse auf die vorliegenden Daten anwendet. Dabei kann man die ROC-Kurve alternativ auf Basis der Diskriminanzwerte oder der Klassifizierungswahrscheinlichkeiten erstellen.

wird, während die Modelle 1 und 2 sich, wie gezeigt, mit der linearen Regression rechnen lassen.

Bei mehr als einer unabhängigen Variablen spricht man von multipler logistischer Regression (analog zur multiplen Regressionsanalyse, die in Abschnitt 1.2.2.2 behandelt wurde). Formel (5.16) für die einfache logistischen Regression (Modell 3) lässt sich leicht erweitern zum Modell für die multiple logistische Regression.

Es sei $\mathbf{x}_k = (x_{1k}, x_{2k}, ..., x_{Jk})$ eine Menge von Werten (Beobachtungen) der J unabhängigen Variablen. Y_k bezeichnet wie bisher eine binäre Zufallsvariable (0,1-Variable), deren Eintrittswahrscheinlichkeit $\pi(x_k)$ durch x_k determiniert wird:

$$\pi(\mathbf{x}_k) = Prob(Y_k = 1|x_k) \tag{5.19}$$

Das *multiple logistische Regressionsmodell* lautet damit (unter Vernachlässigung von Index k):

$$\pi(\mathbf{x}) = \frac{1}{1 + e^{-(\alpha+\beta_1 x_1+...+\beta_J x_J)}} \tag{5.20}$$

Alternativ können auch die folgenden Formulierungen gewählt werden:

$$\pi(\mathbf{x}) = \frac{e^{\alpha+\sum_j \beta_j x_j}}{1 + e^{\alpha+\sum_j \beta_j x_j}} = \frac{e^{\alpha+\mathbf{x}\boldsymbol{\beta}'}}{1 + e^{\alpha+\mathbf{x}\boldsymbol{\beta}'}} = \frac{1}{1 + e^{-(\alpha+\mathbf{x}\boldsymbol{\beta}')}}$$

Dabei bezeichnen $\mathbf{x} = (x_{1k}, x_{2k}, ..., x_{Jk})$ und $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_J)$ jeweils Zeilenvektoren.

Für unser Beispiel mit den Daten in Abbildung 5.3 beziehen wir jetzt zusätzlich die Variable „Geschlecht“ mit ein. Das Modell lautet daher in allgemeiner Formulierung

$$Y = f(Einkommen, Geschlecht)$$

und wir erhalten gemäß (5.20) das folgende logistische Modell:

$$\pi(\mathbf{x}_k) = \frac{1}{1 + e^{-(\alpha+\beta_1 x_{1k}+\beta_2 x_{2k})}} \tag{5.21}$$

Die Schätzung mit der Maximum-Likelihood-Methode liefert für die Parameter die folgenden Schätzwerte:

$$\begin{aligned} a &= -5,635 \\ b_1 &= 2,351 \\ b_2 &= 1,751 \end{aligned}$$

und damit die folgende logistische Regressionsfunktion:

$$p(\mathbf{x}_k) = \frac{1}{1 + e^{-(-5,635+2,351x_{1k}+1,751x_{2k})}} (k = 1, ..., K) \tag{5.22}$$

Für die erste Person in Abbildung 5.3, eine Frau mit einem Einkommen von 2.530 Euro, ergibt sich für die systematische Komponente des Modells:

$$z = -5,635 + 2,351 \cdot 2,530 + 1,751 \cdot 0 = 0,313$$

und man erhält damit die Kaufwahrscheinlichkeit:

$$p = \frac{1}{1+e^{-z}} = \frac{1}{1+e^{-0,313}} = 0,578$$

Der positive Koeffizient für die Variable „Geschlecht“ weist darauf hin, dass die Gourmet-Butter (wohl infolge ihres herzhaft salzigen Geschmacks) von Männern stärker präferiert wird als von Frauen. Dies ist eine Information, die unseren Produktmanager sehr interessiert, um die Werbung für das Produkt zielgenau ausrichten zu können. Für einen Mann mit gleichem Einkommen würde sich die folgende Kaufwahrscheinlichkeit ergeben:

$$p = \frac{1}{1+e^{-(0,313+1,751)}} = 0,887$$

Das Diagramm in Abbildung 5.20 zeigt die logistischen Regressionsfunktionen für Männer und Frauen.

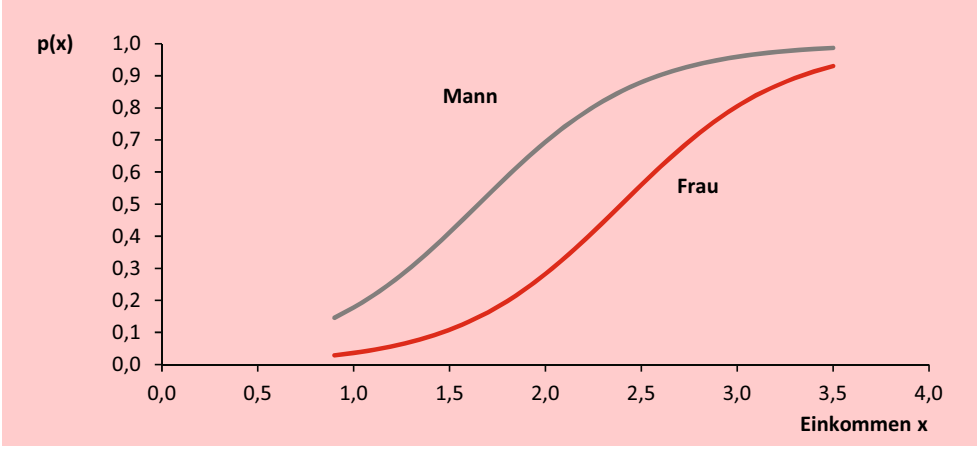


Abbildung 5.20: Logistische Regressionsfunktionen

Die Prognose mit dem geschätzten Modell liefert die Klassifizierungstabelle in Abbildung 5.21. Durch die Einbeziehung der Variable „Geschlecht“ hat sich die Prognosegüte des Modells erheblich verbessert. Der verwendete Trennwert $p^* = 0,5$ ist hier auch derjenige Trennwert, für den die Trefferquote maximal wird. Dies ist auch aus der ROC-Kurve ersichtlich, die in Abbildung 5.22 dargestellt ist. Mit $AUC = 0,813$ (Fläche unter der ROC-Kurve) ist die Prognosegüte des Modells als „exzellent“ einzustufen.

	Prognose				
Gruppe	0=NK	1=Kauf	Summe	Anteil Richtige	
0 = NK	**11**	3	14	0,786	Spezifität
1 = Kauf	2	**14**	16	0,875	Sensitivität
Gesamt	13	17	30	**0,833**	Treffer

Abbildung 5.21: Klassifizierungstabelle für das multiple logistische Modell ($p^* = 0,50$)

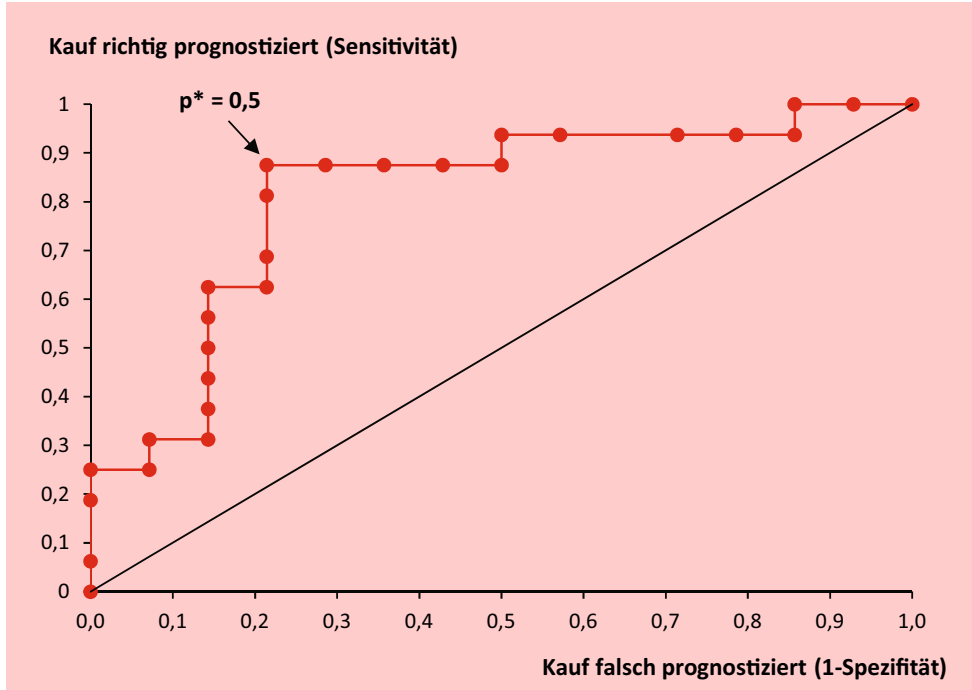


Abbildung 5.22: ROC-Kurve für die multiple logistische Regression (AUC = 0,813)

5.2.2 Schätzung der logistischen Regressionsfunktion

ML-Methode

Zur Schätzung der logistischen Regressionsfunktion ist infolge deren Nichtlinearität anstelle der Kleinst-Quadrate-Methode die *Maximum-Likelihood-Methode* (ML-Methode) anzuwenden.
Das ML-Prinzip besagt: Bestimme die Schätzwerte für die unbekannten Parameter so, *dass die realisierten Daten maximale Plausibilität (Likelihood) erlangen.*[22]

Für die Schätzung des Logistischen Regressionsmodells bedeutet dies, dass für eine Person k die Wahrscheinlichkeit $p(x_k)$ möglichst groß sein soll, falls $y_k = 1$, und möglichst klein, falls $y_k = 0$. Dies lässt sich zusammenfassen durch den folgenden Ausdruck, der möglichst groß werden soll:

$$p(x_k)^{y_k} \cdot [1 - p(x_k)]^{1-y_k} \tag{5.23}$$

Da laut Annahme des Modells die Y_k über die Personen unabhängig voneinander verteilt sein sollen, lässt sich die gemeinsame Wahrscheinlichkeit für alle Personen als Produkt der einzelnen Wahrscheinlichkeiten ausdrücken. Damit erhält man die

[22] Das Prinzip der ML-Methode geht zurück Daniel Bernoulli (1700–1782), einem Neffen von Jakob Bernoulli. Ronald A. Fisher (1890–1962) analysierte die statistischen Eigenschaften der ML-Methode und bereitete so den Weg für ihre praktische Anwendung und Verbreitung. Sie bildet neben der KQ-Methode das wichtigste statistische Schätzprinzip.

folgende *Likelihood-Funktion*, die zu maximieren ist:

$$L(a,b) = \prod_{k=1}^{K} p(x_k)^{y_k} \cdot [1 - p(x_k)]^{1-y_k} \rightarrow Max! \tag{5.24}$$

mit $y_k = 1$ für Kauf und 0 für Nicht-Kauf.
Die Parameter a und b sollen so bestimmt werden, dass die Likelihood maximal wird. Für die praktische Berechnung ist es von Vorteil, die Wahrscheinlichkeiten zu logarithmieren und damit das Produkt in eine Summe umzuwandeln. Man erhält damit die sog. *Log-Likelihood-Funktion*:

$$LL(a,b) = \sum_{k=1}^{K} \ln [p(x_k)] \cdot y_k + \ln [1 - p(x_k)] \cdot (1 - y_k) \rightarrow Max! \tag{5.25}$$

Da der Logarithmus eine streng monoton steigende Funktion ist, führt die Maximierung beider Funktionen zum gleichen Ergebnis.

LL kann nur negative Werte annehmen, da der Logarithmus einer Wahrscheinlichkeit negativ ist. Die Maximierung von LL bedeutet also, dass man dem Wert 0 möglichst nahe kommt. LL = 0 würde sich ergeben, wenn die Wahrscheinlichkeiten der gewählten Alternativen alle 1 und somit die für die nicht gewählten Alternativen 0 werden.

Abbildung 5.23 veranschaulicht den Verlauf von LL bei Variation des Koeffizienten b im Beispiel für die einfache logistische Regression (Modell 3). Für b = 1 ergibt sich für LL der Wert -28. Das Maximum ist LL = -18,027. Es wird bei b = 1,83 erreicht.

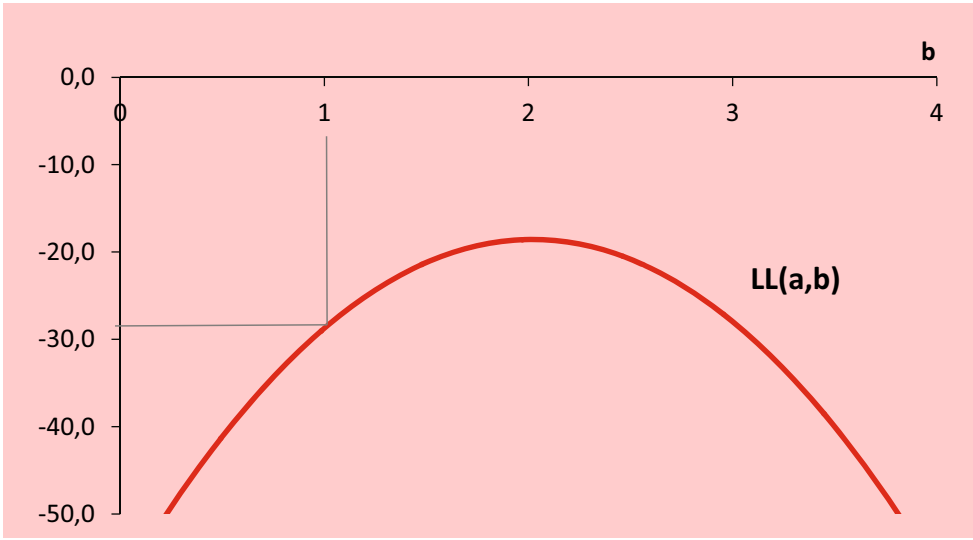


Abbildung 5.23: Maximierung der LL-Funktion

Die Lösung dieses Optimierungsproblems, d.h. die Maximierung der Log-Likelihood-Funktion, erfordert die Anwendung *iterativer Algorithmen*. In Frage kom-

men hierfür *Quasi-Newton-Verfahren* oder *Gradientenverfahren*.[23] Diese sind zwar sehr rechenaufwendig, was aber angesichts der Rechenleistung heutiger Computer kaum ins Gewicht fällt. Problematischer ist dagegen, dass iterative Algorithmen möglicherweise nicht konvergieren oder in einem lokalen Optimum hängen bleiben. Diese Gefahr besteht hier nicht, da die LL-Funktion konvex ist und somit nur ein globales Optimum existiert.[24]

5.2.3 Interpretation der Regressionskoeffizienten

Infolge der Nichtlinearität des logistischen Modells fällt die Interpretation der Koeffizienten als Maß für die Wirkung einer unabhängigen Variablen schwerer als bei den zuvor behandelten Verfahren. Das Problem ergibt sich daraus, dass die Wirkungen nicht konstant sind, sondern sich auch mit der abhängigen Variable ändern. Man kann daher generell nur sagen, wie sich bei Änderung eines Parameters oder einer unabhängigen Variablen die abhängige Variable ändert, aber nicht, um wie viel sie sich verändert. In Abbildung 5.24 sind zur Veranschaulichung der Wirkungen von Veränderungen der Parameter im logistischen Modell verschiedene Verlaufsformen der einfachen logistischen Funktion

$$p = \frac{1}{1 + e^{-(a+bx)}}$$

zusammengestellt.

Abbildung 5.24 enthält drei Diagramme:

a) Änderung des konstanten Terms (Glieds)
Eine Änderung des konstanten Terms bewirkt im logistischen Modell eine horizontale Verschiebung der Kurve über der x-Achse. Bei Vergrößerung von a verschiebt sich die Kurve nach links und die Wahrscheinlichkeit bei einem gegebenen Wert x wird größer.

b) Änderung des Koeffizienten b
Eine Vergrößerung des Koeffizienten b bewirkt einen steileren Anstieg der Kurve im mittleren Bereich. Da die Kurve S-förmig ist, muss ein steilerer Anstieg im mittleren Bereich zu einem flacheren Anstieg in den äußeren Bereichen führen. Für $b = 0$ flacht die Kurve zu einer horizontalen Linie ab.

c) Änderung des Vorzeichens von b
Ein negatives Vorzeichen des Koeffizienten b bewirkt einen abfallenden Verlauf.

[23] Für die Logistische Regression kommen primär Quasi-Newton-Verfahren zur Anwendung, die recht schnell konvergieren. Diese Verfahren basieren auf der Methode von Newton zum Auffinden der Nullstelle einer Funktion. Sie benutzen zur Auffindung des Optimums die ersten und zweiten partiellen Ableitungen der LL-Funktion nach den unbekannten Parametern. Die Ableitungen werden, je nach Verfahren, unterschiedlich approximiert. Spezielle Verfahren sind die Gauss-Newton-Methode und deren Weiterentwicklung, die Newton-Raphson-Methode. Verbreitete Anwendung findet inzwischen auch die Methode der Iteratively Reweighted Least Squares (IRLS). Siehe dazu z.B. Agresti (2013), S. 149 ff.; Fox (2015), S. 431 ff.; Press et al. (2007), S. 521 ff.

[24] McFadden (1974) hat nachgewiesen, dass bei linearer systematischer Komponente des logistischen Modells die LL-Funktion global konvex verläuft, was die Maximierung sehr erleichtert.

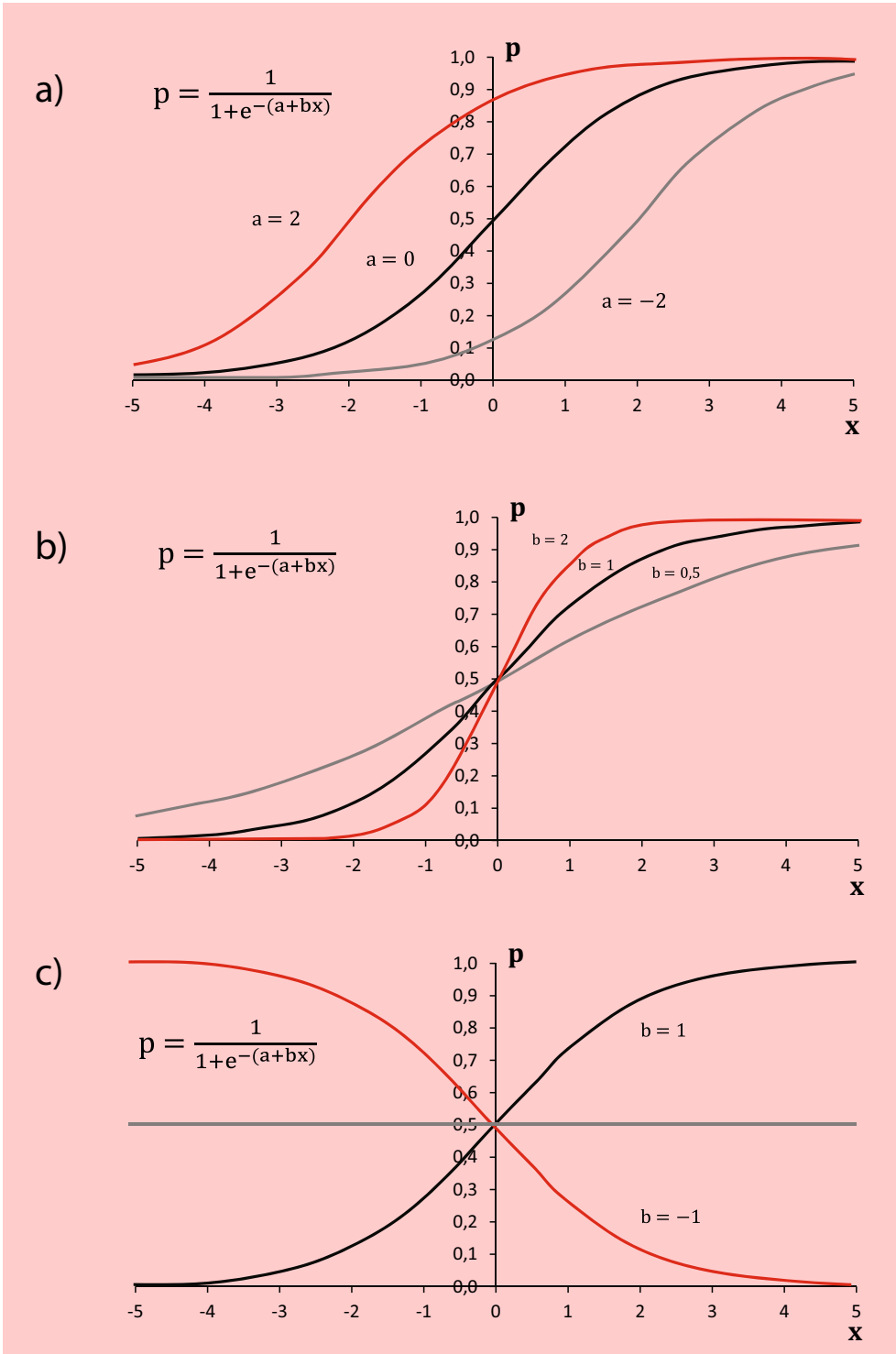


Abbildung 5.24: Verlaufsformen der logistischen Funktion bei unterschiedlichen Werten der Parameter

Der Koeffizient b bestimmt, wie die unabhängige Variable x auf die abhängige Variable p wirkt. Eine Schwierigkeit der Interpretation beim logistischen Modell resultiert daraus, wie schon erwähnt, dass die Wirkungen nicht konstant sind im Unterschied zur linearen Regression:

- Bei der linearen Regression bewirkt eine Änderung von x um eine Einheit eine Änderung der abhängigen Variablen um b.
- Bei der logistischen Regression ist die Wirkung einer Änderung von x auch abhängig vom Wert der abhängigen Variablen p. Am größten ist die Wirkung, wenn $p = 0,5$ ist, und sie wird um so geringer, je mehr p von 0,5 abweicht.

An der Stelle p beträgt die Steigung der logistischen Funktion $p(1-p)b$ und damit $0,25b$ für $p = 0,5$. An der Stelle $p = 0,01$ oder $p = 0,99$ dagegen beträgt die Steigung nur $0,01b$. Wegen der Krümmung der logistischen Funktion entspricht die Steigung nur angenähert der Veränderung von p bei einer Erhöhung von x um eine Einheit. Die Annäherung ist um so besser, je kleiner die Einheit von x gewählt wird.

Ein Rechenbeispiel soll die Wirkungen einer Änderung von x verdeutlichen. Für die einfache logistische Regression (Modell 3) hatten wir folgende Funktion geschätzt:

$$p = \frac{1}{1 + e^{-(a+bx)}} = \frac{1}{1 + e^{-(-3,67+1,827x)}} \tag{5.26}$$

Bei einem Einkommen von 2000 Euro gilt annähernd $p = 0,5$. Hier hat eine Veränderung des Einkommens die größte Wirkung auf p. Wenn das Einkommen in Schritten von jeweils 1 (Tsd. Euro) erhöht wird, dann vergrößert sich p mit abnehmenden Zuwächsen.

In Abbildung 5.25 zeigt Spalte 1a, wie sich p erhöht, und Spalte 2a zeigt die Zuwächse, die abnehmen.

Einkommen [Tsd. Euro]	Werte der Logistischen Funktion			Differenz zum vorherigen Wert		
x	1a p(x)	1b Odds	1c Logits	2a p(x)	2b Odds	2c Logits
2	0,496	0,984	-0,016			
3	0,859	6,115	1,811	0,364	5,132	1,827
4	0,974	38,01	3,638	0,115	31,899	1,827
5	0,996	236,3	5,465	0,021	198,291	1,827

Abbildung 5.25: Rechenbeispiel

Wie in Abschnitt 5.2.1.2 gezeigt, lässt sich das logistische Modell alternativ auch in Odds und in Logits ausdrücken. Damit erleichtert sich die Interpretation. Analog zu (5.8) und (5.9) schreiben wir das geschätzte Modell 3 alternativ wie folgt:

$$\text{Odds: } \frac{p}{1-p} = e^{a+bx} = e^{-3,67+1,827x} \tag{5.27}$$

$$\text{Logit: } ln\left(\frac{p}{1-p}\right) = a + bx = -3,67 + 1,827x \tag{5.28}$$

Odds

Der englische Begriff „Odds“ lässt sich übersetzen mit Chancen (auch Gewinnchancen oder Gewinnquote, z.B. beim Pferderennen).[25] Die Odds ergeben sich aus dem Verhältnis einer Wahrscheinlichkeit p zu ihrer Gegenwahrscheinlichkeit (1-p). Beträgt z.B. die Wahrscheinlichkeit für einen Gewinn 75%, so erhält man:

$$odds = \frac{p}{1-p} = \frac{0,75}{1-0,75} = 3 \tag{5.29}$$

Die Chancen für einen Gewinn stehen damit 3 zu 1. Bei einem Würfel dagegen betragen die Chancen (Odds), eine Sechs zu würfeln, nur 1 zu 5. Die Odds sind immer positiv und haben im Unterschied zu den Wahrscheinlichkeiten keine obere Grenze (siehe Spalte 1b in Abbildung 5.25). Durch die Odds ist umgekehrt auch die Wahrscheinlichkeit p bestimmt. Es gilt:

$$p = \frac{odds}{odds+1} = \frac{3}{3+1} = 0,75 \tag{5.30}$$

Zur Verdeutlichung der Wirkung einer Änderung von x auf die Odds setzen wir in (5.27) anstelle von x jetzt $x+1$ ein:

$$odds(x+1) = e^{a+b(x+1)} = e^{a+bx+b} = e^{a+bx} \cdot e^b = odds(x) \cdot e^b \tag{5.31}$$

Für das *Odds-Ratio* (OR) ergibt sich damit bei der logistischen Regression: Odds-Ratio

$$OR = \frac{odds(x+1)}{odds(x)} = e^b \tag{5.32}$$

Die Odds erhöhen sich um den Faktor e^b, wenn x um eine Einheit erhöht wird. Der Faktor e^b wird daher als Odds-Ratio oder auch als *Effekt-Koeffizient* bezeichnet. Der Wert dieses Faktors wird gewöhnlich in Statistik-Programmen zur logistischen Regression (z.B. SPSS) für jede unabhängige Variable ausgegeben. Für unser Beispiel ergibt sich: Effekt-Koeffizient

$$OR = e^b = e^{1,827} = 6,216 \tag{5.33}$$

Diesen Wert erhält man auch (abgesehen von Rundungsfehlern), wenn man in den Spalten 1b und 2b von Abbildung 5.25 die Werte durch den jeweils vorhergehenden Wert dividiert, z.B.

$$\frac{odds(3)}{odds(2)} = \frac{6,115}{0,984} = 6,214$$

Inhaltlich heißt das im Beispiel, dass sich die Chancen für einen Kauf versechsfachen, wenn das Einkommen einer Person um 1 [Tsd. Euro] Euro steigt. Die Odds erhöhen sich bei Vergrößerung des Einkommens um eine Einheit nicht um einen konstanten Betrag, sondern um einen konstanten Faktor. Bei negativem Regressionskoeffizienten ergibt sich

$$e^{-b} = \frac{1}{e^b} \leq 1$$

[25] „Odds“ im Sinne von Chancen existiert im Englischen nur im Plural, während das Wort „odd“ eine ganz andere Bedeutung hat. Im Deutschen dagegen werden die Begriffe „Chance“ und „Chancen“ oft synonym verwendet.

d.h. die Odds verringern sich um diesen Faktor bei einer Vergrößerung von x um eine Einheit. Für $b = 0$ wird der Faktor 1 und eine Änderung von x hat dann keine Wirkung. Diese Interpretationen gelten in gleicher Weise auch für die multiple logistische Regression, wenn eine der unabhängigen Variablen verändert wird und die übrigen konstant gehalten werden.

Logit

Für den Logit einer Wahrscheinlichkeit p gilt per definitionem:

$$logit(p) = ln\left(\frac{p}{1-p}\right) \tag{5.34}$$

Logit steht als Kurzform für logarithmierte Odds (logarithmic odds, Log-Odds), denn es gilt gleichermaßen:

$$logit(p) = ln(odds(p))$$

Durch die Transformation der Wahrscheinlichkeit p in Logits wird der Wertebereich von $[0, 1]$ auf den Wertebereich $[-\infty, +\infty]$ ausgedehnt, wie in Abbildung 5.7 dargestellt. Gemäß (5.32) erhöht sich $logit(p(x))$ um b, wenn sich x auf $x + 1$ erhöht:

$$logit(p(x)) = a + b(x + 1) = a + bx + b$$

Im Beispiel erhöht sich $logit(p(x))$ um 1,827, wenn sich das Einkommen um 1 Einheit [Tsd. Euro] erhöht. Vergleiche dazu die Spalten 1c und 2c in Abbildung 5.25.

Der logistische Regressionskoeffizient b lässt sich folglich als der marginale Effekt des Einkommens auf die Logits interpretieren. Die Logits erhöhen sich um b Einheiten, wenn sich das Einkommens um eine Einheit erhöht. Diese Interpretation ist besonders einfach, aber leider wenig nützlich, da wohl nur Wenige in Logit-Einheiten denken (s. dazu Abbildung 5.7. In Abbildung 5.26 sind die oben beschriebenen Wirkungen zusammengefasst.

Es gelten folgende Umkehrfunktionen:

$$p = \frac{odds}{odds + 1}, p = \frac{1}{1 + e^{-logit}} \text{ und odds} = e^{logit}$$

	Eine Erhöhung von x auf $x + 1$ bewirkt im Fall von	
	$b > 0$	$b < 0$
p	eine Erhöhung um annähernd $p(1-p)b$	eine Verminderung um annähernd $p(1-p)\|b\|$
odds	eine Erhöhung um den Faktor e^b	eine Verminderung um den Faktor $e^{-\|b\|} = \frac{1}{e^{\|b\|}}$
logit	eine Erhöhung um den Betrag b	eine Verminderung um den Betrag $\|b\|$
Odds -Ratio	Es gilt: $e^b > 1$	Es gilt: $e^b < 1$

Abbildung 5.26: Wirkungen einer Vergrößerung von x um eine Einheit bei positivem und negativem Regressionskoeffizienten

Odds-Ratio und Relatives Risiko

Wenn man das Odds-Ratio für eine metrische Variable berechnet, dann hängt seine Größe von der Maßeinheit dieser Variablen ab und ist somit wenig aussagekräftig. Anders sieht es bei binären Variablen aus. Im Modell 4 hatten wir das Geschlecht der Personen als unabhängige Variable einbezogen und die folgende Funktion (5.22) geschätzt:

$$p = \frac{1}{1 + e^{-(a+b_1 x_{1k}+b_2 x_{2k})}} = \frac{1}{1 + e^{-(-5{,}635+2{,}351 x_{1k}+1{,}751 x_{2k})}} \tag{5.35}$$

Bei einem mittleren Einkommen von 2 [Tsd. Euro] ergeben sich für Frauen und Männer die folgenden Wahrscheinlichkeiten:

Frau: $p_w = \frac{1}{1+e^{-(-5{,}635+2{,}351 \cdot 2+1{,}751 \cdot 0)}} = 0,283$
Mann: $p_m = \frac{1}{1+e^{-(-5{,}635+2{,}351 \cdot 1+1{,}751 \cdot 1)}} = 0,694$

Das Verhältnis der Wahrscheinlichkeiten für zwei Gruppen wird als Relatives Risiko (RR) bezeichnet und man erhält:

$$RR_m = \frac{p_m}{p_w} = \frac{0,694}{0,283} = 2,5$$
$$RR_w = \frac{p_w}{p_m} = \frac{0,283}{0,694} = 0,41$$

Ein Mann kauft 2,5 mal wahrscheinlicher als eine Frau bei dem gegebenen Einkommen. Umgangssprachlich wird mit dem Begriff Risiko eher die Vorstellung von negativen Ereignissen verbunden, wie z.B. Unfall, Krankheit oder Tod.

Für das entsprechende *Odds-Ratio* (*Chancenverhältnis*) ergibt sich:

$$OR_m = \frac{odds_m}{odds_w} = \frac{p_m/(1-p_m)}{p_w/(1-p_w)} = \frac{0,392}{2,267} = 5,8$$

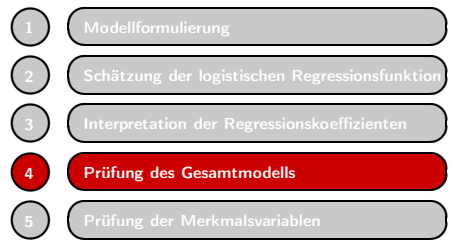
Bei einem Mann sind hier die Odds für einen Kauf 5,8 mal so hoch wie bei einer Frau. Das Odds-Ratio erhalten wir auch durch:

$$OR_m = e^{b_2} = e^{1,751} = 5,8$$
$$OR_w = e^{-b_2} = e^{-1,751} = 0,17$$

Das Odds-Ratio liefert hier bedeutend größere Werte als das Relative Risiko (bzw. bedeutend kleinere Werte bei Quotienten < 1). Unserer Vorstellung entsprechen eher die Werte von RR als die von OR. Ein Vorteil des Odds-Ratio aber liegt darin, dass es auch in Situationen angewendet werden kann, in denen sich keine Wahrscheinlichkeiten und damit kein RR berechnen lassen.[26] Das Odds-Ratio besitzt daher eine breiteres Anwendungsspektrum als das relative Risiko.

[26] Z.B. in sog. Fall-Kontroll-Studien (case-control studies), wie sie häufig in der Epidemiologie, Medizin oder Biologie als Alternative zu Kohorten-Studien durchgeführt werden. Siehe dazu z.B. Agresti (2013), 42 f.; Hosmer/Lemeshow/Sturdivant (2013), S. 229 f.

5.2.4 Prüfung des Gesamtmodells



Nachdem ein Modell geschätzt wurde, ist dessen Güte (goodness of fit) zu prüfen. Wie auch bei den zuvor behandelten Verfahren trennen wir dabei zwischen der globalen Prüfung des Modells und der Prüfung des Einflusses einzelner unabhängiger Variablen, die im nachfolgenden Abschnitt erfolgen soll.

Da zur Schätzung der logistischen Regressionsfunktion die Maximum-Likelihood-Methode (ML-Methode) verwendet wurde, ist es naheliegend, den Wert der maximierten Log-Likelihood LL (vgl. Abbildung 5.23) als Basis für die Beurteilung der Güte zu verwenden. Hierauf bauen auch die verschiedenen Gütemaße auf. Anstelle von LL wird dabei der Wert $-2LL = -2 \cdot LL$ verwendet. Da LL immer negativ ist, ist $-2LL$ positiv. Ein kleiner Wert für $-2LL$ deutet damit auf eine gute Anpassung (fit) des Modells an die vorliegenden Daten hin. Die „2“ rührt daher, dass eine Chi-Quadrat-verteilte Teststatistik angestrebt wird (siehe unten).
Für Modell 4 mit der systematischen Komponente

$$z = a + b_1 x_1 + b_2 x_2$$

ergibt sich: $-2LL = 2 \cdot 16,053 = 32,105$

Die absolute Größe dieses Wertes sagt aber wenig aus, da LL ja gemäß (5.25) eine Summe ist. Der Wert von LL und damit $-2LL$ ist also abhängig von der Anzahl K der Beobachtungen. Beide Werte würden sich folglich verdoppeln, wenn man die Beobachtungen doppeln würde, ohne dass sich dabei die Schätzwerte ändern. Die Größe $-2LL$ ist vergleichbar mit der Summe der quadrierten Residuen (SSR), die bei der KQ-Methode minimiert wird. Beide Größen werden Null bei perfekter Anpassung. Die ML-Schätzung könnte man auch durchführen, indem man die Größe $-2LL$ minimiert anstatt LL zu maximieren.

Die -2LL-Statistik kann aber zum Vergleich eines Modells mit anderen Modellen (bei gleichem Datensatz) verwendet werden. Für Modell 3, also die einfache logistische Regression mit der unabhängigen Variablen Einkommen, reduziert sich die systematische Komponente zu

$$z = a + b_1 x_1$$

und man erhält: $-2LL = 2 \cdot 18,027 = 36,054$. Es ergibt sich also bei Weglassung von Variable 2 (Geschlecht) eine Vergrößerung von $-2LL$ und damit eine Verschlechterung der Anpassung. Ein noch simpleres Modell ergibt sich mit der systematischen Komponente

$$z = a$$

Man erhält hierfür: $-2LL = 2 \cdot 20,728 = 41,455$.

In diesem Modell ist der konstante Term a der einzige Parameter, der geschätzt wird und es erzielt folglich eine noch schlechtere Anpassung. Man nennt dieses Modell

das 0-Modell (constant-only model) und es hat für sich genommen keine Bedeutung. Aber es dient zur Konstruktion des wichtigsten Tests für die Prüfung der Güte eines logistischen Modells, dem *Likelihood-Ratio-Test* (auch Likelihood-Quotienten-Test).

Likelihood-Ratio-Test

5.2.4.1 Likelihood-Ratio-Test (LR-Test)

Die Likelihood-Ratio-Statistik lautet:

$$LLR = -2 \cdot ln\left(\frac{\text{Likelihood des 0-Modells}}{\text{Likelihood des vollständigen Modells}}\right) \tag{5.36}$$
$$= -2 \cdot ln\left(\frac{L_0}{L_v}\right) = -2 \cdot (LL_0 - LL_v)$$

mit

LL_0 : Maximierte Log-Likelihood für das 0-Modell (constant-only model)
LL_v : Maximierte Log-Likelihood für das vollständige Modell (full model).

Der Likelihood-Quotient ist also gleich der Differenz der Log-Likelihoods. Unter der Nullhypothese $H_0 : \beta_1 = \beta_2 = ... = \beta_J = 0$ ist LLR angenähert χ^2-verteilt mit J Freiheitsgraden (df).[27]
In unserem Beispiel für die multiple logistische Regression (Modell 4) ergibt sich mit obigen Werten:

$$LLR = -2 \cdot (LL_0 - LL_v) = -2(-20,728 + 16,053) = 9,350 \tag{5.37}$$

Der tabellierte χ^2-Wert für $\alpha = 0,05$ und 2 Freiheitsgrade ist laut Tabelle A.4 im Anhang 5,99. Da $LLR = \chi^2_{emp} = 9,35 > 5,99$, ist die Nullhypothese abzulehnen und das Modell als statistisch signifikant anzusehen. Der p-Wert (empirisches Signifikanzniveau) beträgt nur 0,009 und das Modell ist damit sogar als hoch signifikant anzusehen.[28]

Abbildung 5.27 veranschaulicht die Log-Likelihood-Werte, die im LR-Test verwendet werden. Der LR-Test ist in seiner Bedeutung mit dem F-Test der linearen Regressionsanalyse vergleichbar.

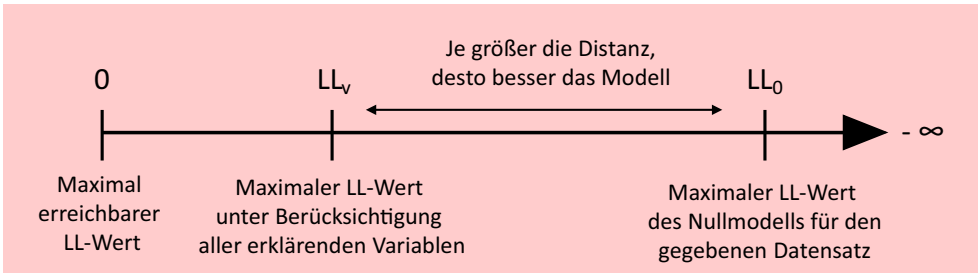


Abbildung 5.27: Log-Likelihood-Werte im LR-Test

[27] Siehe dazu z.B. Agresti (2013), S. 11; Fox (2015), S. 426 ff.
[28] Der p-Wert lässt sich in Excel mittels 1-CHIQU.VERT(x;df;1) berechnen. Hier gilt: 1-CHIQU.VERT(9,35;2;1) = 0,009.

Vergleich verschiedener Modelle

Modellbildung sollte immer um Sparsamkeit (parsimony) bemüht sein. Mit Hilfe des LR-Tests lässt sich auch überprüfen, ob ein komplexeres Modell eine signifikante Verbesserung erbringt.[29] Im Beispiel könnte man z.B. untersuchen, ob die Einbeziehung weiterer Merkmale wie Alter oder Gewicht der Personen eine bessere Modellanpassung liefern würde. Umgekehrt können wir hier untersuchen, ob Modell 4 durch Einbeziehung der Variable Geschlecht eine signifikante Verbesserung gegenüber Modell 3 erbringt. Zur Überprüfung bilden wir die folgende Likelihood-Ratio-Statistik:

$$LLR = -2 \cdot ln\left(\frac{L_r}{L_v}\right) = -2 \cdot (LL_r - LL_v) \tag{5.38}$$

mit

LL_0: Maximierte Log-Likelihood für das reduzierte Modell (Modell 3)
LL_v: Maximierte Log-Likelihood für das vollständige Modell (Modell 4).

Wir erhalten mit obigen Werten:

$$LLR = -2 \cdot (LL_r - LL_v) = -2(-18,027 + 16,053) = 3,949$$

LLR ist wiederum angenähert χ^2-verteilt, wobei die Zahl der Freiheitsgrade sich aus der Differenz der Anzahl der Parameter zwischen beiden Modellen ergibt.[30] In diesem Fall gilt df = 1. Wir erhalten damit den p-Wert 0,047. Die Verbesserung von Model 4 gegenüber Model 3 erweist sich damit bei $\alpha = 0,05$ als statistisch signifikant.

5.2.4.2 Pseudo-R-Quadrat-Statistiken

Bei der linearen Regression ist das Bestimmtheitsmaß R^2 ein sehr anschauliches Gütemaß, da es angibt, welcher Anteil der Streuung der abhängigen Variable durch das Modell erklärt wird. Ein solches Maß existiert leider für die logistische Regression nicht, da die abhängige Variable nicht metrisch ist. Das Ergebnis von Bemühungen, ähnliche Maße für die logistische Regression zu kreieren, sind die sog. Pseudo-R^2-Statistiken. Sie ähneln R^2 insofern, als sie nur Werte zwischen 0 und 1 annehmen können, und wie bei R^2 bedeutet ein höherer Wert eine bessere Anpassung.

Aber leider bedeuten sie nicht das, was R^2 bedeutet, nämlich der Anteil der erklärten Streuung an der gesamten Streuung der abhängigen Variable. Sie basieren nicht auf einer Streuungszerlegung bzw. dem Verhältnis von zwei Streuungen, sondern auf dem Verhältnis von zwei Wahrscheinlichkeiten, der Likelihood eines 0-Modells und der des vollständigen Modells (wie auch die LR-Statistik).

(a) McFadden's R^2

$$McF - R^2 = 1 - \left(\frac{LL_v}{LL_0}\right) = 1 - \frac{-16,053}{-20,728} = 0,226 \tag{5.39}$$

Im Unterschied zur LR-Statistik wird hier ein Quotient der Log-Likelihoods gebildet. Bei einem geringen Unterschied zwischen den beiden Modellen ist der Quotient nahe 1

[29] Vgl. Agresti (2013), S. 136 f.
[30] Voraussetzung für die Chi-Quadrat-Verteilung ist, dass es sich um ineinander verschachtelte Modelle (nested models) handelt. Die Variablen eines der Modelle müssen eine Untermenge der Variablen des anderen Models bilden. Siehe dazu Agresti (2013), S. 515 f.; Menard (2002), S. 22.

und $McF - R^2$ folglich nahe 0. Bei einem großen Unterschied ist es genau umgekehrt, wobei das Erreichen der 1 bei realen Datensätzen nahezu unmöglich ist. Dazu müsste die Likelihood 1 und damit die Log-Likelihood 0 werden (perfekte Anpassung). Die Werte liegen daher in der Praxis viel niedriger als bei R^2. Als Faustregel gilt, dass bereits Werte von 0,2 bis 0,4 eine gute Modellanpassung bedeuten.[31]

(b) Cox & Snell-R^2

$$R^2_{CS} = 1 - \left(\frac{L_0}{L_v}\right)^{\frac{2}{K}} = 1 - \left(\frac{exp(-20,728)}{exp(-16,053)}\right)^{\frac{2}{30}} = 0,268 \tag{5.40}$$

Das Cox & Snell-R^2 kann nur Werte < 1 annehmen, da L_0 immer > 0 sein wird, d.h. es liefert auch bei perfekter Anpassung Werte < 1.

(c) Nagelkerke's R^2

$$R^2_N = \frac{R^2_{CS}}{1 - L_0^{2/K}} = \frac{0,268^2}{1 - exp(-20,728)^{2/30}} = 0,358 \tag{5.41}$$

Das Pseudo-R^2 von Nagelkerke basiert auf der Statistik von Cox und Snell. Es modifiziert sie so, dass auch der Maximalwert 1 erreicht werden kann.
Für unser Modell liefern alle drei Pseudo-R-Quadrate recht niedrige Werte, obgleich das Modell eine recht gute Anpassung und hohe Signifikanz erzielt. Die Werte liegen weit unter dem, was man vom Bestimmtheitsmaß R^2 erwarten würde.

5.2.4.3 Beurteilung der Klassifizierung

Die Erstellung von Klassifizierungstabellen, die wir oben bereits behandelt haben, bildet eine weitere und besonders anschauliche Möglichkeit zur Beurteilung der Güte eines Modells. Leider führen diese alternativen Ansätze nicht immer zu übereinstimmenden Ergebnissen. Ein Modell kann eine gute Anpassung zeigen, aber dennoch schlechte Prognosen (Klassifizierungen) liefern.[32]

Bei der Beurteilung der Trefferquote der erzielten Klassifizierung ist immer in Betracht zu ziehen, welche Trefferquote man bei einer rein *zufälligen Zuordnung* der Elemente erwarten würde. Bei zwei Gruppen wäre durch Werfen einer Münze oder durch Würfeln eine Trefferquote von 50% zu erwarten. Dieselbe Trefferquote erzielt man auch bei gleicher Größe der Gruppen, wenn man blindlings alle Elemente einer der beiden Gruppen zuordnet. Eine noch höhere Trefferquote erzielt man mit dieser naiven Klassifizierung bei ungleicher Größe der Gruppen. Beträgt z.B. das Verhältnis der Gruppen 80 zu 20, so würde man eine Trefferquote von 80% erzielen, wenn man alle Elemente der größeren Gruppe zuordnet.

Weiterhin ist zu berücksichtigen, dass die Trefferquote immer überhöht ist, wenn sie, wie allgemein üblich, auf Basis derselben Stichprobe berechnet wird, die auch für die Schätzung der logistischen Regressionsfunktion verwendet wurde.[33] Da die logistische Regressionsfunktion immer so ermittelt wird, dass die Trefferquote in der verwendeten Stichprobe maximal wird, ist bei Anwendung auf eine andere Stichprobe mit einer niedrigeren Trefferquote zu rechnen. Dieser *Stichprobeneffekt* vermindert sich allerdings mit zunehmendem Umfang der Stichprobe. Er wird allerdings größer mit

[31] Vgl. Urban (1993), S.62.
[32] Vgl. Menard (2002), S. 37.
[33] Siehe dazu z.B. Morrison (1969), S. 158.

der Zahl der Variablen im Modell. Deshalb ist Sparsamkeit (parsimony) ein wichtiges Kriterium der Modellbildung.

Holdout-Sample

Eine *bereinigte Trefferquote* lässt sich gewinnen, indem die verfügbare Stichprobe zufällig in zwei Unterstichproben aufgeteilt wird und zwar in eine Lernstichprobe und eine Kontrollstichprobe (*Holdout-Sample*). Die Lernstichprobe wird zur Schätzung der logistischen Regressionsfunktion verwendet, mit deren Hilfe sodann die Elemente der Kontrollstichprobe klassifiziert werden und hierfür die Trefferquote berechnet wird. Diese Vorgehensweise ist allerdings nur dann zweckmäßig, wenn eine hinreichend große Stichprobe zur Verfügung steht, da sich mit abnehmender Größe der Lernstichprobe die Zuverlässigkeit der geschätzten logistischen Regressionsfunktion reduziert. Außerdem wird die vorhandene Information nur unvollständig genutzt.[34]

Leave-one-out-Methode

Bessere Möglichkeiten zur Erzielung von unverzerrten Trefferquoten bieten Methoden der Kreuz-Validierung (Cross-Validation). Ein einfacher Spezialfall ist die *Leave-one-out-Methode* (LOO). Man sondert dabei ein Element der Stichprobe aus und klassifiziert es mit Hilfe derjenigen logistischen Regressionsfunktion, die auf Basis der übrigen Elemente geschätzt wurde. Dies wird dann für alle Elemente der Stichprobe wiederholt. Auf diese Art lässt sich unter vollständiger Nutzung der vorhandenen Information eine unverzerrte Klassifizierungstabelle erzielen. Die Methode ist allerdings recht aufwendig und daher nur bei kleinem Stichprobenumfang praktikabel. In unserem Beispiel werden bei Anwendung der LOO-Methode 3 Treffer weniger erzielt, womit sich die Trefferrate von 83,3% auf 73,3% vermindert.

ROC-Kurve

Zur Beurteilung von Klassifizierungstabellen wurden diverse statistische Gütemaße entwickelt.[35] Bezüglich der Güteprüfung des zugrundeliegenden Modells ergibt sich dabei das Problem, dass die Klassifizierungstabelle und damit auch die Trefferquote sich mit der Veränderung des gewählten Trennwertes ändern können. Als ein verallgemeinertes Konzept hatten wir daher oben die ROC-Kurve (Receiver Operating Characteristic) herangezogen, die eine Zusammenfassung der Klassifizierungstabellen über die möglichen Trennwerte bildet. Ein Maß für die Güte der Prognose- bzw. Klassifizierungsfähigkeit des Modells bildet die Fläche unter der ROC-Kurve, die als AUC (Area under Curve) bezeichnet wird (siehe oben). Auch dieses Gütemaß ist normiert auf Werte zwischen 0 und 1.

Auf einen interessanten Aspekt weisen Christensen et al. (2014) hin. Mittels Simulationsstudien fanden sie heraus, dass Likelihood-basierte Gütemaße bei Stichproben mit ungleichen Gruppengrößen dazu tendieren, die Modellgüte zu unterschätzen. Die ROC-Kurve dagegen weist diesbezüglich eine hohe Stabilität auf.

5.2.5 Prüfung der Merkmalsvariablen

Um zu überprüfen, ob eine bestimmte unabhängige Variable einen signifikanten Einfluss auf die abhängige Variable hat, wird untersucht, ob ihr Koeffizient sich signifikant von 0 unterscheidet. In der linearen Regressionsanalyse wird hierfür der t-Test verwendet. Alternativ könnte man auch den F-Test verwenden und beide Tests würden zu identischen Ergebnissen führen.

[34] Zu dieser und weiteren Methoden siehe z.B. Melvin/Perreault (1977), S. 60-68; Hastie/Tibshirani/Friedman (2009), S. 241 ff.
[35] Siehe dazu z.B. Menard (2002), S. 30 ff.; Hair et al. (2010), S. 367 f.

In der logistischen Regressionsanalyse ist das etwas anders. Hier werden vornehmlich zwei Tests angewendet, die nicht immer zu gleichen Ergebnissen führen, und zwar der *Wald-Test* und der *Likelihood-Ratio-Test*, den wir schon zur Prüfung der Gesamtgüte des Modells verwendet hatten. Beide Tests werden z.B. in SPSS verwendet, der LR-Test allerdings nur in der Prozedur NOMREG für die multinomiale logistische Regression.

(a) Wald-Test

Der Wald-Test[36] ähnelt dem t-Test (genauer gesagt bildet der t-Test einen Spezialfall des Wald-Tests). Die Wald-Statistik lautet

$$W = \left(\frac{b_j}{s_{b_j}} \right)^2 \tag{5.42}$$

mit: s_{b_j} = Standardfehler von $b_j (j = 0, 1, 2, ..., J)$.

Formal ist die Wald-Statistik identisch mit dem Quadrat der t-Statistik. Unter der Nullhypothese $H_0 : \beta_i = 0$ ist sie asymptotisch χ^2-verteilt mit einem Freiheitsgrad (im Gegensatz zur t-Statistik, die Student-verteilt ist). Abbildung 5.28 zeigt für unser Beispiel die Werte der Wald-Statistik und deren p-Werte. Der Koeffizient der Variable Geschlecht ist hier nicht signifikant bei $\alpha = 0,05$.

	b_j	SE	Wald	p-Wert
Konstante	-5,635	2,417	5,436	0,041
Einkommen	2,351	1,040	5,114	0,021
Geschlecht	1,751	0,953	3,380	0,060

Abbildung 5.28: Prüfung der Regressionskoeffizienten mit dem Wald-Test

	b_j	LL_{0j}	LL_v	LLR_j	p-Wert
Konstante	-5,635	-19,944	-16,053	7,783	0,005
Einkommen	2,351	-19,643	-16,053	7,181	0,007
Geschlecht	1,751	-18,027	-16,053	3,949	0,047

Abbildung 5.29: Prüfung der Regressionskoeffizienten mit dem Likelihood-Ratio-Test

(b) Likelihood-Ratio-Test

Analog zur Verwendung des Likelihood-Ratio-Tests für die globale Güteprüfung des Modells lässt sich dieser auch verwenden, um die geschätzten Regressionskoeffizienten auf Signifikanz zu prüfen. Hierzu wird die Likelihood des vollständigen Modells LL_v mit der reduzierten Likelihood verglichen, die man erhält, wenn man den zu prüfenden Koeffizienten b_j auf 0 setzt und die Maximierung der Likelihood für die übrigen Parameter durchführt. LL_{0j} sei der Maximalwert, den man so erhält. Für die

[36] Benannt nach dem ungarischen Mathematiker Abraham Wald (1902–1950). Siehe dazu z.B. Agresti (2013), S. 10; Hosmer/Lemeshow/Sturdivant (2013), S. 42 ff.

Likelihood-Statistik zur Prüfung des Koeffizienten b_j gilt damit:

$$LLR_j = -2 \cdot (LL_{0j} - LL_v) \tag{5.43}$$

wobei LL_v wiederum den Wert für das vollständige Modell bezeichnet, der sich nicht ändert. Unter der Hypothese $H_0 : b_j = 0$ ist LLR_j asymptotisch χ^2-verteilt mit einem Freiheitsgrad. Die Ergebnisse für unser Beispiel zeigt Abbildung 5.29.

Ein Vergleich der Ergebnisse zeigt, das beim Wald-Test hier die p-Werte generell höher ausfallen, als beim Likelihood-Ratio-Test. Bei $\alpha = 0,05$ erweisen sich beim Likelihood-Ratio-Test alle Koeffizienten als signifikant, während dies beim Wald-Test für die Variable Geschlecht nicht der Fall ist.

Der Likelihood-Ratio-Test ist rechnerisch sehr viel aufwendiger als der Wald-Test, da zur Prüfung jedes der Koeffizienten eine separate ML-Schätzung durchgeführt werden muss. Aus diesem Grund wird oft dem Wald-Test der Vorzug gegeben. Der Wald-Test aber kann irreführend sein, da er systematisch größere p-Werte liefert, als der Likelihood-Ratio-Test.[37] Folglich kann er versagen, die Signifikanz eines Koeffizienten anzuzeigen (bzw. eine falsche Nullhypothese abzulehnen), so wie dies hier für die Variable Geschlecht der Fall ist. Der Likelihood-Ratio-Test ist daher der eindeutig bessere Test. Der Wald-Test sollte nur für große Stichproben zur Anwendung kommen, da sich dann die Ergebnisse der beiden Tests angleichen.

Es sei noch darauf hingewiesen, dass der LR-Test zur Signifikanzprüfung von b_2 (Koeffizient der Variable Geschlecht) identisch ist mit dem Test, den wir gemäß Formel (5.38) für den Vergleich von Modell 4 mit Modell 3 durchgeführt hatten. Die Signifikanz der Verbesserung von Modell 4 gegenüber Modell 3 durch die Einbeziehung der Variable Geschlecht ist gleich der Signifikanz des Koeffizienten der Variable Geschlecht.

5.2.6 Residuen-Analyse

Wenngleich die logistische Regression als relativ unempfindlich gegenüber Ausreißern gilt, ist es dennoch immer zweckmäßig, die Daten diesbezüglich zu kontrollieren. Möglicherweise handelt es sich um fehlerhafte Werte, die eliminiert werden sollten, da sie das Ergebnis verfälschen können. Zur Aufdeckung von Ausreißern sind die Residuen zu berechnen.

Bei der linearen Regression werden die Residuen durch $e_k = y_k - \hat{y}_k$ (beobachteter minus geschätzter Wert) berechnet. Analog werden hier die Residuen durch

$$e_k = y_k - p_k$$

berechnet (beobachteter Wert minus geschätzte Wahrscheinlichkeit). Da y nur die Werte 0 oder 1 annehmen kann und p nur Werte von 0 bis 1, können die Residuen nur Werte von -1 bis +1. Wie bei der linearen Regression ist die Summe der Residuen gleich Null.

[37] Der Grund ist, dass der Standardfehler zu groß wird, insbesondere wenn der absolute Wert des Koeffizienten groß ist. Die Wald-Statistik wird damit zu klein und der p-Wert zu groß. Siehe dazu Hauck und Donner (1977). Agresti (2013), S. 169, weist darauf hin, dass der Likelihood-Ratio-Test mehr Information nutzt als der Wald-Test und deshalb vorzuziehen ist.

Aussagekräftiger sind die standardisierten Residuen, die man erhält, wenn man die Residuen durch die Standardabweichung der Bernoulli-Verteilung dividiert:

$$z_k = \frac{y_k - p_k}{\sqrt{p_k(1-p_k)}} \tag{5.44}$$

Bei großem Stichprobenumfang K sind die standardisierten Residuen annähernd normal-verteilt mit Mittelwert 0 und Standardabweichung 1. Sie werden auch als *Pearson-Residuen* bezeichnet. Abbildung 5.30 zeigt die berechneten Residuen und Abbildung 5.31 zeigt ein Streudiagramm der standardisierten Residuen.

Pearson-Residuen

Person	Einkommen	Geschl.	Kauf y	p	y-p	z
1	2.530	0	1	0,578	0,422	0,855
2	**2.370**	**1**	0	**0,844**	**-0,844**	**-2,326**
3	2.720	1	1	0,925	0,075	0,285
4	2.540	0	0	0,583	-0,583	-1,183
5	3.200	1	1	0,974	0,026	0,162
6	2.940	0	1	0,782	0,218	0,528
7	3.200	0	1	0,869	0,131	0,389
8	2.720	1	1	0,925	0,075	0,285
9	2.930	0	1	0,778	0,222	0,534
10	2.370	0	0	0,484	-0,484	-0,969
11	2.240	1	1	0,799	0,201	0,501
12	1.910	1	1	0,647	0,353	0,738
13	2.120	0	1	0,343	0,657	1,385
14	1.830	1	1	0,603	0,397	0,811
15	1.920	1	1	0,653	0,347	0,730
16	2.010	0	0	0,287	-0,287	-0,635
17	2.010	0	0	0,287	-0,287	-0,635
18	2.230	1	0	0,796	-0,796	-1,973
19	1.820	0	0	0,205	-0,205	-0,508
20	2.110	0	0	0,338	-0,338	-0,714
21	1.750	1	1	0,557	0,443	0,891
22	1.460	1	0	0,389	-0,389	-0,798
23	**1.610**	**0**	1	**0,136**	**0,864**	**2,522**
24	1.570	1	0	0,452	-0,452	-0,908
25	1.370	0	0	0,082	-0,082	-0,299
26	1.410	1	0	0,362	-0,362	-0,752
27	1.510	0	0	0,111	-0,111	-0,353
28	1.750	1	1	0,557	0,443	0,891
29	1.680	1	1	0,516	0,484	0,968
30	1.620	0	0	0,139	-0,139	-0,401
Sum:			16	16	0,0	0,0

Abbildung 5.30: Residuenanalyse

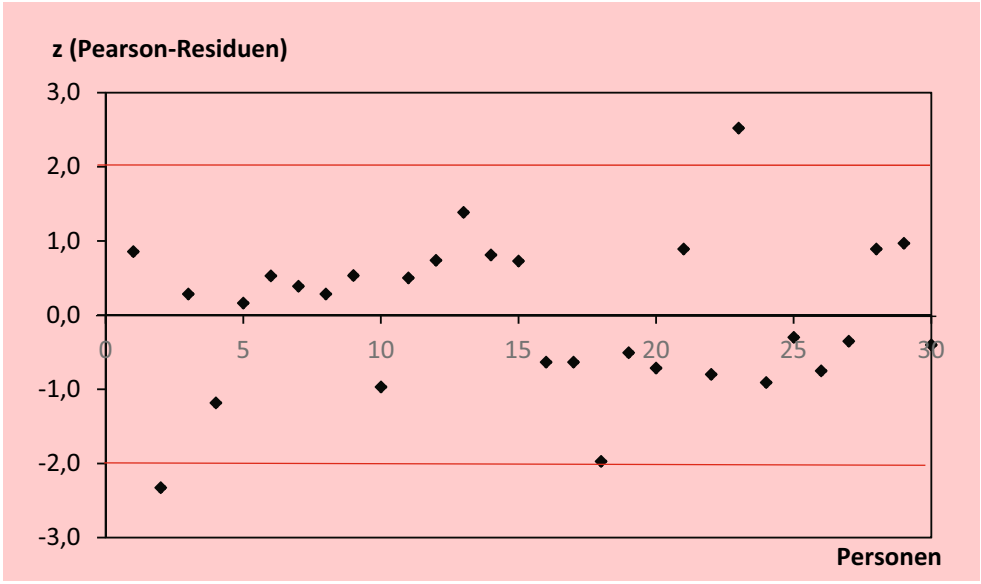


Abbildung 5.31: Standardisierte Residuen

Die Summe der quadrierten Pearson-Residuen ergibt die Pearson-Chi-Quadrat-Statistik:

$$X^2 = \sum_{k=1}^{K} \frac{(y_k - p_k)^2}{p_k(1-p_k)} = 30,039 \tag{5.45}$$

Sie wird gewöhnlich auf Basis gruppierter Daten, z.B. bei der Auswertung von Kontingenztabellen (siehe Kapitel 6), berechnet. Im Rahmen der logistischen Regression findet sie auch als Gütemaß Verwendung (siehe mathematischer Anhang). Der Wert von X^2 liegt nahe bei dem Wert 32,105, den wir für $-2LL$ erhalten hatten. Sowohl $-2LL$ wie auch X^2 sind vergleichbar mit der Summe der quadrierten Residuen (SSR) bei der linearen Regression.

Werte der standardisierten Residuen, die absolut > 2 (Standardabweichungen) sind, sollten untersucht werden. Das ist hier bei Person 2 und Person 23 der Fall.
Für die beiden auffälligen Personen gilt:

- Person 2: Mann mit hohem Einkommen, kein Kauf
- Person 23: Frau mit niedrigem Einkommen, Kauf

Die Personen entsprechen zwar nicht dem Befund, dass die Gourmet-Butter eher von Personen mit höherem Einkommen und eher von Männern gekauft wird, aber offensichtlich handelt es sich hier um Zufallsschwankungen und nicht um Ausreißer.

Man kann sich aber fragen, welchen Effekt eine auffällige Person (oder ein Ausreißer) auf die Schätzung hat. Die Wirkung hängt dabei nicht nur von der Größe des Residuums ab, sondern auch von der Distanz der zugehörigen x-Werte von ihrem Mittelpunkt (Leverage-Effekt). Am einfachsten stellt man die Wirkung fest, indem man die auffällige Person aus dem Datensatz eliminiert und die Schätzung des Modells wiederholt.

Entfernen wir z.B. Person 23 mit $z = 2,522$, so vergrößern sich die Regressionskoeffizienten wie folgt:

- Einkommen: $b_1 = 3,203$, Differenz: $3,203 - 2,351 = 0,852$
- Geschlecht: $b_2 = 2,551$, Differenz: $2,551 - 1,751 = 0,800$

Bei Person 2 sind die Änderungen nicht ganz so stark. Das Beispiel aber zeigt, das die Eliminierung eines Ausreißers bei einem kleinen Datensatz einen erheblichen Effekt haben kann.

5.2.7 SPSS-Output

In SPSS existieren, wie schon erwähnt, zwei Prozeduren zur Durchführung von logistischen Regressionsanalysen, die Prozedur LOGISTIC REGRESSION für binäre logistische Regression und die Prozedur NOMREG für multinomiale logistische Regression. Beide Prozeduren erreicht man unter dem Menüpunkt „Analysieren / Regression“. Da für das Fallbeispiel die Prozedur NOMREG verwendet wird, soll hier auszugsweise der Output der Prozedur LOGISTIC REGRESSION wiedergegeben und auf Unterschiede der Prozeduren hingewiesen werden.

Die Prozedur LOGISTIC REGRESSION erreicht man unter dem Menüpunkt „Analysieren / Regression / Binär logistisch...“. Dort ist im Dialogfenster als abhängige Variable die binäre Variable „Kauf“ anzugeben und als Kovariaten sind die Variablen „Einkommen“ und „Geschlecht“ anzugeben. Unter „Optionen“ kann man z.B. „Klassifizierungsdiagramme“ und einen „Klassifizierungstrennwert“ wählen. Klickt man danach auf „OK“, so erhält man u.a. den folgenden Output.

Omnibus-Tests der Modellkoeffizienten

		Chi-Quadrat	df	Sig.
Schritt 1	Schritt	9,350	2	,009
	Block	9,350	2	,009
	Modell	9,350	2	,009

Modellzusammenfassung

Schritt	-2 Log-Likelihood	Cox & Snell R-Quadrat	Nagelkerkes R-Quadrat
1	32,105[a]	,268	,358

a. Schätzung beendet bei Iteration Nummer 5, weil die Parameterschätzer sich um weniger als ,001 änderten.

Abbildung 5.32: Globale Gütemaße

Abbildung 5.32 zeigt im oberen Teil unter „Omnibus-Tests der Modellkoeffizienten“ das Ergebnis des Likelihood-Ratio-Tests gemäß (5.36), und zwar den Wert von LLR (Chi-Quadrat) und den zugehörigen p-Wert (Signifikanzniveau). Im unteren Teil sind als Gütemaße der Wert von $-2LL$ und die Werte von zwei Pseudo-R^2-Statistiken,

dem Cox & Snell-R^2 und Nagelkerke's R^2, angegeben. Vergleiche dazu (5.40) und (5.41). McFadden's R^2 wird nur von der Prozedur NOMREG ausgegeben. Außerdem wird angegeben, dass für die ML-Schätzung 5 Iterationen erforderlich waren.

Abbildung 5.33 zeigt die geschätzten Regressionskoeffizienten und den Wald-Test und entspricht Abbildung 5.28. Außerdem sind ganz rechts noch die Odds-Ratios gemäß (5.32) angegeben. Den Likelihood-Ratio-Test der Regressionskoeffizienten (vgl. Abbildung 5.29) erhält man in SPSS nur mit der Prozedur NOMREG.

Variablen in der Gleichung

		RegressionskoeffizientB	Standardfehler	Wald	df	Sig.	Exp(B)
Schritt 1[a]	Einkommen	2,351	1,040	5,114	1	,024	10,495
	Geschlecht	1,751	,953	3,380	1	,066	5,762
	Konstante	-5,635	2,417	5,436	1	,020	,004

a. In Schritt 1 eingegebene Variablen: Einkommen, Geschlecht.

Abbildung 5.33: Geschätzte Regressionskoeffizienten mit Wald-Test und Odds-Ratios

Klassifizierungstabelle[a]

			Vorhergesagt		
			Kauf		
	Beobachtet		Nicht-Kauf	Kauf	Prozentsatz der Richtigen
Schritt 1	Kauf	Nicht-Kauf	11	3	78,6
		Kauf	2	14	87,5
	Gesamtprozentsatz				83,3

a. Der Trennwert lautet ,500

Abbildung 5.34: Klassifizierungstabelle

Abbildung 5.34 zeigt die Klassifizierungstabelle und stimmt mit Abbildung 5.21 überein. Unter der Tabelle ist der Trennwert angegeben. Nur in der Prozedur LOGISTIC REGRESSION lässt sich der Trennwert verändern.

Um zur Prüfung der Klassifizierungstabelle eine ROC-Kurve zu erzeugen, müssen zunächst die geschätzten Wahrscheinlichkeiten, wie sie in Abbildung 5.13 für das einfache Modell 3 angegeben sind, erzeugt und in der Arbeitsdatei abgespeichert werden. Dazu ist im Dialogfenster der Prozedur LOGISTIC REGRESSION die Option „Speichern“ und dann „Wahrscheinlichkeiten“ zu wählen. Damit wird nach Durchführung der Analyse eine Variable „PRE_1“ erzeugt, die die geschätzten Wahrscheinlichkeiten $p_k = P(Y = 1)$ enthält.

Unter dem Menüpunkt „Analysieren / ROC-Kurve“ erreicht man die Prozedur ROC. Dort ist als Testvariable die Variable „PRE_1“ anzugeben und als Zustandsvariable die Variable „Kauf“. Außerdem ist unter „Wert der Zustandsvariablen“ anzugeben, für welchen Wert von Y die Wahrscheinlichkeiten gelten. Das ist hier 1. Außerdem sollten noch die Optionen „Mit diagonaler Bezugslinie“ und „Standardfehler und Konfidenzintervall“ gewählt werden. Man erhält dann den in Abbildung 5.35 und Abbildung 5.36 wiedergegeben Output.

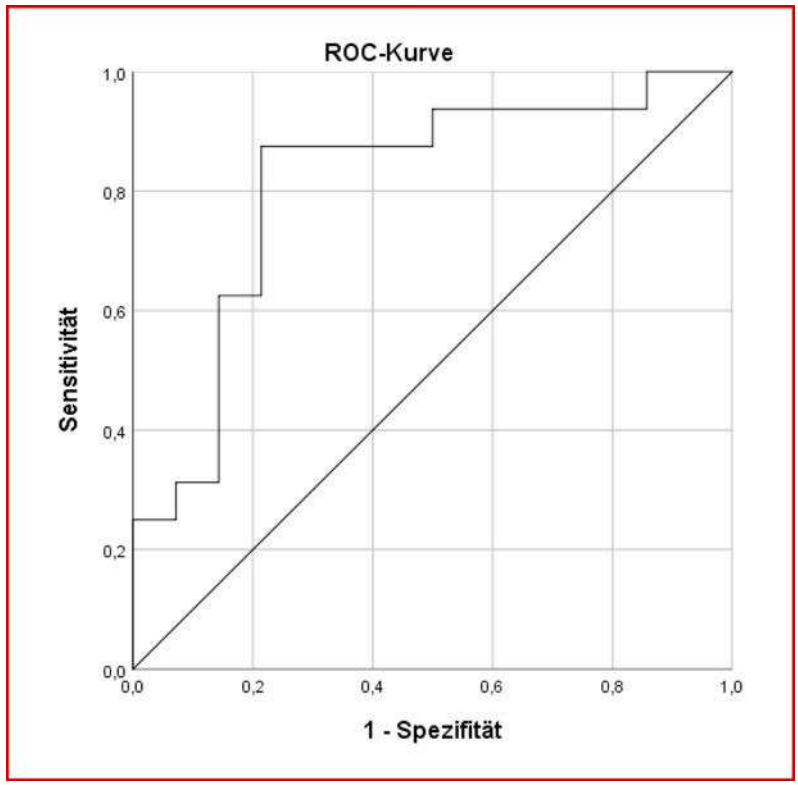


Abbildung 5.35: ROC-Kurve für Modell 4

Fläche unter der Kurve

Variable(n) für Testergebnis: Vorhergesagte Wahrscheinlichkeit

Fläche	Std.-Fehler[a]	Asymptotische Signifikanz[b]	Asymptotisches 95% Konfidenzintervall	
			Untergrenze	Obergrenze
,813	,083	,004	,649	,976

a. Unter der nichtparametrischen Annahme

b. Nullhypothese: Wahrheitsfläche = 0.5

Abbildung 5.36: Fläche unter der ROC-Kurve (AUC) mit p-Wert und Konfidenzintervall

5.3 Multinomiale logistische Regression

Ein logistisches Modell, bei dem die abhängige kategoriale Variable mehr als zwei Ausprägungen annehmen kann, wird als *multinomiales logistisches Modell* bezeichnet. Es sei wie bisher $\mathbf{x}_k = (x_{1k}, x_{2k}, ..., x_{Jk})$ eine Menge von Werten der J unabhängigen Variablen, die beobachtet wurden. Y_k bezeichne jetzt eine multinomiale Zufallsvariable, die die Ausprägungen $g = 1, ..., G$ annehmen kann. Deren Eintrittswahrscheinlichkei-

ten $\pi_g(\mathbf{x}_k)$ werden durch x_k determiniert und es gelte:

$$\pi_g(\mathbf{x}_k) = Prob(Y_k = g|x_k) \qquad (g = 1, ..., G) \tag{5.46}$$

Außerdem muss gelten:

$$\sum_{g=1}^{G} \pi_g(\mathbf{x}_k) = 1.$$

Analog zu (5.20) können wir damit das Modell der *multinomialen logistischen Regression* (MLR) unter Vernachlässigung von Index k wie folgt formulieren:[38]

$$\pi_g(\mathbf{x}) = \frac{e^{\alpha_g + \beta_{g1}x_1 + ... + \beta_{gJ}x_J}}{\sum_{h=1}^{G} e^{\alpha_h + \beta_{h1}x_1 + ... + \beta_{hJ}x_J}} \qquad (g = 1, ..., G) \tag{5.47}$$

Referenzkategorie

Da sich die Wahrscheinlichkeiten über die G Kategorien (Gruppen) zu Eins addieren müssen, sind die Parameter einer der Kategorien redundant. Zwecks Identifizierbarkeit der Parameter wird eine der G Kategorien als Referenzkategorie (baseline category) gewählt, deren Parameter auf Null gesetzt werden.[39] Gewöhnlich ist das die letzte Kategorie G. Wir erhalten damit:

$$\pi_G(\mathbf{x}) = \frac{e^{\alpha_g + \beta_{g1}x_1 + ... + \beta_{gJ}x_J}}{1 + \sum_{h=1}^{G-1} e^{\alpha_h + \beta_{h1}x_1 + ... + \beta_{hJ}x_J}} \qquad (g = 1, ..., G-1) \tag{5.48}$$

Für die Referenzkategorie G gilt: $\alpha_G = \beta_{G1} = ... = \beta_{GJ} = 0$ und damit

$$\pi_G(\mathbf{x}) = \frac{1}{1 + \sum_{h=1}^{G-1} e^{\alpha_h + \beta_{h1}x_1 + ... + \beta_{hJ}x_J}} \qquad (g = 1, ..., G-1) \tag{5.49}$$

Die Parameter der Kategorien $g = 1$ bis $G - 1$ drücken den relativen Effekt in Bezug auf die Referenzkategorie G aus.

Auch bei der binären logistischen Regression existiert immer eine Referenzkategorie, wenngleich sie nicht explizit in Erscheinung tritt. Im obigen Kaufbeispiel hatten wir $\pi(\mathbf{x}) = P(Y = 1|\mathbf{x})$ definiert, womit $Y = 0$ (Nicht-Kauf) zur Referenzkategorie wird mit $P(Y = 0|\mathbf{x}) = 1 - \pi(\mathbf{x})$. Von SPSS wird gewöhnlich die Kategorie mit dem höchsten Index als Referenzkategorie gewählt. Rechnet man das Kaufbeispiel mit der Prozedur NOMREG für multinomiale logistische Regression, so werden die Parameter für die Kategorie 0 = „Nicht-Kauf“ geschätzt und die Kategorie 1 = „Kauf“ wird als Referenzkategorie gewählt. Das hat zur Folge, dass die Parameter mit umgekehrten Vorzeichen ausgegeben werden, während ansonsten alles gleich bleibt.

Für jede Kategorie, mit Ausnahme der Referenz Kategorie, sind $J + 1$ Parameter zu schätzen (bei Einbeziehung des konstanten Terms). Damit sind insgesamt $(J + 1) \cdot (G - 1)$ Parameter zu schätzen. Bei zwei unabhängigen Variablen und drei Kategorien wären das z.B. $3 \cdot 2 = 6$ Parameter. Die Schätzung aller Parameter erfolgt simultan.

Für jede Kategorie des multinomialen logistischen Modells ist eine logistische Regressionsfunktion zu bilden, gemäß (5.48) und (5.49). Und jede dieser Funktionen umfasst alle Parameter, auch die Funktion für die Referenz Kategorie, für die keine Parameter zu schätzen sind.

[38] Vgl. z.B. Agresti (2013), S. 296; Maddala (1983), S. 34ff.; Tutz (2000), S. 162 f.

[39] Eine Alternative besteht darin, die Parameter zu zentrieren, sodass ihre Summe über die Kategorien jeweils Null ergibt.

5.3.1 Maximum-Likelihood-Schätzung

Zur Schätzung der Parameter des multinomialen logistischen Modells ist wie bisher die Log-Likelihood Funktion über die Beobachtungen $k = 1, ..., K$ zu maximieren:[40]

$$LL = \sum_{k=1}^{K} \sum_{g=1}^{G} ln\left[p_g(x_k)\right] \cdot y_{gk} \rightarrow Max! \tag{5.50}$$

mit $y_{gk} = 1$, wenn im Fall k die Kategorie g beobachtet wurde und $y_{gk} = 0$ andernfalls. und mit

$$p_g(\mathbf{x}_k) = \frac{e^{a_g + b_{g1}x_{1k} + ... + b_{gJ}x_{JK}}}{1 + \sum_{h=1}^{G-1} e^{a_h + b_{h1}x_{1k} + ... + b_{hJ}x_{JK}}} \; (g = 1, ..., G)$$

mit $a_G = b_{G1} = ... = b_{GJ} = 0$.

5.3.2 Beispiel und Interpretation

Zur Veranschaulichung modifizieren wir unser obiges Kaufbeispiel wie folgt: Es sollen den Testpersonen jetzt zwei Buttersorten A und B zur Auswahl stehen, womit sich insgesamt drei Response-Kategorien ergeben:

1. Kauf A (Kauf von Buttersorte A)
2. Kauf B (Kauf von Buttersorte B)
3. Kein Kauf

Als unabhängige Variable soll nur das Einkommen berücksichtigt werden. Damit sind $(J + 1)(G - 1) = 2 \cdot 2 = 4$ Parameter zu schätzen. Die logistischen Regressionsfunktionen für die drei Kategorien lauten:

1. Kauf A: $p_{1k} = \frac{e^{a_1 + b_1 x_k}}{1 + \sum_{h=1}^{2} e^{a_h + b_h x_k}}$
2. Kauf B: $p_{2k} = \frac{e^{a_2 + b_2 x_k}}{1 + \sum_{h=1}^{2} e^{a_h + b_h x_k}}$
3. Kein Kauf: $p_{3k} = \frac{1}{1 + \sum_{h=1}^{2} e^{a_h + b_h x_k}} = 1 - p_{1k} - p_{2k}$

Folgende Werte wurden für die vier Parameter geschätzt:

1. Kauf A: $a_1 = -14,489 \quad b_1 = 3,924$
2. Kauf B: $a_2 = 7,929 \quad b_2 = -2,772$

Außerdem gilt:

3. Kein Kauf: $a_3 = 0 \quad b_3 = 0$

Die Konstanten Terme a_1 und a_2 sagen aus, dass generell eher die Sorte B gekauft wird, als die Sorte A. Das negative Vorzeichen von a_1 besagt, dass die Wahrscheinlichkeit für einen Kauf von Sorte A generell noch niedriger ist als die Wahrscheinlichkeit dafür,

[40] Vgl. z.B. Maddala (1983), S. 36; Agresti (2013), S. 253.

dass kein Kauf erfolgt. Die Koeffizienten b_1 und b_2 sagen aus, dass eine Person mit höherem Einkommen mehr zu Sorte A tendiert, während eine Person mit niedrigerem Einkommen eher die Sorte B kaufen würde.

Für gegebenes Einkommen lässt sich die Wahrscheinlichkeit für jede der drei Alternativen berechnen. Z.B. erhält man für eine Person mit einem Einkommen von 4 [Tsd. Euro] die folgenden Wahrscheinlichkeiten:

$$p_1 = 0,763, p_2 = 0,01, p_3 = 0,228$$

Für eine Person mit einem niedrigerem Einkommen von 1,5 [Tsd. Euro] verlagert sich die Wahrscheinlichkeitsmasse weitgehend auf die zweite Buttersorte:

$$p_1 = 0,0, p_2 = 0,977, p_3 = 0,023$$

5.3.3 Das Baseline-Logit-Modell

Die einfache logistische Regressionsfunktion lässt sich für den binären Fall gemäß (5.28) wie folgt als Logit ausdrücken:

$$ln\left(\frac{p}{1-p}\right) = a + bx$$

Wir definieren jetzt für das Ausgangsbeispiel 1 = Kauf, wie bisher, und 2 = Nicht-Kauf und wir wählen die Kategorie 2 als die Baseline (Referenzkategorie). Dann lässt sich das binäre Modell wie folgt als sog. *Baseline-Logit-Modell* formulieren:

$$ln\left(\frac{P(Y=1|x)}{P(Y=2|x)}\right) = ln\left(\frac{p_1(x)}{p_2(x)}\right) = a + bx \tag{5.51}$$

Es ändert sich praktisch nichts. Aber diese Formulierung lässt sich zum *multinomialen Baseline-Logit-Modell* erweitern:

$$ln\left(\frac{p_g(\mathbf{x})}{p_G(\mathbf{x})}\right) = a_g + b_{g1}x_1 + ... + b_{gJ}x_J \; (g = 1, ..., G-1) \tag{5.52}$$

Baseline-Logit-Modell

Für den multinomialen Fall umfasst das Baseline-Logit-Modell eine Menge von $G-1$ Logit-Gleichungen. Jede Gleichung beschreibt den Effekt der unabhängigen Variablen auf die abhängige Variable, und zwar relativ zur Baseline-Kategorie. Im binären Fall gibt es nur eine derartige Gleichung. Während die logistischen Regressionsfunktionen, die zur Bestimmung der Wahrscheinlichkeiten benötigt werden, immer die Parameter für alle Kategorien umfassen, benötigt der Baseline-Logit für eine bestimmte Kategorie nur die Parameter dieser Kategorie.

In unserem Beispiel mit $G = 3$ Kategorien ergeben sich zwei Gleichungen:

$$ln\left(\frac{p_1(x)}{p_3(x)}\right) = a_1 + b_1x$$
$$ln\left(\frac{p_2(x)}{p_3(x)}\right) = a_2 + b_2x$$

Für die Person mit einem Einkommen von 4 [Tsd. Euro] erhält man mit den obigen Schätzwerten die folgenden Logits:

$$ln\left(\frac{p_1(x)}{p_3(x)}\right) = z_1 = -14,489 + 3,924 \cdot 4 = 1,209$$
$$ln\left(\frac{p_2(x)}{p_3(x)}\right) = z_2 = 7,929 - 2,772 \cdot 4 = -3,159$$

Daraus lassen sich die Odds ableiten. Für einen Kauf der Sorte A gegenüber der Alternative „Kein-Kauf“ betragen die Chancen:

$$odds_1 = e^{z_1} = e^{1,209} = 3,35$$

und für einen Kauf der Sorte B gegenüber „Kein-Kauf“ sinken die Chancen auf:

$$odds_2 = e^{z_2} = e^{-3,159} = 0,04 \text{ bzw. } 1 \text{ zu } 24.$$

Mit Hilfe der systematischen Komponente $z_g = a_g + b_g x$ für eine Gruppe g lassen sich damit recht einfach die Odds gegenüber der Referenzkategorie berechnen. Für jedes andere Paar von Kategorien, von denen keine die Referenzkategorie ist, erhält man die Logits durch:

$$ln\left(\frac{p_g(x)}{p_h(x)}\right) = ln\left(\frac{p_g(x)}{p_G(x)}\right) - ln\left(\frac{p_h(x)}{p_G(x)}\right) \tag{5.53}$$

$$= z_g - z_h = a_g - a_h + (b_g - b_h)x \tag{5.54}$$

und die Odds durch: $odds_{gh} = e^{z_g - z_h}$

Im Beispiel ergibt sich:

$$ln\left(\frac{p_1(x)}{p_2(x)}\right) = a_1 - a_2 + (b_1 - b_2)x = z_1 - z_2$$

$$\text{und } z_1 - z_2 = 1,209 - (-3,159) = 4,368$$

Damit erhält man: $odds_{12} = e^{z_1 - z_2} = 79$

Die Chancen, dass eine Person mit einem Einkommen von 4 [Tsd. Euro] die Buttersorte 1 der Sorte 2 vorzieht, stehen also 79 zu 1. Den gleichen Wert erhält man auch durch $p_1/p_2 = 0,763/0,01$, wenn man mit genügend Stellen hinter dem Komma rechnet. Die Berechnung der Odds über die Wahrscheinlichkeiten ist aber bedeutend aufwendiger.

Wenn man mit SPSS rechnet, dann kann man jede Kategorie als die Referenzkategorie wählen und so die Odds mittels des Baseline-Logits bestimmen. Trifft der Benutzer keine Wahl, so wählt SPSS (Prozedur NOMREG) immer die letzte Kategorie als Referenzkategorie.

5.3.4 Gütemaße

Zur Prüfung der Güte eines multinomialen logistischen Modells können dieselben Gütemaße und Tests verwendet werden, wie für das binäre logistische Modell. Zur Prüfung der globalen Güte hatten wir oben (Abschnitt 5.2.4) die $-2LL$-Statistik, die Likelihood-Ratio-Statistik $LLR = -2 \cdot (LL_0 - LL)$ und die darauf basierenden Pseudo-R-Quadrat-Statistiken verwendet.

Devianz
Informationskriterien

Zusätzlich werden von SPSS bei der multinomialen logistischen Regression zwei weitere Gütemaße verwendet, die *Pearson-Chi-Quadrat-Statistik* und die *Devianz* (Abweichung) und außerdem werden noch *Informationskriterien* für die Modellauswahl angegeben. Hierauf soll kurz eingegangen werden.

Die Pearson-Chi-Quadrat-Statistik und die Devianz sind beide angenähert Chi-Quadrat-verteilt und können damit, wie die LR-Statistik, zur Prüfung von Hypothesen verwendet werden. Im Unterschied zur LR-Statistik und den Pseudo-R-Quadrat-Statistiken, die mit größeren Werten eine bessere Anpassung (goodness of fit) indizieren, ist es bei der Pearson-Statistik und der Devianz umgekehrt. Bei besserer Anpassung werden sie kleiner und im Extremfall Null. Es handelt sich daher um „inverse“ Gütemaße (Badness-of-Fit-Maße).[41] Beim Hypothesentest ist daher auch nicht die Ablehnung der Null-Hypothese erwünscht, sondern deren Beibehaltung. Ein großer p-Wert ist also besser.

Pearson-Chi-Quadrat-Statistik

Die Pearson-Statistik hatten wir bereits in Abschnitt 5.2.6 zur Residuen-Analyse erwähnt. In der Form gemäß Formel (5.38) ist sie allerdings nicht Chi-Quadrat-verteilt. Die approximative Chi-Quadrat-Verteilung ergibt sich nur für gruppierte Daten, wie sie insbesondere in der Kontingenzanalyse behandelt werden.[42] Hierzu wird die Stichprobe in Teilgesamtheiten und Zellen unterteilt (siehe dazu die Ausführungen im mathematischen Anhang). Die Pearson-Statistik lautet dann im Unterschied zu (5.45):

$$X^2 = \sum_{i=1}^{I} \sum_{g=1}^{G} \frac{(m_{ig} - \hat{m}_{ig})^2}{\hat{m}_{ig}} \tag{5.55}$$

mit

m_{ig} = Ereignisse in Zelle ig
Häufigkeit, mit der $Y_i = g$ in Teilgesamtheit i beobachtet wurde.
$\hat{m}_{ig}$ = geschätzte Fallzahl für Zelle ig: $\hat{m}_{ig} = n_i p_{ig}$
n_i = Fallzahl in Teilgesamtheit i

Es gilt: $n_i = \sum_{g=1}^{G} m_{ig}$ und $K = \sum_{i=1}^{I} n_i$

Je besser die geschätzten Häufigkeiten $\hat{m}_{ig}$ mit den beobachteten Häufigkeiten m_{ig} übereinstimmen, desto kleiner wird X^2 und desto größer wird der p-Wert.

Für die Bildung der Zellen sind zum einen die Anzahl der Response-Kategorien G und zum anderen die Merkmalskombinationen $(x_{1k}, x_{2k}, ..., x_{Jk})$ der unabhängigen Variablen bestimmend. Wenn wir zwei kategoriale unabhängige Variablen haben, von denen die erste 3 Ausprägungen und die zweite 2 Ausprägungen hat, dann ergeben sich $I = 3 \cdot 2 = 6$ mögliche Merkmalskombinationen. Mit $G = 3$ ergeben sich dann $I \cdot G = 6 \cdot 3 = 18$ Zellen.

[41] Das trifft auch für die $-2LL$-Statistik zu, die allerdings nicht Chi-Quadrat-verteilt ist, was manchmal übersehen wird. Siehe dazu Agresti (2013), S. 137 ff.; McChullagh/Nelder (1989), S. 118.
[42] Vgl. Agresti (2013), S. 130.

Im Rahmen der logistischen Regression wird eine Merkmalskombination $(x_{1k}, x_{2k}, ..., x_{Jk})$ als Kovariatenmuster (covariate pattern) bezeichnet und I sei die Anzahl der unterschiedlichen Kovariatenmuster. Die Menge der Beobachtungen, die auf ein bestimmtes Kovariatenmuster i entfällt, heißt Teilgesamtheit und ihre Fallzahl sei n_i. Die m_{ig} sind dann die Fallzahlen, die in der Teilgesamtheit i auf die Response-Kategorie g entfallen. Bei metrischen Kovariaten wird es i.d.R. sehr viele unterschiedliche Kovariatenmuster und damit auch Zellen geben, und die Fallzahlen m_{ig} werden sehr klein und oft sogar Null sein. SPSS gibt dann eine Warnmeldung aus und benennt die Anzahl der leeren Zellen.

Kovariatenmuster

Im Extremfall ist jeder Fall unterschiedlich und die Anzahl der verschiedenen Kovariatenmuster ist gleich der Anzahl der Beobachtungen ($I = K$). Es ergeben sich dann $K \cdot G$ Zellen. Nur in K dieser Zellen kann dann eine Beobachtung (Response) fallen und $(G-1) \cdot K$ Zellen bleiben leer.

Die approximative Chi-Quadrat-Verteilung der Pearson-Statistik ergibt sich nur bei mehrfacher Besetzung der Zellen. Je größer die m_{ig}, desto besser ist die Approximation. Die Verwendung der Pearson-Statistik als Gütemaß macht daher bei metrischen unabhängigen Variablen keinen Sinn.

Devianz

Als Devianz (Abweichung) wird die Differenz zwischen der maximierten Log-Likelihood LL eines zu prüfenden Modells und der Log-Likelihood LLs für ein sog. saturiertes Modell bezeichnet.[43]

$$D = -2(LL - LLs) \tag{5.56}$$

Während man bei der LR-Statistik das zu prüfende Modell mit einem möglichst schlechten Modell, dem Null-Modell, vergleicht, wird bei der Devianz das zu prüfende Modell mit dem „bestmöglichen“ Modell (in Bezug auf die Anpassung) verglichen. Dabei ist das „bestmögliche“ Modell, das saturierte Modell, kein gutes Modell, da es separate Parameter für jede Beobachtung besitzt und damit natürlich eine perfekte Anpassung erzielt. Ein derartiges Modell bringt keine Vereinfachung gegenüber der Realität und genügt somit nicht dem Anspruch der Sparsamkeit. Es kann aber, wie auch das Null-Modell, als Basis zum Vergleich mit anderen Modellen dienen, um deren Güte zu beurteilen.

Die Berechnung der Devianz erfolgt, wie die der Pearson-Statistik, auf Basis der Zellen für die Kovariatenmuster (siehe dazu die Ausführungen im mathematischen Anhang). Wenn jede Beobachtung ein eigenes Kovariatenmuster hat) wird die Likelihood des saturierten Modells immer 1 und damit $LLs = 0$. Die Devianz degeneriert dann zu $D = -2LL$ und ist damit nicht mehr Chi-Quadrat-verteilt. Ihre Anwendung ist daher bei Modellen mit metrischen Kovariaten meist nicht sinnvoll, da sich zu viele unterschiedliche Kovariatenmuster und damit leere Zellen ergeben. Für die Schätzung des logistischen Modells hat dies allerdings keine Bedeutung, da dafür die Bildung der Zellen nicht benötigt wird.

[43] Vgl. Agresti (2013), S. 116 ff.; Hosmer/Lemeshow/Sturdivant (2013), S. 42 ff.

Informationskriterien für die Modellauswahl

Wenn ein Modell durch die Einbeziehung weiterer unabhängiger Variablen erweitert wird, dann wird $-2LL$ immer kleiner werden, so wie auch bei der linearen Regression die Summe der quadrierten Residuen (SSR) kleiner wird. Das Modell passt sich dadurch immer besser an die Daten der Stichprobe an. Das bedeutet aber nicht, dass das Modell besser wird. Ein Modell sollte bestmöglich die Realität (Grundgesamtheit) wiederspiegeln und nicht nur die Daten der Stichprobe.

Ein eher simples Modell mit einer akzeptablen Anpassung wird daher meist bessere Prognosen für Fälle außerhalb der Stichprobe liefern, als ein komplexeres Modell, dass innerhalb der Stichprobe besser prognostiziert. Deshalb bildet Sparsamkeit (model parsimony) ein wichtiges Kriterium der Modellbildung.

Neben Signifikanztests (wie z.B. dem Likelihood-Ratio-Test) wurden daher weitere Kriterien entwickelt, die von Nutzen sind, um Modelle mit unterschiedlicher Anzahl von Variablen zu vergleichen und zwischen diesen auszuwählen. Hierzu gehören das Akaike Informationskriterium (AIC) und das Bayessche Informationskriterium (BIC). Wie beim korrigierten Bestimmtheitsmaß der linearen Regression wird zunehmende Modellkomplexität durch eine Korrekturgröße „bestraft“ (vgl. Kapitel 1). Hier wird die Korrekturgröße hinzu addiert. Ein Modell mit kleinerem Wert des Informationskriteriums ist das bessere Modell.

Akaike Informationskriterium (AIC)

$$AIC = -2LL + 2 \cdot \text{ Zahl der Parameter}$$

Bayessches Informationskriterium (BIC)

$$BIC = -2LL + ln(K) \cdot \text{ Zahl der Parameter}$$

mit K = Stichprobenumfang.

Für die Zahl der zu schätzenden Parameter (= Freiheitsgrade) gilt:

$$\text{Zahl der Parameter} = [(G-1)(J+1)]$$

mit

- J = Anzahl der unabhängigen Variablen
- G = Anzahl der Kategorien der abhängigen Variable

In obigem Beispiel mit $G = 3$ und $K = 50$ erhalten wir für das Modell mit nur einer unabhängigen Variable, dem Einkommen, für $-2LL$ den Wert $45,5$. Für die Zahl der Parameter (Freiheitsgrade) gilt bei Einbeziehung eines konstanten Terms:

$$[(G-1)(J+1)] = [(3-1)(1+1)] = 4$$

Man erhält damit:

$$AIC = 45,5 + 2 \cdot 4 = 45,5 + 8 = 53,5$$
$$BIC = 45,5 + ln(50) \cdot 4 = 45,5 + 15,6 = 61,1.$$

Bezieht man jetzt die Variable Geschlecht in das Modell mit ein, so verringert sich der Wert von $-2LL$ auf $23,6$. Die Zahl der Parameter aber steigt auf 6 und damit erhöht sich der Bestrafungseffekt:

$$AIC = 23,6 + 2 \cdot 6 = 23,6 + 12 = 35,6$$
$$BIC = 23,6 + ln(50) \cdot 6 = 23,5 + 23,6 = 47,1$$

Beide Maße verringern sich durch die Aufnahme der zusätzlichen Variable Geschlecht in das Modell. Durch die Verminderung der Likelihood wird der Bestrafungseffekt aber mehr als kompensiert. Die Modellerweiterung ist hier also vorteilhaft.

AIC und BIC sind nur für den Vergleich von Modellen, die auf demselben Datensatz basieren, geeignet. Das Modell mit dem niedrigsten Wert ist das beste Modell. Allerdings führen AIC und BIC nicht immer zum gleichen Ergebnis. Wie sich ersehen lässt, ist der Bestrafungseffekt beim BIC größer als beim AIC. BIC favorisiert also in stärkerem Maße sparsame Modelle. Welches Kriterium „richtiger“ ist, das ist nicht eindeutig zu entscheiden. Je größer die Stichprobe ist, desto eher hat man die Gewähr, mit BIC das beste Modell auszuwählen. Bei kleinen Stichproben dagegen besteht die Gefahr, dass mittels BIC ein zu simples Modell ausgewählt wird.[44]

5.4 Fallbeispiel

5.4.1 Problemstellung

Nachfolgend soll eine multinomiale logistische Regression an dem bereits bekannten Fallbeispiel zum Margarinemarkt unter Anwendung des Computer-Programms SPSS durchgeführt werden. Wir verwenden dabei den gleichen Datensatz wie bei der Diskriminanzanalyse, um so Gemeinsamkeiten und Unterschiede zwischen beiden Verfahren besser verdeutlichen zu können. Folgende Fragen sollen untersucht werden:

- Bestehen signifikante Unterschiede in der Wahrnehmung verschiedener Marken?
- Welche Eigenschaften sind für die unterschiedliche Wahrnehmung der Marken relevant?

Bei entsprechenden Daten ließe sich die Untersuchung ausweiten auf die wichtige Frage, mit welcher Wahrscheinlichkeit bestimmte Marken gekauft werden. Für derartige Fragestellungen, wie wir sie in obigen Beispielen angesprochen haben, ist die Logistische Regression besonders geeignet.

Zur Gewinnung der Daten wurde eine Befragung von 18 Personen durchgeführt, wobei diese veranlasst wurden, 11 Butter- und Margarinemarken jeweils bezüglich 10 verschiedener Merkmalsvariablen auf einer siebenstufigen Rating-Skala zu beurteilen (vgl. Abbildung 5.37). Da nicht alle Personen alle Marken beurteilen konnten, umfasst der Datensatz nur 127 Markenbeurteilungen anstelle der vollständigen Anzahl von 198 Markenbeurteilungen (18 Personen x 11 Marken). Eine Markenbeurteilung umfasst dabei die Skalenwerte der 10 Merkmalsvariablen.

[44] Vgl. dazu Hastie/Tibshirani/Friedman (2009), S. 235.

Von den 127 Markenbeurteilungen sind nur 92 vollständig, während die restlichen 35 Beurteilungen fehlende Werte, sog. Missing Values, enthalten. Missing Values bilden ein unvermeidliches Problem bei der Durchführung von Befragungen (z. B. weil Personen nicht antworten können oder wollen oder als Folge von Interviewerfehlern). Die unvollständigen Beurteilungen sollen hier nicht berücksichtigt werden, sodass sich die Fallzahl auf 92 verringert.

Emulsionsfette (Butter- und Margarinemarken)		Merkmalsvariablen (subjektive Beurteilungen)	
1	Sanella	1	Streichfähigkeit
2	Homa	2	Preis
3	SB	3	Haltbarkeit
4	Delicado	4	Anteil ungesättigter Fettsäuren
5	Holländische Markenbutter	5	Back- und Brateignung
6	Weihnachtsbutter	6	Geschmack
7	Du darfst	7	Kaloriengehalt
8	Becel	8	Anteil tierischer Fette
9	Botteram	9	Vitamingehalt
10	Flora Soft	10	Natürlichkeit
11	Rama		

Abbildung 5.37: Untersuchte Marken und Merkmalsvariablen im Fallbeispiel

Zwecks besserer Überschaubarkeit des Marktes wurden die 11 Marken zu drei Gruppen (Marktsegmenten) zusammengefasst. Die Gruppenbildung wurde durch Anwendung einer Clusteranalyse vorgenommen (vgl. Kapitel 8). In Abbildung 5.38 ist die Zusammensetzung der Gruppen angegeben. Mittels logistischer Regression soll nun untersucht werden, ob und wie sich diese Gruppen unterscheiden. Damit erfolgt auch eine Überprüfung, wie gut die Clusteranalyse gelungen ist. Darüber hinaus soll ermittelt werden, mit welchen Wahrscheinlichkeiten sich eine bestimmte Markenbeurteilung den drei Marktsegmenten zuordnen lässt. Die abhängige Variable Y bildet hier also die Zugehörigkeit zu einem der drei Segmente und die unabhängigen Variablen sind die 10 Merkmalsvariablen, hinsichtlich derer die Markenbeurteilungen erfolgten.

Gruppen (Marktsegmente)	Marken im Segment
1	Du darfst, Becel
2	Sanella, Homa, SB, Botteram, Flora Soft, Rama
3	Delicado, Holländische Markenbutter, Weihnachtsbutter

Abbildung 5.38: Definition der Gruppen

5.4.2 Menü-Eingaben

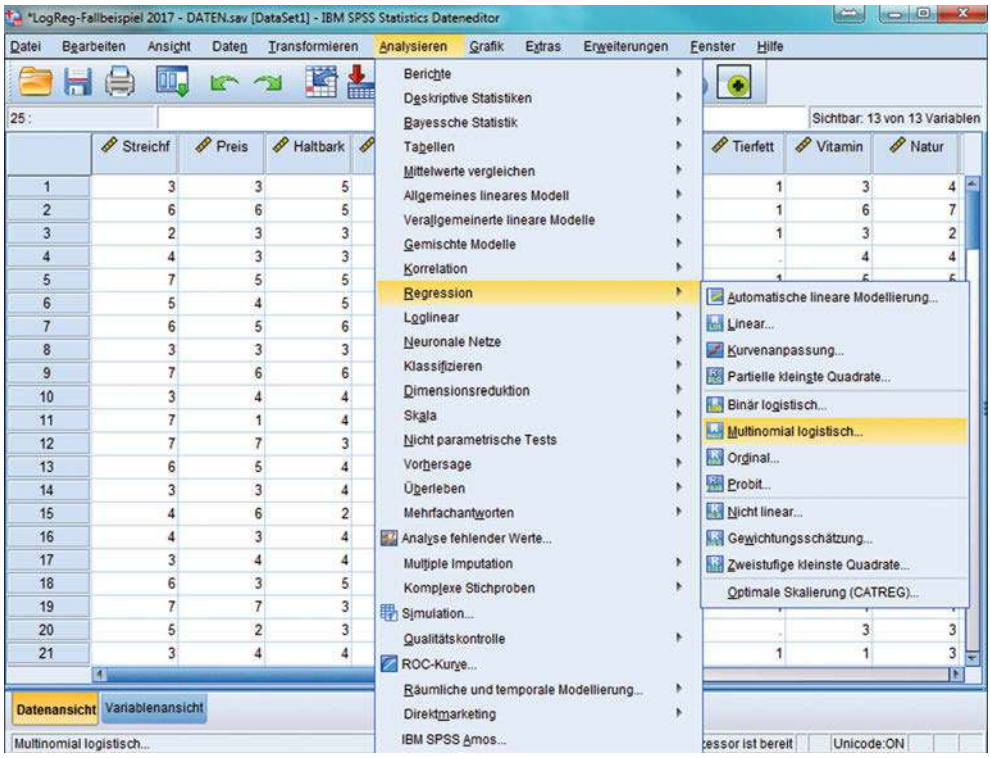


Abbildung 5.39: Daten-Editor mit Auswahl der Prozedur NOMREG (Multinomial logistische Regression)

Im Folgenden wird zunächst erläutert, wie mittels der Menüsteuerung von SPSS eine multinomiale logistische Regressionsanalyse durchgeführt werden kann. Abbildung 5.39 zeigt, wie die Prozedur NOMREG zur multinomialen logistischen Regression unter dem Menüpunkt „Analysieren“ als eine der möglichen Regressionsprozeduren aufgerufen wird. Das Untermenü zu „Regression“ zeigt gleichzeitig auch die Auswahlmöglichkeit „Binär logistische Regression“, die im Abschnitt 5.2.7 angesprochen wurde.

Nachdem die Prozedur „Multinomial logistisch“ (NOMREG) ausgewählt wurde, wird das in Abbildung 5.40 wiedergegebene Dialogfenster geöffnet. Die Variable „Segment“ ist hier die abhängige Variable und ist somit aus der linken Variablenliste auszuwählen und in das Feld „Abhängige Variable“ zu verschieben. Die Bezeichnung „Letzte“ gibt an, dass das dritte Segment von SPSS als Referenzkategorie ausgewählt wurde. Der Benutzer kann hier, wie schon erwähnt. auch ein anderes Segment als Referenzkategorie auswählen.

Wir hatten eingangs erwähnt, dass bei der logistischen Regression (wie bei der Varianzanalyse) die unabhängigen Variablen gewöhnlich als Kovariaten bezeichnet werden, wenn es sich um metrische Variablen handelt, und als Faktoren, wenn es sich um kategoriale Variablen handelt. Entsprechend finden sich im Dialogfenster zwei Felder für die Spezifikation der unabhängigen Variablen. Da es sich bei den hier zu betrachtenden unabhängigen Variablen um metrische Variablen handelt, sind sie aus der Variablenliste in das mit „Kovariate(n)“ überschriebene Feld zu verschieben.

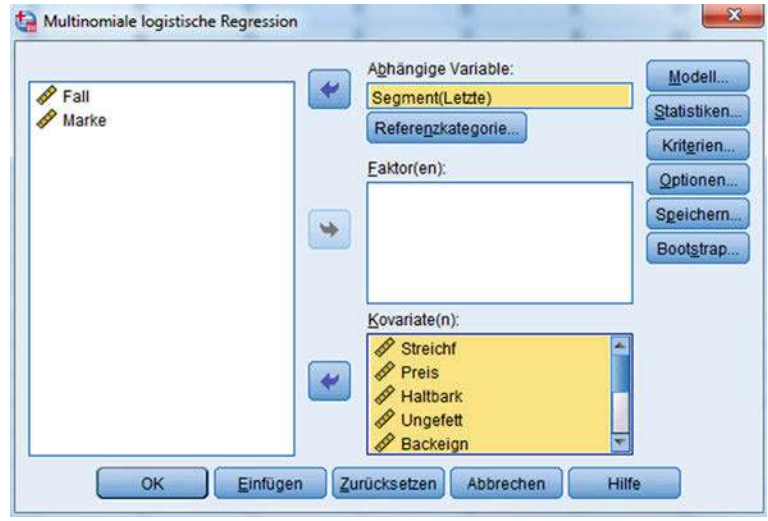


Abbildung 5.40: Dialogfenster „Multinomiale logistische Regression“

Mittels weiterer Dialogfenster, die man über die Schaltflächen rechts im Hauptmenü erreicht, lassen sich Einstellungen zum Modell und zum Output vornehmen. Das Dialogfenster „Modell“ (Abbildung 5.41) stellt drei Optionen zur Verfügung: Wird der Menüpunkt „Haupteffekte“ ausgewählt, so werden nur die Koeffizienten der im Hauptmenü ausgewählten unabhängigen Variablen (Faktoren und/oder Kovariaten)

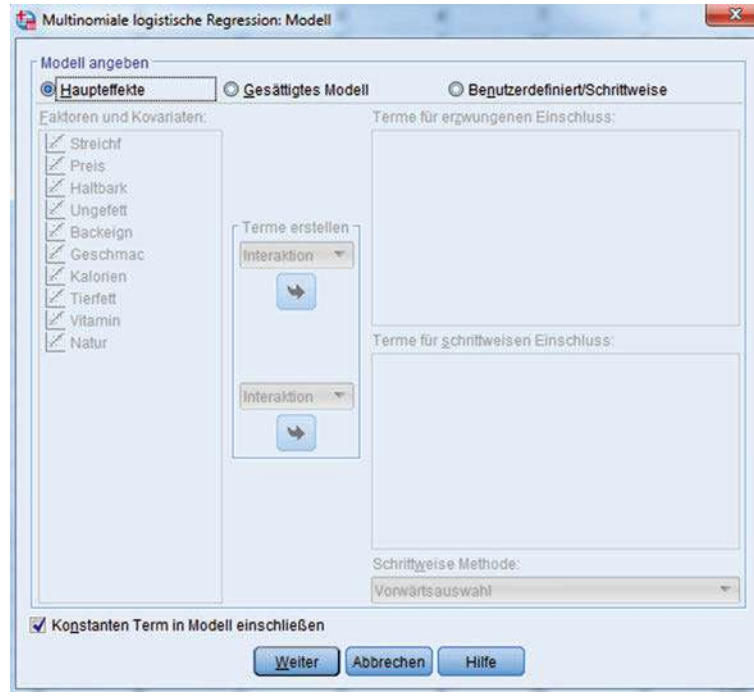


Abbildung 5.41: Dialogfenster „Modell“

geschätzt. Wird hingegen der Menüpunkt „Gesättigtes Modell“ gewählt, so werden auch alle möglichen Interaktionseffekte zwischen den ausgewählten Variablen in das Modell einbezogen. Die Wahl ist nur bei kategorialen unabhängigen Variablen möglich. Über die Option „Benutzerdefiniert/Schrittweise“ kann der Benutzer bestimmen, welche Interaktionseffekte er in das Modell einbeziehen möchte. Außerdem kann er hier angeben, ob eine schrittweise Regression durchgeführt werden soll und unter welchen Variablen das Programm dabei auswählen soll (vgl. dazu Kapitel 1, Abschnitt 1.3.2). Schließlich kann der Benutzer noch auswählen, ob das Modell einen konstanten Term enthalten soll oder nicht. Mit der Schaltfläche „Weiter“ kommt man wieder in das Hauptmenü zurück.

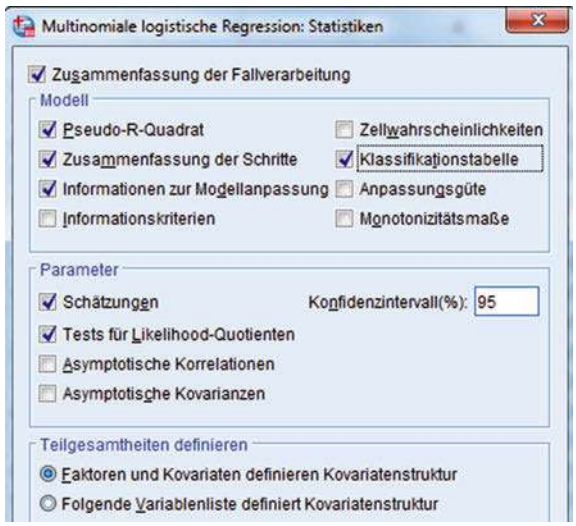


Abbildung 5.42: Dialogfenster „Statistiken“

Im Dialogfenster 'Statistiken' (Abbildung 5.42) kann der Benutzer wählen, welche Auswertungen und Gütemaße im Output erscheinen sollen:

- Die „Zusammenfassung der Fallverarbeitung“ liefert Informationen über die Fallzahlen gegliedert nach Segmenten oder solche mit fehlenden Werten.
- Unter „Modell“ lassen sich globale Gütemaße wie die Pseudo-R-Quadrat-Statistiken von McFadden, Cox & Snell sowie Nagelkerke, der Likelihood-Ratio-Test oder die Klassifizierungstabelle anfordern.
- Unter „Parameter“ lässt sich eine Tabelle mit den geschätzten Koeffizienten inklusive Standardfehler, Wald-Test und Odds-Ratio anfordern (ähnlich zu Abbildung 5.28). Mittels „Tests für Likelihood-Quotienten“ lassen sich Like-lihood-Ratio-Tests der Koeffizienten abrufen (analog zu Abbildung 5.29). Für die Odds-Ratios werden auch Konfidenzintervalle erstellt, deren Vertrauenswahrscheinlichkeit der Benutzer spezifizieren kann.

Unter der Option „Anpassungsgüte“ lassen sich Pearsons Chi-Quadrat-Statistik und die Devianz (Abweichung) anfordern. Man erhält aber bei Wahl dieser Option für die vorliegenden Daten im Output eine Warnung, da die unabhängigen Variablen hier metrisch sind und daher fast jeder Fall ein eigenes Kovariatenmuster besitzt. Insgesamt

bilden die vorliegenden 92 Fälle 89 unterschiedliche Kovariatenmuster (d.h. nur drei Fälle haben Kovariatenmuster, die noch einmal vorkommen), sodass sich 267 Zellen ergeben, von denen 176 leer sind. SPSS gibt daher im Output eine Warnmeldung.

Über das Dialogfenster „Kriterien“ kann der Benutzer Einstellungen für den Iterationsprozess zur Durchführung der Maximum-Likelihood-Schätzung vornehmen und ein Iterationsprotokoll anfordern. Über das Dialogfenster „Optionen“ lassen sich Parameter zur Durchführung einer schrittweisen Durchführung der logistischen Regressionsanalyse einstellen (vgl. dazu Kapitel 1, Abschnitt 1.3.2). Und über das Dialogfenster „Speichern“ lassen sich individuelle Ergebnisse, wie die geschätzten Wahrscheinlichkeiten oder die jeweils prognostizierte Kategorie in der Arbeitsdatei speichern, indem sie als neue Variable angehängt werden.

5.4.3 Ergebnisse

Der Output der multinomialen logistischen Regression gibt zunächst eine Übersicht über die Fallzahlen des Datensatzes (Abbildung 5.43). Von den 127 eingegebenen Fällen enthalten 35 Fälle fehlende Daten und werden daher ausgesondert. Es verbleiben 92 Fälle, die sich im Verhältnis 19:51:22 auf die Segmente A, B und C verteilen. Die Bemerkung unter der Tabelle bezieht sich auf die oben erwähnte Bildung einer Kreuztabelle, die SPSS durchführt und die aber für metrische unabhängige Variablen keinen Sinn macht.

Verarbeitete Fälle

		Anzahl	Rand-Prozentsatz
Segment	Segment A	19	20,7%
	Segment B	51	55,4%
	Segment C	22	23,9%
Gültig		92	100,0%
Fehlend		35	
Gesamt		127	
Teilgesamtheit		89[a]	

a. Die abhängige Variable hat nur einen in 89 (100,0%) Teilgesamtheiten beobachteten Wert.

Abbildung 5.43: Verarbeitete Fälle

Globale Güteprüfung des geschätzten Modells

Abbildung 5.44 zeigt im oberen Teil die Durchführung des Likelihood-Ratio-Tests (Likelihood-Quotienten-Test), den wir in Abschnitt 5.2.4.1 beschrieben hatten. Die erste Spalte zeigt die Werte von $-2LL_0$ und $-2LL_v$. Die Differenz ergibt die Likelihood-Ratio-Statistik $LLR = 183,058 - 96,684 = 86,384$, die hier als Chi-Quadrat bezeichnet ist. Mit $J \cdot (G-1) = 20$ Freiheitsgraden erhält man einen p-Wert von praktisch gleich Null. Das Modell ist also statistisch hoch signifikant, d.h. die Merkmalsvariablen trennen deutlich zwischen den drei Segmenten.

Informationen zur Modellanpassung

Modell	Kriterien für die Modellanpassung -2 Log-Likelihood	Likelihood-Quotienten-Tests Chi-Quadrat	Freiheitsgrade	Signifikanz
Nur konstanter Term	183,068			
Endgültig	96,684	86,384	20	,000

Pseudo-R-Quadrat

Cox und Snell	,609
Nagelkerke	,705
McFadden	,472

Abbildung 5.44: Globale Güteprüfung des Modells

Auch die Werte der drei Pseudo-R-Quadrat-Statistiken weisen auf eine sehr gute Modellanpassung hin. Mit den obigen Werten ergibt sich z.B. für McFadden's R^2:

$$McF - R^2 = 1 - \left(\frac{LL_v}{LL_0}\right) = 1 - \left(\frac{-2LL_v}{-2LL_0}\right) = 1 - \frac{96,684}{183,068} = 0,472$$

Schätzung der Parameter

Da das Segment $G = 3$ hier die Referenzkategorie bildet, sind die Parameter für die Segmente A und B mittels der ML-Methode gemäß (5.50) zu schätzen. Unter Einbeziehung der konstanten Terme sind hier $(G-1)(J+1) = 22$ Parameter zu bestimmen. Sie sind in Abbildung 5.45 gemeinsam mit ihren Standardfehlern wiedergegeben.

Die Koeffizienten in den Segmenten A und B haben mit Ausnahme des Koeffizienten für das Merkmal „Backeignung“ alle das gleiche Vorzeichen. Man kann daraus schließen, dass die Segmente A und B sich stärker gegenüber dem Referenz-Segment C unterscheiden als untereinander. Am stärksten unterscheiden sich die Segmente A und B voneinander bezüglich des Merkmals „Streichfähigkeit“ (siehe unten).

In Abbildung 5.45 sind außerdem die Werte der Wald-Statistik gemäß (5.42) und die zugehörigen p-Werte angegeben. Die Mehrzahl der Koeffizienten ist hier nicht signifikant bei $\alpha = 0,05$. Eine hohe Signifikanz weist in beiden Segmenten das Merkmal „Natur“ auf. Das negative Vorzeichen der Koeffizienten weist aber darauf hin, das ein Fall mit hoher Ausprägung dieses Merkmals eher der Referenzgruppe C zuzuordnen ist. Das Gleiche gilt für das Merkmal „Kalorien“. Die Merkmale, die in positiver Beziehung zu den Segmenten A und B stehen, sind alle nicht signifikant bei $\alpha = 0,05$, mit Ausnahme von „Haltbarkeit“ bezüglich Segment B. Wie aber schon oben erwähnt, sind die Signifikanzwerte mit Skepsis zu betrachten, da der Wald-Test bei kleinen Stichproben systematisch zu große p-Werte liefert.[45]

Schließlich sind in der rechten Spalte von Abbildung 5.45 noch die Odds-Ratios (Effekt-Koeffizienten) in Bezug auf die Referenzkategorie angegeben. Für positive Werte der Regressionskoeffizienten sind sie > 1 und für negative Werte < 1 (vgl.

[45] Vgl. Hauck und Donner (1977); Agresti (2013), S. 169.

Abschnitt 5.2.3). SPSS gibt außerdem noch die Konfidenzintervalle der Odds-Ratios an, auf deren Wiedergabe wir hier verzichtet haben.

Mittels Formel (5.53) lassen sich auch die Odds-Ratios zwischen den Segmenten A und B bezüglich einzelner Variablen ermitteln. Für das Merkmal 1 („Streichfähigkeit“), das am stärksten zwischen diesen beiden Segmenten diskriminiert, ergibt sich der folgende Logit:

$$ln\left(\frac{p_1(b_{11})}{p_2(b_{21})}\right) = (b_{11} - b_{21}) = 0,708 - 0,109 = 0,599$$

Damit erhält man das Odds-Ratio: $odds_{12} = e^{0,599} = 1,82$

Parameterschätzer

Segment[a]		B	Standard Fehler	Wald	Freiheitsgrade	Signifikanz	Exp(B)
Segment A	Konstanter Term	4,606	3,704	1,547	1	,214	
	Streichf	,708	,515	1,887	1	,169	2,030
	Preis	-,469	,452	1,079	1	,299	,625
	Haltbark	1,214	,678	3,202	1	,074	3,367
	Ungefett	,523	,473	1,225	1	,268	1,687
	Backeign	,118	,386	,094	1	,759	1,125
	Geschmac	-,773	,746	1,073	1	,300	,462
	Kalorien	-,979	,410	5,686	1	,017	,376
	Tierfett	-,095	,245	,150	1	,698	,909
	Vitamin	,743	,570	1,697	1	,193	2,102
	Natur	-1,861	,645	8,329	1	,004	,156
Segment B	Konstanter Term	8,238	3,395	5,890	1	,015	
	Streichf	,109	,451	,058	1	,810	1,115
	Preis	-,331	,420	,619	1	,431	,719
	Haltbark	1,537	,628	5,996	1	,014	4,651
	Ungefett	,761	,444	2,940	1	,086	2,140
	Backeign	-,096	,342	,078	1	,779	,909
	Geschmac	-1,080	,707	2,330	1	,127	,340
	Kalorien	-,820	,399	4,231	1	,040	,440
	Tierfett	-,214	,227	,882	1	,348	,808
	Vitamin	,206	,521	,156	1	,693	1,229
	Natur	-1,413	,599	5,562	1	,018	,243

a. Die Referenzkategorie lautet: Segment C.

Abbildung 5.45: Geschätzte Parameter der Regressionsfunktionen für Segment A und B

Es ist geringer als die größten Odds-Ratios der Segmente A und B in Bezug auf Segment C, was bestätigt, dass sich die Segmente A und B stärker gegenüber der Referenzkategorie unterscheiden als untereinander.

Prüfung der Merkmalsvariablen

Um die Einflussstärke der Merkmalsvariablen auf die Segmentierung zu prüfen, kann wie schon im binären Fall wiederum der Likelihood-Ratio-Test eingesetzt werden.

Likelihood-Quotienten-Tests

Effekt	Kriterien für die Modellanpassung: -2 Log-Likelihood für reduziertes Modell	Likelihood-Quotienten-Tests: Chi-Quadrat	Freiheitsgrade	Signifikanz
Konstanter Term	106,258	9,573	2	,008
Streichf	101,116	4,432	2	,109
Preis	97,841	1,157	2	,561
Haltbark	105,520	8,836	2	,012
Ungefett	100,707	4,022	2	,134
Backeign	97,339	,655	2	,721
Geschmac	99,971	3,286	2	,193
Kalorien	103,822	7,138	2	,028
Tierfett	97,900	1,216	2	,544
Vitamin	100,392	3,708	2	,157
Natur	109,235	12,551	2	,002

Die Chi-Quadrat-Statistik stellt die Differenz der -2 Log-Likelihoods zwischen dem endgültigen Modell und einem reduziertem Modell dar. Das reduzierte Modell wird berechnet, indem ein Effekt aus dem endgültigen Modell weggelassen wird. Hierbei liegt die Nullhypothese zugrunde, nach der alle Parameter dieses Effekts 0 betragen.

Abbildung 5.46: Prüfung der Merkmalsvariablen mit dem Likelihood-Ratio-Test

Das Ergebnis zeigt Abbildung 5.46. Anders als im binären Fall wird jetzt nicht jeweils ein einzelner Koeffizient geprüft, sondern es werden jeweils alle Koeffizienten einer Merkmalsvariablen geprüft. Im binären Fall existiert ja nur ein Koeffizient je Merkmalsvariable, sodass die Prüfung von Koeffizient und Variable identisch ist. In unserem Beispiel existieren jetzt zwei Koeffizienten je Merkmalsvariable.

In Abbildung 5.46 ist für jede Variable der $-2LL$-Wert des reduzierten Modells angegeben, der sich ergibt, wenn die beiden Koeffizienten auf Null gesetzt werden und die Maximierung der Likelihood für die übrigen Parameter durchführt wird. Sei LL_{0j} der Maximalwert für das um die Koeffizienten b_{1j} und b_{2j} reduzierte Modell, dann erhält man die Likelihood-Ratio-Statistik zur Prüfung der Merkmalsvariable j gemäß (5.43) durch:

$$LLR_j = -2 \cdot (LL_{0j} - LL_v)$$

Dabei bezeichnet LL_v wiederum den Wert für das vollständige Modell, den wir Abbildung 5.44 entnehmen können. Es ergibt sich damit z.B. für Variable 1 (Streichfähigkeit):

$$LLR_j = -2LL_{0j} + 2LL_v = 101,116 - 96,684 = 4,432$$

Unter der Hypothese $H_0 : b_{1j} = b_{2j} = 0$ ist LLR_j asymptotisch χ^2-verteilt mit 2 Freiheitsgraden. Der zughörige p-Wert beträgt 0,109, sodass der Einfluss der Merkmalsvariable „Streichfähigkeit“ nicht als signifikant angesehen werden kann. Den stärksten Einfluss weisen die Merkmale „Natur“, „Haltbarkeit“ und „Kalorien“ auf, die alle

statistisch signifikant sind. Den geringsten Einfluss dagegen hat die Variable „Backeignung“ mit einem p-Wert von 0,721.

Klassifizierungsergebnisse

Mittels des multinomialen logistischen Modells lassen sich für jeden Fall und jeweils alle Kategorien gemäß (5.48) und (5.49) die Wahrscheinlichkeiten für die Zuordnung von Fällen zu Segmenten ableiten. In SPSS kann man die geschätzten Wahrscheinlichkeiten über das Dialogfenster „Speichern“ anfordern. Sie werden dann in der Arbeitsdatei als neue Variablen angehängt werden. Abbildung 5.47 zeigt einen Ausschnitt aus der Arbeitsdatei im Dateneditor mit den kreierten Variablen EST1, EST2 und EST3 für die Wahrscheinlichkeiten drei Segmente. Die Variable PRE gibt die vorhergesagte Kategorie an. Das ist jeweils die Kategorie mit der größten Wahrscheinlichkeit. Für Fall 3 wird z.B. richtig die Zugehörigkeit zu Segment 2 „prognostiziert“, für Fall 12 wird dagegen falsch die Zugehörigkeit zu Segment 1 prognostiziert.

Durch Gegenüberstellung von beobachteten und vorhergesagten Segmenten lässt sich die Klassifizierungstabelle in Abbildung 5.48 erstellen, die jetzt 9 Zellen enthält. In den diagonalen Zellen stehen die Treffer. Von den 19 Fällen in Segment A werden nur 6 Fälle richtig vorhergesagt (31,6%), von den 51 Fällen von Segment B dagegen 45 (88, 2%) und von den 22 Fällen in Segment werden 19 (86,4%) richtig prognostiziert. Insgesamt ergibt sich eine Trefferquote von 76,1%.

Fall	Marke	Segment	EST1_1	EST2_1	EST3_1	PRE_1
1	1	2,00	,011	,989	,000	2
3	1	2,00	,142	,432	,426	2
4	1	2,00	,041	,948	,012	2
7	1	2,00	.	.	.	.
11	1	2,00	,232	,739	,029	2
12	1	2,00	,856	,144	,001	1
16	1	2,00	.	.	.	.
18	1	2,00	,047	,951	,001	2
2	2	2,00	,775	,209	,016	1
4	2	2,00	,041	,811	,148	2
7	2	2,00	,216	,783	,001	2
8	2	2,00	,115	,883	,002	2

Abbildung 5.47: Geschätzte Wahrscheinlichkeiten (Ausschnitt der Arbeitsdatei)

Klassifikation

Beobachtet	Vorhergesagt			
	Segment A	Segment B	Segment C	Prozent richtig
Segment A	6	12	1	31,6%
Segment B	4	45	2	88,2%
Segment C	1	2	19	86,4%
Prozent insgesamt	12,0%	64,1%	23,9%	76,1%

Abbildung 5.48: Klassifizierungstabelle für das Fallbeispiel

Die Überprüfung der Klassifizierung mit der ROC-Kurve muss separat für jedes Segment erfolgen. Für die Fläche unter der ROC-Kurve (AUC) erhält man folgende Werte für die drei Segmente: $AUC1 = 0,867, AUC2 = 0,901, AUC3 = 0,961$. Sie unterstreichen die Klassifizierungsfähigkeit des Modells. Den Wert von 76,1% für die Trefferquote hatten wir auch mit der Diskriminanzanalyse (unter Berücksichtigung ungleicher Streuungen) erzielt. Bezüglich der Vorhersagen für die Segmente A und B unterscheiden sich die Klassifizierungstabellen aber deutlich. Für Segment A wurde mit der Diskriminanzanalyse eine Trefferquote von 68,4% erzielt und für Segment B eine Trefferquote von 74,5%. Die Diskriminanzanalyse lieferte also für die kleinste Gruppe A eine höhere Trefferquote und für die größte Gruppe B eine niedrigere Trefferquote. Bezüglich der Bedeutung der Merkmalsvariablen lieferten beide Verfahren ein ähnliches Bild. Wie auch hier, ergab sich bei der Diskriminanzanalyse für die Variable „Natur“ die größte diskriminatorische Bedeutung und für die Variable „Backeignung“ die geringste. Diskriminanzanalyse und logistische Regression unterscheiden sich zwar sehr stark hinsichtlich ihrer Methodik und statistischen Annahmen, erbringen aber sehr ähnliche Ergebnisse. Kritisch ist zu der hier durchgeführten Analyse anzumerken, dass das geschätzte Modell sehr viele Variablen enthält. Das entspricht nicht dem Gebot der Sparsamkeit bei der Modellbildung. Ziel war aber, zunächst mal herauszufinden, welche Variablen überhaupt zur Unterscheidung zwischen den Markensegmenten relevant sind. Überlässt man SPSS die Modellbildung, indem man eine schrittweise Regression durchführt, so wird ein Modell mit nur drei Variablen gebildet. Ausgewählt werden die Variablen „Natur“, „Vitamin“ und „Tierfett“. Die beiden letzten Variablen waren im vollständigen Modell nicht signifikant, was auf Multikollinearität zwischen den unabhängigen Variablen hindeutet. Mit dem jetzt reduziertem Modell erzielt man eine Trefferquote von 68,5%.

Erweiterungen der logistischen Regression

Die logistische Regression ist durch die Verwendung von e-Funktionen ein sehr elegantes und flexibles Verfahren. Sie liefert direkt Schätzungen von Wahrscheinlichkeiten, mittels derer sich Klassifizierungen oder Prognosen vornehmen lassen. Durch die Bildung von Odds und Logits werden vereinfachte Modellformulierungen ermöglicht. Numerisch ist sie etwas schwieriger handhabbar und rechnerisch aufwendiger, als die Diskriminanzanalyse.

Große Bedeutung hat die logistische Regression für die Analyse von Entscheidungsverhalten bei diskreten Alternativen (Discrete Choice Analysis) erlangt.[46] In die dabei verwendeten sog. Logit-Choice-Modelle werden unabhängige Variablen einbezogen, die nicht (nur) über die Fälle (Personen) sondern auch über die Kategorien (Entscheidungsalternativen) variieren. So werden z.B. in der Realität die Kaufwahrscheinlichkeiten für konkurrierende Marken nicht nur von soziodemographischen Merkmalen der Käufer abhängen, sondern vielmehr auch von den Preisen und Eigenschaften der Produkte. Deren Berücksichtigung ermöglicht daher realitätsnahe Modelle, die wichtige Informationen zu liefern vermögen. Durch die Möglichkeit der Spezifizierung von generischen Koeffizienten, die konstant über die Alternativen sind, lassen sich sehr effiziente Modelle bilden. Bei den dabei verwendeten Daten kann es sich um reale Marktdaten handeln (z.B. aus Scanner-Panels)[47], um simuliertes Kaufverhal-

Logit-Choice-Modell

[46] Siehe dazu z.B. Ben-Akiva/Lerman (1985); Urban (1993); Tutz (2000); Hensher/Rose/Greene (2015); Train (2009).
[47] Siehe dazu z.B. Guadagni/Little (1983); Jain/Vilcassim/Chintagunta (1994).

ten in Testmarktsimulationen[48] oder um durch Abfrage von Auswahlentscheidungen (z.B. im Rahmen von Conjoint-Analysen) gewonnene Daten.[49] Das logistische Modell besitzt damit ein sehr breites Anwendungsspektrum und wichtige Bedeutung für Wissenschaft und Praxis.

5.4.4 SPSS-Kommandos

In Abbildung 5.49 ist abschließend die Syntaxdatei mit den SPSS-Kommandos für das Fallbeispiel wiedergegeben.

```
* MVA: Fallbeispiel Multinomiale Logistische Regression.
* DATENDEFINITION.
DATA LIST FIXED
  /Streichf 8 Preis 10  Haltbark 12  Ungefett 14
   Backeign 16  Geschmac 18 Kalorien  20  Tierfett 22
   Vitamin 24  Natur 26  Fall 27-29 Marke 30-32.

* DATENMODIFIKATION
* Definition der Segmente (Gruppen):
*   A: Du darfst, Becel
*   B: Sanella, Homa, SB, Botteram, Flora Soft, Rama
*   C: Delicado, Hollaendische Butter, Weihnachtsbutter.

COMPUTE Segment = Marke.
RECODE  SEGMENT (7,8=1) (1,2,3,9,10,11=2) (4,5,6=3).
VALUE LABELS Segment 1 "Segment A"
                    2 "Segment B"
                    3 "Segment C".

BEGIN DATA
  1    3 3 5 4 1 2 3 1 3 4  1  1
  2    6 6 5 2 2 5 2 1 6 7  3  1
  3    2 3 3 3 2 3 5 1 3 2  4  1
.....................................

127    5 4 4 1 4 4 1 1 1 4 18 11
END DATA.

* PROZEDUR.
* Multinomiale Logistische Regression für den Margarinemarkt.
NOMREG
  Segment WITH Streichf Preis Haltbark Ungefett Backeign Geschmac
   Kalorien Tierfett Vitamin Natur
  /CRITERIA = CIN(95) DELTA(0) MXITER(100) MXSTEP(5) CHKSEP(20)
   LCONVERGE(0) PCONVERGE(1.0E-6) SINGULAR(1.0E-8)
  /MODEL
  /STEPWISE=PIN(.05) POUT(0.1) MINEFFECT(0) RULE(SINGLE)
   ENTRYMETHOD(LR) REMOVALMETHOD(LR)
  /INTERCEPT = INCLUDE
  /PRINT=CLASSTABLE FIT PARAMETER SUMMARY LRT CPS STEP MFI
  /SAVE ESTPROB PREDCAT.

*  Prüfung der Klassifizierung für Segment 1 mit der ROC-Kurve.
ROC EST1  1 BY Segment (1)
  /PLOT=CURVE(REFERENCE)
  /CRITERIA=CUTOFF(INCLUDE) TESTPOS(LARGE)
    DISTRIBUTION(FREE) CI(95)
  /MISSING=EXCLUDE.
```

Abbildung 5.49: SPSS-Kommandodatei für das Fallbeispiel

Im ersten Teil erfolgt die Zusammenfassung von Marken zu Segmenten. Nach den Daten, die hier in der Syntax-Datei enthalten sind, erfolgt der Aufruf der Prozedur NOMREG für die multinomiale logistische Regression. Anschließend erfolgt eine Überprüfung der Klassifizierung für Segment 1 mit der ROC-Kurve. Vergleiche zur Syntax auch die Ausführungen im einleitenden Kapitel dieses Buches.

[48] Siehe dazu z.B. Erichson (2008).
[49] Siehe dazu z.B. Backhaus/Erichson/Weiber (2015), Kapitel 4.

5.5 Anwendungsempfehlungen

Nachfolgend werden einige Empfehlungen für die Durchführung einer logistischen Regression zusammengestellt, die nach den Aspekten Anforderungen an das Datenmaterial, Schätzung der Regressionskoeffizienten und globale Gütemaße differenziert sind.

Anforderungen an das Datenmaterial

- Die logistische Regression stellt relativ geringe statistische Anforderungen an die Daten. Die wichtigste Annahme des logistischen Modells besteht darin, dass die kategoriale abhängige Variable eine Zufallsvariable ist (Bernoulli- oder multinomial-verteilt). Über die Beobachtungen hinweg sollten die abhängigen Variablen unabhängig voneinander verteilt sein.
- Der logistischen Regression ist damit gegenüber der Diskriminanzanalyse der Vorzug zu geben, wenn Unsicherheit über die Verteilung der unabhängigen Variablen besteht, insbesondere wenn kategoriale unabhängige Variable vorhanden sind.
- Für die logistische Regression werden größere Fallzahlen benötigt als z.B. für die lineare Regression oder die Diskriminanzanalyse. Die Fallzahl sollte pro Gruppe (Kategorie der abhängigen Variable) nicht kleiner als 25 sein und damit mindestens 50 betragen.
- Bei größerer Zahl an unabhängigen Variablen sind auch größere Fallzahlen pro Gruppe erforderlich. Es sollten wenigstens 10 Fälle pro zu schätzendem Parameter vorhanden sein.
- Die unabhängigen Variablen sollten weitgehend frei von Multikollinearität sein. (keine linearen Abhängigkeiten, vgl. Kapitel 1)

Schätzung der Regressionskoeffizienten

- Zur Prüfung der Regressionskoeffizienten ist dem Likelihood-Ratio-Test gegenüber dem Wald-Test der Vorzug zu geben, da der Wald-Test bei kleinen Stichproben zu hohe p-Werte liefert. Damit werden eventuell relevante Einflussfaktoren nicht als signifikant erkannt. Der LR-Test für die Koeffizienten wird in SPSS allerdings nur von der Prozedur zur multinomialen logistischen Regression durchgeführt.
- Es ist zu beachten, dass bei einer Kodierung der Ausprägungen der abhängigen Variablen mit Null und Eins die Prozedur zur binären logistischen Regression die Gruppe Null als Referenzkategorie wählt, während jene zur multinomialen logistischen Regression stets die Gruppe mit der höchsten Kodierung –hier die Gruppe Eins– als Referenzkategorie wählt. Die geschätzten Parameter unterscheiden sich dadurch in ihren Vorzeichen, nicht jedoch in ihrem Betrag.

Globale Gütemaße

- Der Likelihood-Ratio-Test zur Beurteilung der Signifikanz des Gesamtmodells ist der beste verfügbare Test und immer anwendbar. Er ist vergleichbar mit dem F-Test der linearen Regressionsanalyse. Andere globale Gütemaße, wie die die

Pearson-Statistik oder die Devianz (Abweichung) sollten bei metrischen unabhängigen Variablen (wenn sich viele Kovariatenmuster ergeben) skeptisch betrachtet werden, da sie in diesen Fällen meist nicht Chi-Quadrat-verteilt sind.

- Niedrige Werte der Pseudo-R-Quadrat-Statistiken sollten nicht zu Enttäuschung Anlass geben, da deren Werte regelmäßig niedriger liegen, als man sie vom Bestimmtheitsmaß R2 der linearen Regression erwartet.
- Bei der Prüfung des Modells mittels einer Klassifizierungstabelle ist zu beachten, dass die Trefferquoten von der Wahl des Trennwertes abhängen. Ein davon unabhängiges Maß für die Güte der Prognose- bzw. Klassifizierungsfähigkeit des Modells bildet die Fläche unter der ROC-Kurve (AUC), die auf Werte zwischen 0 und 1 normiert ist.
- Generell wird eine Ausreißerdiagnostik auf Basis der Pearson-Residuen pro Beobachtung empfohlen. Im Fall eines multinomialen Modells werden sie von SPSS nicht pro Beobachtung, sondern zellenweise berechnet und ausgegeben (Option „Zellwahrscheinlichkeiten“, Abbildung 5.42). Siehe dazu die Ausführungen im mathematischen Anhang.

5.6 Mathematischer Anhang

Um die Berechnung der Pearson-Statistik und der Devianz näher zu erläutern, modifizieren wir die Daten unseres Ausgangsbeispiels in Abbildung 5.3, indem wir die Personen in drei Einkommensklassen gruppieren und die Einkommenswerte in den Klassen durch die drei Mittelwerte ersetzen, die wir in Abschnitt 5.2.1.2 für die gruppierte Analyse verwendet hatten (Abbildung 5.50).

Mit diesen Daten können wir, wie oben unter Abschnitt 5.2.1.4 (Modell 4) eine Individualanalyse vornehmen. Mit den Originaldaten ergaben sich für $LL = -16.053$ (Maximalwert der Log-Likelihood-Funktion) die Schätzwerte

$$a = -5,635, b_1 = 2,351, b_2 = 1,751$$

Für die modifizierten Daten erhält man jetzt den Maximalwert $LL = -18,150$ und die Schätzwerte:

$$a = -3,662, b_1 = 1,462, b_2 = 1,465$$

Bedingt durch den Informationsverlust haben sich die Werte der Koeffizienten etwas verringert. Auch die die Trefferquote sinkt von 83,3% auf 70%.

Durch die Modifikation der Daten hat sich die Anzahl der Merkmalskombinationen (Kovariatenmuster) erheblich verringert, während die Gesamtzahl der Beobachtungen ($K = 30$) gleich geblieben ist. Zuvor waren es 27 Kovariatenmuster (3 Muster waren doppelt). Jetzt sind es nur noch $3 \cdot 2 = 6$ Kovariatenmuster. Die Menge gleicher Kovariatenmuster wird in SPSS als Teilgesamtheit bezeichnet. n_i sei die Anzahl der Fälle in Teilgesamtheit i (Fälle mit Kovariatenmuster i). m_i sei die Anzahl der Ereignisse (Käufe) in Teilgesamtheit i. In Abbildung 5.51 sind die 6 Kovariatenmuster mit den zugehörigen Fall- und Kaufzahlen wiedergegeben. Jede Zeile steht für ein Kovariatenmuster.

Person	Einkommen [Tsd. Euro]	Geschlecht 0=w, 1=m	Kauf 1=ja, 0=nein
1	2.752	0	1
2	2.752	1	0
3	2.752	1	1
4	2.752	0	0
5	2.752	1	1
6	2.752	0	1
7	2.752	0	1
8	2.752	1	1
9	2.752	0	1
10	2.752	0	0
11	2.020	1	1
12	2.020	1	1
13	2.020	0	1
14	2.020	1	1
15	2.020	1	1
16	2.020	0	0
17	2.010	0	0
18	2.020	1	0
19	2.020	0	0
20	2.020	0	0
21	1.573	1	1
22	1.573	1	0
23	1.573	0	1
24	1.573	1	0
25	1.573	0	0
26	1.573	1	0
27	1.573	0	0
28	1.573	1	1
29	1.573	1	1
30	1.573	0	0

Abbildung 5.50: Individualdaten mit Gruppenmittelwerten für das Einkommen

i	Einkommen x_{1i}	Geschlecht x_{2i}	Fälle n_i	Käufe m_i
1	**2.752**	**1**	4	3
2	**2.020**	**1**	5	4
3	**1.573**	**1**	6	3
4	**2.752**	**0**	6	4
5	**2.020**	**0**	5	1
6	**1.573**	**0**	4	1
Summe:			30	16

Abbildung 5.51: Vergleich von geschätzten Wahrscheinlichkeiten

Berechnung der Pearson-Chi-Quadrat-Statistik

Im hier vorliegenden binären Fall lässt sich die Pearson-Chi-Quadrat-Statistik wie folgt berechnet

$$X^2 = \sum_{i=1}^{I} r_i^2 = \sum_{i=1}^{I} \frac{(m_i - n_i p_i)^2}{n_i p_i (1 - p_1)} \tag{5.57}$$

mit

n_i = Fallzahl in Teilgesamtheit $i, K = \sum_{i=1}^{I} n_i$
m_i = Käufe in Teilgesamtheit i,
Häufigkeit, mit der $y_i = 1$ in Teilgesamtheit i beobachtet wurde
p_i = geschätzte Wahrscheinlichkeit für Fälle in Teilgesamtheit i
$p_i = \frac{1}{1+e^{-(a+b_1 x_{1i} + b_2 x_{2i})}}$
I = Anzahl der Teilgesamtheiten (Kovariatenmuster)

Pearson-Residuen

Die Produkte $\hat{m}_i = n_i p_i$ sind die geschätzten (erwarteten) Käufe. Damit lassen sich die sog. *Pearson-Residuen* wie folgt ausdrücken:

$$r_i = \frac{m_i - \hat{m}_i}{\sqrt{n_i p_i (1 - p_i}} \qquad (i = 1, ..., I) \tag{5.58}$$

Die Pearson-Residuen ergeben sich also aus den Differenzen zwischen den beobachteten und den erwarteten Käufen. Die Größe im Nenner ist die Standardabweichung der Binomialverteilung. Wenn die Y_i Bernoulli-verteilt sind, dann sind die m_i binomialverteilt. Für hinreichend große m_i sind die Pearson-Residuen standardnormalverteilt, wenn das Modell korrekt ist.

Die standardisierten Pearson-Residuen gemäß (5.58) können für diagnostische Zwecke verwendet werden. Absolutwerte > 3 sollten untersucht werden. Von SPSS werden im multinomialen Fall die Pearson-Residuen zellenweise (Kovariatenmuster · Response-Kategorien) berechnet und ausgegeben (Option „Zellwahrscheinlichkeiten“, Abbildung 5.42).

i	Fälle n_i	Käufe m_i	p_i	$\hat{m}_i = n_i p_i$	r_i	r_i^2
1	4	3	0,861	3,445	-0,644	0,414
2	5	4	0,680	3,402	0,574	0,329
3	6	3	0,526	3,153	-0,125	0,016
4	6	4	0,589	3,536	0,385	0,148
5	5	1	0,330	1,649	-0,617	0,381
6	4	1	0,204	0,815	0,229	0,053
Summe:	30	16		16	0,0	1,341

Abbildung 5.52: Berechnungstabelle für die Pearson-Statistik (binärer Fall)

Abbildung 5.52 zeigt die Berechnungstabelle für obige Daten. Es ergibt sich hier für die Pearson-Chi-Quadrat-Statistik:

$$X^2 = \sum_{i=1}^{I} r_i^2 = 1,341$$

Unter der Nullhypothese, dass kein Unterschied zwischen den beobachteten und den geschätzten Fallzahlen besteht (perfekte Anpassung), ist sie approximativ Chi-Quadrat-verteilt mit $df = I - (J + 1) = 6 - 3 = 3$ Freiheitsgraden. Man erhält damit den p-Wert 0,719. Der p-Wert ist so groß, dass kein Grund besteht, die Nullhypothese abzulehnen.

Um die Berechnung der Pearson-Chi-Quadrat-Statistik für den multinomialen Fall zu verdeutlichen, wiederholen wir nachfolgend die Berechnung für den binären Fall in etwas umständlicherer Form. Durch Kombination der I Kovariatenmuster und der G Gruppen (Response-Kategorien) lassen sich $I \cdot G = 6 \cdot 2 = 12$ Zellen bilden. Sie sind in Abbildung 5.53 analog zu Abbildung 5.51 mit den zugehörigen Fallzahlen dargestellt. Jede Zeile steht jetzt für eine Zelle. In der rechten Spalte steht jetzt die Anzahl der Beobachtungen (Käufe und Nicht-Käufe) und es gilt:

$$m_{ig} = \text{ Häufigkeit, mit der } Y_i = g \text{ in Zelle ig beobachtet wurde}$$

mit g=1 für Kauf und g=2 für Nicht-Kauf.

Die Fallzahlen summieren sich zu Gesamtzahl der Beobachtungen. Es gilt:

$$n_i = \sum_{g=1}^{G} m_{ig} \text{ und} \qquad K = \sum_{i=1}^{I} n_i$$

i	Einkom. x_{1i}	Gschl. x_{2i}	Gruppe g	Fälle n_{ig}	Beob. m_{ig}
1	2.752	1	1	4	3
2	2.020	1	1	5	4
3	1.573	1	1	6	3
4	2.752	0	1	6	4
5	2.020	0	1	5	1
6	1.573	0	1	4	1
1	2.752	1	2	4	1
2	2.020	1	2	5	1
3	1.573	1	2	6	3
4	2.752	0	2	6	2
5	2.020	0	2	5	4
6	1.573	0	2	4	3
Summe:				60	30

Abbildung 5.53: Zellen mit zugehörigen Fallzahlen

Die Berechnung der Residuen für die Zellen ig erfolgt jetzt anders als oben durch:[50]

$$r_{ig} = \frac{m_{ig} - n_i p_{ig}}{\sqrt{n_i p_{ig}}} \tag{5.59}$$

[50] In dieser Art erfolgt die Berechnung der Pearson-Chi-Quadrat-Statistik in der Kontingenzanalyse (vgl. Kapitel 6). Der Nenner in (5.59) entspricht der Standardabweichung der Poisson-Verteilung. Die Residuen je Zelle gemäß (5.59) sind kleiner als die standardisierten Residuen gemäß (5.58). Sie werden nicht von SPSS ausgegeben.

Man erhält damit die Pearson-Chi-Quadrat-Statistik:

$$X^2 = \sum_{i=1}^{I}\sum_{g=1}^{G} r_{ig}^2 = \sum_{i=1}^{I}\sum_{g=1}^{G} \frac{(m_{ig} - n_i p_{ig})^2}{n_i p_{ig}} = \sum_{i=1}^{I}\sum_{g=1}^{G} \frac{(m_{ig} - \hat{m}_{ig})^2}{\hat{m}_{ig}} \tag{5.60}$$

Die Berechnung (Abbildung 5.54) gemäß Formel (5.52) liefert wiederum den Wert 1,341. Die Freiheitsgrade berechnen sich durch

$$df = I \cdot (G - 1) - \text{Zahl der Parameter}$$

Es ergibt sich wiederum 6. Damit erhalten wir auch wieder den gleichen p-Wert. Im Unterschied zu Formel (5.57) ist Formel (5.60) jetzt auch für den multinomialen Fall mit G $>$ 2 anwendbar.

Aus Abbildung 5.53 oder 5.54 ist ersichtlich, dass hier in allen 12 Zellen wenigsten eine Beobachtung vorliegt. Je größer die Fallzahlen m_{ig}, desto näher kommt die Verteilung der Pearson-Chi-Quadrat-Statistik einer Chi-Quadrat-Verteilung. Ohne die Reduzierung der Einkommenswerte auf drei Klassen hätten wir rund 60 Zellen erhalten, von denen die Hälfte leer geblieben wäre.

i	g	n_{ig}	m_{ig}	p_{ig}	$\hat{m}_{ig}$	r_{ig}	r_{ig}^2
1	1	4	3	0,86	3,44	-0,24	0,06
2	1	5	4	0,68	3,40	0,32	0,11
3	1	6	3	0,53	3,15	-0,09	0,01
4	1	6	4	0,59	3,54	0,25	0,06
5	1	5	1	0,33	1,65	-0,51	0,26
6	1	4	1	0,20	0,82	0,20	0,04
1	2	4	1	0,14	0,56	0,60	0,36
2	2	5	1	0,32	1,60	-0,47	0,22
3	2	6	3	0,47	2,85	0,09	0,01
4	2	6	2	0,41	2,46	-0,30	0,09
5	2	5	4	0,67	3,35	0,35	0,13
6	2	4	3	0,80	3,18	-0,10	0,01
$\sum$		60	30		30		1,341

Abbildung 5.54: Berechnungstabelle für die Pearson-Statistik (12 Zellen)

Berechnung der Devianz

Als Devianz (Abweichung) wird die Differenz zwischen der maximierten Log-Likelihood LL eines zu prüfenden Modells und der Log-Likelihood LLs für ein sog. saturiertes Modell bezeichnet.[51]

$$D = -2(LL - LL_s) \tag{5.61}$$

Das saturierte Modell enthält für jede Beobachtung einen Parameter und ermöglicht damit eine perfekte Anpassung. Wie oben bemerkt, ist das kein gutes Modell, da es

[51] Vgl. McChullagh/Nelder (1989), S. 33 ff.; Agresti (2013), S. 116 ff.; Hosmer/Lemeshow/Sturdivant (2013), S. 12 f.

keine Vereinfachung gegenüber der Realität erbringt. Es soll lediglich als Basis zur Beurteilung der Güte eines zu prüfenden Modells dienen.

Für individuelle Daten wird $LLs = 0$ und damit $D = -2LL$. D ist damit nicht mehr Chi-Quadrat-verteilt. Entsprechend der hier vorhandenen 6 Kovariatenmuster aber lassen sich die Daten gruppieren (in Teilgesamtheiten zerlegen). Nachfolgend soll die Berechnung der Devianz an obigem Beispiel demonstriert werden.

Für gruppierte Daten ist dem logistischen Modell die Binomial-Verteilung zugrunde zu legen. Es sei wie oben

n_i= Anzahl der Fälle in Gruppe i (Fälle mit Kovariatenmuster i)
m_i= Anzahl der Ereignisse (Käufe) in Teilgesamtheit i

mit $$m_i = \sum_{s=1}^{n_i} y_{is} \text{ und } K = \sum_{i=1}^{I} n_i$$

und I = Anzahl der Teilgesamtheiten (Kovariatenmuster)

Wenn die individuellen Ereignisse Y_{is} in den Gruppen i Bernoulli-verteilt sind mit den Wahrscheinlichkeiten π_i, dann ergibt sich für die Summen m_i jeweils eine *Binomial-Verteilung*:

$$P(m_i) = \binom{n_i}{m_i} \pi_i^{m_i} (1-\pi_i)^{n_i - m_i} \tag{5.62}$$

mit dem Erwartungswert $n_i\pi_i$ und der Varianz $n_i\pi_i(1-\pi_i)$.

Die Logarithmierung von (5.49) ergibt:

$$lnP(m_i) = m_i ln(\pi_i) + (n_i - m_i) ln(1-\pi_i) + ln\binom{n_i}{m_i}$$

Für die Bildung der Log-Likelihood-Funktion kann der rechte Summand entfallen, da er unabhängig von der Schätzung der π_i. Für die unbekannten Wahrscheinlichkeiten π_i setzten wir die geschätzten Wahrscheinlichkeit für Fälle in Gruppe i ein:

$$p_i = \frac{1}{1 + e^{-(a + b_i x_{1i} + b_2 x_{2i})}}$$

Durch Summierung über die I Teilgesamtheiten erhält man die folgende Log-Likelihood-Funktion:

$$LL = \sum_{i=1}^{I} m_i ln(p_i) + (n_i - m_i) ln(1-p_i) \tag{5.63}$$

Im rechten Summanden stehen die Fallzahlen für die Nicht-Käufe und deren Wahrscheinlichkeiten. Der Ausdruck lässt sich noch vereinfachen, wenn man über die

$$I \cdot G = 6 \cdot 2 = 12 \text{ Zellen}$$

summiert, da in den Zellen für $G = 2$ hier die Fallzahlen für die Nicht-Käufe stehen. Vergleiche dazu die Berechnungstabelle in Abbildung 5.55. Man erhält dann

$$LL = \sum_{i=1}^{I} \sum_{g=1}^{G} m_{ig} ln(p_{ig}) = -18,150 \tag{5.64}$$

i	g	n_{ig}	m_{ig}	p_{ig}	$\frac{m_{ig}}{n_{ig}}$	LL	LLs
1	1	4	3	0,86	0,75	-0,45	-0,86
2	1	5	4	0,68	0,80	-1,54	-0,89
3	1	6	3	0,53	0,50	-1,93	-2,08
4	1	6	4	0,59	0,67	-2,12	-1,62
5	1	5	1	0,33	0,20	-1,11	-1,61
6	1	4	1	0,20	0,25	-1,59	-1,39
1	2	4	1	0,14	0,25	-1,98	-1,39
2	2	5	1	0,32	0,20	-1,14	-1,61
3	2	6	3	0,47	0,50	-2,24	-2,08
4	2	6	2	0,41	0,33	-1,78	-2,20
5	2	5	4	0,67	0,80	-1,60	-0,89
6	2	4	3	0,80	0,75	-0,68	-0,86
$\sum$	6	60	30	6	6	**-18,150**	**-17,481**

Abbildung 5.55: Berechnungstabelle für die Devianz

Der Wert, der sich jetzt für die gruppierten Daten ergibt, ist identisch mit dem Wert, den wir oben für die Analyse auf Basis der Individualdaten erhalten hatten. Für das saturierte Modell sind die geschätzten Häufigkeiten von Käufen und Nicht-Käufen identisch mit den beobachteten Häufigkeiten m_{ig}. Die Wahrscheinlichkeiten für die Zellen ergeben sich damit durch m_{ig}/n_{ig} und man erhält für die Log-Likelihood-Funktion:

$$LLs = \sum_{i=1}^{I}\sum_{g=1}^{G} m_{ig} ln\left(\frac{m_{ig}}{n_{ig}}\right) = -17,481 \tag{5.65}$$

Für die Devianz erhält man damit

$$D = -2(LL - LLs) = -2(-18,150 + 17,481) = 1,339 \tag{5.66}$$

Durch Zusammenfassung der Formeln (5.64) und (5.65) lässt sich die Berechnung der Devianz weiter vereinfachen zu:[52]

$$D = 2\sum_{i=1}^{I}\sum_{g=1}^{G} m_{ig} ln\left(\frac{m_{ig}}{n_i p_{ig}}\right) = 1,339 \tag{5.67}$$

Man erkennt jetzt kaum noch, dass es sich um eine Devianz handelt. Ihre Berechnung erfolgt damit sehr ähnlich der Berechnung der Pearson-Statistik und sie erbringt auch ganz ähnliche Ergebnisse. Sie ist ebenfalls approximativ Chi-Quadrat-verteilt mit den Freiheitsgraden

$df = I \cdot (G - 1)$– Zahl der Parameter

wie die Pearson-Chi-Quadrat-Statistik. Man erhält damit den p-Wert 0,720.

Aus Formel (5.65) und (5.67) ist ersichtlich, dass eine Berechnung nur möglich ist, wenn die m_{ig} (Zahl der Beobachtungen in Zelle ig) alle größer Null sind, da der Logarithmus für Null nicht definiert ist. Es dürfen also keine Zellen leer sein. Für die Schätzung des logistischen Modells hat dies allerdings keine Bedeutung, da dafür die Bildung der Zellen nicht benötigt wird.

[52] Vgl. IBM Corporation (2013), S. 636.

Literaturhinweise

A. Basisliteratur zur Logistischen Regression

Agresti, A. (2013), Categorical Data Analysis, 3. Auflage, New Jersey.

Hair, J.F./ Black, W.C./ Babin, B.J./ Anderson, R.E. (2010), Multivariate Data Analysis, 7. Auflage, Upper Saddle River (N.J.).

Herrmann, A./ Homburg, C./ Klarmann, M. (Hrsg.) (2008), Handbuch Marktforschung, 3. Auflage, Wiesbaden.

Hosmer, D./Lemeshow, S./Sturdivant, R. (2013), Applied Logistic Regression, 3. Aufl., New York u. a.

Menard, S. (2002), Applied Logistic Regression Analysis, 2. Auflage, Sage University Paper Nr. 106, Thousand Oaks (CA).

Schlittgen, R. (2009), Multivariate Statistik, München.

Tutz, G. (2000), Die Analyse kategorialer Daten, München.

B. Zitierte Literatur

Agresti, A. (2013), Categorical Data Analysis, 3. Auflage, New Jersey.

Backhaus, K./Erichson, B./Weiber, R. (2015), Fortgeschrittene Multivariate Analyseverfahren, 3. Auflage, Berlin/Heidelberg.

Ben-Akiva, M./Lerman, S. (1985), Discrete Choice Analysis, Cambridge, MIT Press.

Büchel, F./Matiaske, W. (1996), Ausbildungsadäquanz bei Berufsanfängern mit Hochschulabschluß, in: *Konjunkturpolitik*, Vol. 42, S. 53–83.

Christensen, B./Papies, D./Proppe, D./Clement, M. (2014), Gütemaße der logistischen Regression bei unbalancierten Stichproben, in: *Wissenschaftliches Studium (WiSt)*, Heft 4, S. 211–213.

Erichson, B. (2008), Testmarktsimulation, in: *Herrmann, A./ Homburg, C. (Hrsg.) (2008): Marktforschung*, 3, S. 983–1001.

Fahrmeir, L./Kneib, T./Lang, S. (2009), Regression – Modelle, Methoden und Anwendungen, 2. Auflage, Springer-Verlag, Berlin u. a.

Fox, J. (2015), Applied Regression Analysis and Generalized Linear Models, 3. Auflage, Los Angeles u. a.

Guadagni, P./Little, J. (1983), A Logit Model of Brand Choice Calibrated on Scanner Data, in: *Marketing Science*, Vol. 2, Nr. 3, S. 203–238.

Hair, J./Black, W./Babin, B./Anderson, R. (2010), Multivariate Data Analysis, 7. Auflage, Englewood Cliffs (N.J.).

Hastie, T./Tibshirani, R./Friedman, J. (2009), The Elements of Statistical Learning, 2. Auflage, New York.

Hensher, D./Rose, J./Greene, W. (2015), Applied Choice Analysis, 2. Auflage, Cambridge University Press, Cambridge et al.

Hosmer, D./Lemeshow, S./Sturdivant, R. (2013), Applied Logistic Regression, 3. Auflage, New York u. a.

IBM Corporation (2013), IBM SPSS Statistics 22 Algorithms, ohne Ort.

Jain, D./Vilcassim, N./Chintagunta, P. (1994), A Random-Coeffcients Logit Brand-Choice Model Applied to Panel Data, in: *J. o. Business & Economic Statistics*, Vol. 13, Nr. 3, S. 317–326.

Lim, T./Loh, W./Shih, Y. (2000), A Comparison of Predicting Accuracy, Complexity, and Training Time of Thirty-three Old and New Classification Algorithms, in: *Machine Learning*, Vol. 40, Nr. 3, S. 203–229.

McChullagh, P./Nelder, J. (1989), Generalized Linear Models, 2. Auflage, Chapman and Hall, /CRC, London.

McFadden, D. (1974), Conditional Logit Analysis of Qualitative Choice Behavior, in: *Zarembka, P. (Ed.): Frontiers in Econometrics*, Vol. 40, S. 105–142.

Melvin, R./Perreault, W. (1977), Validation of Discriminant Analysis in Marketing Research, in: *Journal of Marketing Research*, Vol. 14, Nr. 1, S. 60–68.

Menard, S. (2002), Applied Logistic Regression Analysis, 2. Auflage, in: *Sage-University Paper Band 106*.

Michie, D./Spiegelhalter, D./Taylor, C. (1994), Machine Learning, Neural and Statistical Classification, Ellis Horwood Series in Artificial Intelligence.

Morrison, D. (1969), On the Interpretation of Discriminant Analysis, in: *Journal of Marketing Research*, Vol. 6, Nr. 2, S. 156–163.

Press, W./Flannery, B./Teukolsky, S./Vetterling, W. (2007), Numerical Recipes – The Art of Scientific Computing, 3. Auflage, Cambridge/New York et al.

Train, K. (2009), Discrete Choice Methods with Simulation, 2. Auflage, Cambridge University Press, Cambridge et al.

Tutz, G. (2000), Die Analyse kategorialer Daten, München.

Urban, D. (1993), Logit-Analyse. Statistische Verfahren zur Analyse von Modellen mit qualitativen Response-Variablen, Stuttgart u. a.

Weiber, R./Adler, J. (2003), Der Wechsel von Geschäftsbeziehungen beim Kauf von Nutzungsgütern: Das Beispiel Telekommunikation, in: Rese, M./Söllner/A., Utzig, B. (Hrsg.): Relationship Marketing – Standortbestimmung und Perspektiven, Berlin u. a., S. 71–103.

6 Kreuztabellierung und Kontingenzanalyse

K. Backhaus, B. Erichson, W. Plinke, R. Weiber, *Multivariate Analysemethoden*,
https://doi.org/10.1007/978-3-662-56655-8_7

6.1 Problemstellung

Kreuztabellierung und Kontingenzanalyse dienen dazu, Zusammenhänge zwischen *nominal* skalierten Variablen aufzudecken und zu untersuchen. Typische Anwendungsbeispiele sind die Untersuchung von Zusammenhängen zwischen der Einkommensklasse, dem Beruf oder dem Geschlecht von Personen und ihrem Konsumverhalten oder die Überprüfung der Frage, ob der Bildungsstand oder die Zugehörigkeit zu einer sozialen Klasse einen Einfluss auf die Mitgliedschaft in einer bestimmten politischen Partei hat. Dabei auftretende Fragen können z. B. sein:

- Ist ein Zusammenhang zwischen den Variablen erkennbar und signifikant?
- Gibt es weitere Variablen, durch deren zusätzliche Betrachtung das vorherige Untersuchungsergebnis bestätigt, näher erläutert oder revidiert wird?
- Gibt es die Möglichkeit, eine Aussage über Stärke oder gar Richtung des Zusammenhangs zu treffen?

Das folgende fiktive Beispiel möge das verdeutlichen. Aus der Statistik der Todesursachen von Patienten eines Krankenhauses lässt sich folgende Aufgliederung entnehmen (vgl. Abbildung 6.1):

	Lungenkrebs	andere Ursachen	$\sum$
Raucher	12	55	67
Nichtraucher	8	60	68
$\sum$	20	115	135

Abbildung 6.1: Statistik der Todesursachen eines Krankenhauses (Auszug)

Es fällt auf, dass der Tod durch Lungenkrebs bei Nichtrauchern relativ seltener auftritt als bei Rauchern. Kann man hieraus möglicherweise einen *nicht zufälligen* Zusammenhang ableiten? Eine Antwort auf diese Frage ergibt sich vielleicht aus einer weiteren Variablen, z. B. dem Wohnort oder dem Beruf der Patienten, die in diesem Beispiel nicht erfasst sind. Aus sachlogischen Überlegungen könnte sich in Großstädten möglicherweise eine andere Verteilung ergeben als auf dem Lande.

Die Kreuztabellierung dient dazu, die Ergebnisse einer Erhebung tabellarisch darzustellen und auf diese Art und Weise einen möglichen Zusammenhang zwischen den Variablen zu erkennen. Dabei ist allerdings insbesondere auf eine durch den Sachverhalt begründete Auswahl der Variablen und ihrer Ausprägungen zu achten. Andernfalls besteht die Gefahr, Zusammenhänge willkürlich zu konstruieren oder tatsächlich existierende Abhängigkeiten zu verdecken.

Ist ein Zusammenhang aufgedeckt worden, kann mit Hilfe der Kontingenzanalyse der Frage nachgegangen werden, ob die Assoziation zufällig in der Stichprobe aufgetreten ist oder ob ein systematischer Zusammenhang zugrunde liegt. Das bekannteste Instrument hierzu ist der Chiquadrat-Test (χ^2-Test). In einem weiteren Schritt kann gegebenenfalls überprüft werden, wie stark diese Assoziation ist. Ein möglicher Indikator hierfür ist der Phi-Koeffizient (ϕ).

χ^2-Test

Phi-Koeffizient

Ein Grund für die häufige Anwendung der hier vorgestellten Verfahren liegt in der Möglichkeit, Variable mit unterschiedlichem Skalenniveau in einer gemeinsamen Analyse zu betrachten, da Variable höheren Skalenniveaus immer auf nominale Niveaus

herunter transformiert werden können. Diese Transformation ist allerdings mit einem gewissen Informationsverlust verbunden.

Werden mehr als zwei Variable analysiert, entstehen statt zweidimensionaler mehrdimensionale Tabellen. Zur übersichtlichen Darstellung werden hieraus häufig mehrere zweidimensionale Tabellen gebildet, wobei innerhalb einer Tabelle die Merkmalsausprägung der dritten (oder weiterer Variabler) konstant gehalten wird.[1]

Die Analyse einer Kreuztabelle (Kontingenztabelle[2]) kann in unterschiedlicher Form auftreten, und zwar als *Homogenitätsprüfung* von zwei Stichproben oder als *Unabhängigkeitsprüfung* von zwei Variablen.[3] Diese zwei Formen der Analyse sind mit jeweils unterschiedlichen Formen der Datenerhebung verknüpft.

Homogenitätsprüfung

Bei einer *Homogenitätsprüfung* wird untersucht, ob ein Merkmal in zwei oder mehreren Stichproben identisch verteilt ist.[4] Es liegt folgendes Erhebungsdesign zugrunde. X und Y seien zwei kategoriale Variablen. Entsprechend den Ausprägungen von X werden separate Stichproben gezogen, in denen dann Y erhoben wird.

Beispiel: Es sei X = „Raucher/Nichtraucher“ die Klassifikationsvariable und Y = „Tod durch Lungenkrebs“ die Beobachtungsvariable. Aus den beiden Gruppen Raucher und Nichtraucher werden jeweils 100 Sterbefälle eines Krankenhauses zufällig ausgewählt und anschließend die Todesursache anhand der Aufzeichnungen festgestellt. In Abbildung 6.2 bildet X die Zeilen und Y die Spalten der Kreuztabelle. Es soll jetzt untersucht werden, ob Y in den beiden Stichproben (Zeilen) identisch verteilt ist. Zur Prüfung auf Gleichheit (Homogenität) kann der χ^2-Homogenitätstest verwendet werden.

Eine sehr einfache Analyse ermöglicht in diesem Fall die Betrachtung der *Odds* oder „*Chancen*“ (in diesem Fall spricht man besser von Gefahren).[5] Aus den Daten in Abbildung 6.2 ergeben sich folgende Odds:

- Raucher: 2 zu 8
- Nichtraucher: 1 zu 9

Daraus ergibt sich das Odds-Ratio: $\frac{2/8}{1/9} = 2,25$

Die Gefahr, an Lungenkrebs zu sterben, ist also für Raucher mehr als doppelt so groß wie für Nichtraucher. Ein Zusammenhang zwischen Rauchen und Lungenkrebs ist damit evident.

[1] Weitere Möglichkeiten sind die Bildung von Mittelwerten oder Verhältniszahlen. Vgl. Zeisel (1970), Kap.V. Darüber hinaus wurden in den letzten Jahrzehnten weitere Verfahren zur Analyse von mehrdimensionalen Tabellen wie die loglinearen Modelle oder die Konfigurationsfrequenzanalyse entwickelt. Diese Ansätze bieten dem Forscher weitergehende Untersuchungsmöglichkeiten über die reine Unabhängigkeitshypothese hinaus, wie z. B. die Untersuchung des Einflusses einiger Variablen auf einige andere oder die Bestimmung der sogenannten second-order-Effekte. Eine Darstellung würde allerdings den hier gesetzten Rahmen sprengen. Vgl. den Literaturüberblick in Fahrmeier, L./Hamerle, A./Tutz, G. (Hrsg.) (1996).

[2] Der Name stammt von dem englischen Statistiker Karl Pearson (1857–1936), der auch die Chiquadrat-Statistik entwickelt hat.

[3] Vgl. Lienert (1973), S. 386-391; Hartung (2009), S. 425 ff.; aber auch die Ausführungen zur Stichprobenerhebung und die Auswirkungen auf die Güte des Tests bei Fleiss/Levin/Paik (2003), S. 50-55; ähnlich Kendall/Stuart (1979), S. 580-585.

[4] Diese Idee ist dem Leser von den χ^2-Anpassungstests vielleicht bekannt, bei denen eine empirische Verteilung auf Gleichheit mit einer theoretischen Verteilung getestet wird.

[5] Siehe dazu die Ausführungen im vorhergehenden Kapitel.

	Lungenkrebs	andere Ursachen	$\sum$
Raucher	20	80	100
Nichtraucher	10	90	100
$\sum$	30	170	200

Abbildung 6.2: Datensatz für Homogenitätsprüfung

Unabhängigkeits-prüfung

In den meisten Fällen liegt den Daten einer Kontingenztabelle nur eine einfache Stichprobe zugrunde. Dazu werden zufällig aus einer Grundgesamtheit Probanden ausgewählt und bei jedem Probanden werden jeweils zwei oder mehr Merkmale erhoben, die sodann einer Unabhängigkeitsprüfung unterzogen werden können.

Beispiel: Es interessiert wiederum die Frage, ob zwischen der Todesursache Lungenkrebs und dem Rauchen ein Zusammenhang besteht. Zu diesem Zweck werden 200 tödlich verlaufene Krankheitsgeschichten zufällig aus der Statistik eines Krankenhauses gezogen. Bei jedem Patienten werden dann die zwei Merkmale X = Rauchverhalten und Y = Todesursache gleichzeitig erhoben (vgl. Abbildung 6.3). Bei dieser Datenstruktur ist der Zusammenhang zwischen Rauchen und Lungenkrebs nicht so augenfällig. Zum Nachweis eines statistischen Zusammenhangs der beiden Merkmale X und Y kann hier ein χ^2-Unabhängigkeitstest verwendet werden.

	Lungenkrebs	andere Ursachen	$\sum$
Raucher	18	63	81
Nichtraucher	12	107	119
$\sum$	30	170	200

Abbildung 6.3: Datensatz für eine Kontingenzanalyse

Fall-Kontroll-Studie

Das erste Erhebungsdesign bietet sich an, wenn eine der Merkmalsausprägungen in der Grundgesamtheit sehr selten vorkommt. Bei Ziehung einer einzigen Stichprobe wäre dann mit einer sehr kleinen Fallzahl dieser Gruppe in der Stichprobe zu rechnen. Um das zu vermeiden, wird für diese Merkmalsausprägung eine separate Stichprobe erhoben. Man spricht bei diesem Design auch von einer Fall-Kontroll-Studie(case-control study).[6]

Der χ^2-Homogenitätstest und der χ^2-Unabhängigkeitstest bilden letztlich nur unterschiedliche Betrachtungen des gleichen Sachverhaltes und führen immer zum gleichen Ergebnis. Die Teststatistik χ^2 ist in beiden Fällen dieselbe.

Sind die Häufigkeiten von Y für unterschiedliche Ausprägungen von X bis auf zufällige Schwankungen gleich (homogen), so besteht auch kein Grund zur Annahme eines Zusammenhangs zwischen den Variablen X und Y. Unterscheiden sich dagegen die Häufigkeiten von Y deutlich für unterschiedliche Ausprägungen von X, so kann auf einen Zusammenhang zwischen X und Y geschlossen werden. Mittels eines statistischen Tests (χ^2-Test) kann dann die Signifikanz des Zusammenhangs geprüft werden.

[6] Fall-Kontroll-Studien werden besonders häufig in der Epidemiologie, Medizin oder Biologie angewendet, wenn es um seltenen Krankheiten oder Phänomene geht. Nachteilig ist, dass sich aus den Häufigkeiten der Kontingenztabelle nicht auf die Häufigkeiten in der Grundgesamtheit schließen lässt, da die Größen der Teilstichproben durch den Untersucher vorgegeben werden.

Welches von zwei untersuchten Merkmalen X und welches Y ist, hat (im Unterschied zur Regressionsanalyse) auf das Ergebnis einer Kontingenzanalyse keinen Einfluss, sondern ist nur für die Interpretation der Ergebnisse von Bedeutung. Zeilen und Spalten einer Kontingenztabelle können also vertauscht werden. Üblicherweise wird mit X eine mutmaßlich unabhängige (ursächliche) Variable bezeichnet mit Y eine abhängige Variable, an der sich die Wirkungen von X zeigen, oder formal ausgedrückt:

$$X \rightarrow Y$$

Bei nichtexperimentellen Studien (Beobachtungsstudien) ist der Schluss von statistischen Zusammenhängen auf kausale Wirkungsbeziehungen immer mit einem hohen Irrtumsrisiko verbunden. Aus einem statistischen Zusammenhang allein lässt sich nicht auf einen Kausalzusammenhang schließen. Vielmehr bildet der statistische Zusammenhang zwischen X und Y nur eine notwendige Bedingung für einen kausalen Zusammenhang. Es müssen immer noch sachlogische Überlegungen hinzukommen, mittels der sich andere Ursachen für den beobachteten Zusammenhang ausschließen lassen. Im vorliegenden Fall aber besteht eine hohe Plausibilität für die Schlussfolgerung, dass der festgestellte Zusammenhang auch kausal bedingt ist, und auch bezüglich der Richtung des Kausalzusammenhangs bestehen hier kaum Zweifel. Es erscheint logisch, dass das Rauchen die Ursache für den Lungenkrebs bildet und nicht umgekehrt, insbesondere auch deshalb, weil das Rauchen meist schon vor der Erkrankung erfolgte (die Ursache geschieht vor der Wirkung).

Kausalzusammenhang

Abbildung 6.4 zeigt typische Anwendungsbeispiele der Kontingenzanalyse.

	Fragestellung	Variable 1	Variable 2
1.	Gibt es einen Zusammenhang von Studienabbruch und Nebenerwerbstätigkeit von Studenten?	Studienabbruch: Abgang von der Hochschule ohne Abschluss	Berufstätigkeit: unter 15 Std. pro Woche, 15-30 Std. pro Woche, mehr als 30 Std. pro Woche
2.	Ist das Krankheitsbild der Depression bei Selbstmördern häufiger vorzufinden als bei anderen Todesursachen?	Selbstmord: ja/nein	Depression: nach ärztlichem Gutachten schwach ausgeprägt, mittel ausgeprägt, hoch ausgeprägt
3.	Sind einem Testmarkt unterzogene Produkte erfolgreicher als nicht getestete?	Erfolg der Markteinführung: Rücknahme des Produktes aus dem Markt innerhalb 6 Monaten nach Einführung	Testmarktdurchführung: ja/nein
4.	Haben international tätige Konzerne eine andere Organisationsstruktur als national tätige?	Konzernstruktur: divisional, funktional, Matrix	Internationale Tätigkeit: ja/nein
5.	Gibt es einen Zusammenhang zwischen Beruf und Herzinfarkt?	Angestellter, Arbeiter, Beamter, Selbständiger, Unternehmer	Herzinfarkt: ja/nein

Abbildung 6.4: Typische Anwendungsbeispiele der Kontingenzanalyse

6.2 Vorgehensweise

Im Folgenden verdeutlichen wir das Grundprinzip der Kontingenzanalyse am Beispiel von zwei nominalskalierten Variablen mit jeweils mehreren Ausprägungen. Dabei folgen wir einer Vorgehensweise in drei Schritten, die in Abbildung 6.5 dargestellt sind.

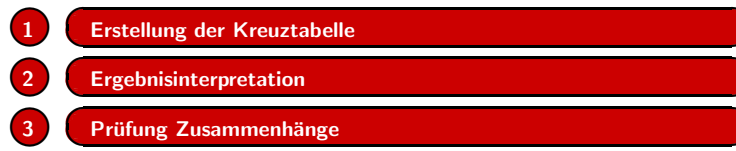


Abbildung 6.5: Ablaufschritte der Kontingenzanalyse

Kontingenztafeln

Die Methoden zur Untersuchung einer zweidimensionalen Kreuztabelle lassen sich auch auf *mehrdimensionale Kontingenztafeln* übertragen. Auch im mehrdimensionalen Fall ist die statistische Abhängigkeit der Variablen durch eine sog. Chi-Quadrat-Statistik prüfbar, die auf den Differenzen zwischen beobachteten und erwarteten Werten beruht. Jedoch sind zwischen den einzelnen Variablen unterschiedliche Abhängigkeiten denkbar. So könnte im Fall der dreidimensionalen Tafel eine Variable unabhängig von zwei anderen sein, die voneinander abhängig sind, oder die Abhängigkeit von zwei Variablen könnte sich in Abhängigkeit von den Ausprägungen der dritten Variablen unterschiedlich darstellen. Da sich diese Fragestellungen mit zum zweidimensionalen Fall analogen Überlegungen überprüfen lassen[7], beschränken sich die nachfolgenden Betrachtungen auf die Untersuchung von zwei Variablen.

Es sei an dieser Stelle darauf hingewiesen, dass sich in den letzten Jahren weitere Verfahren zur Untersuchung solcher Sachverhalte durchgesetzt haben. Durch eine der Varianzanalyse ähnelnde Modelldarstellung, in der die Effekte der einzelnen Merkmalsstufen additiv die beobachteten Zellhäufigkeiten erklären, sind die sogenannten log-linearen Modelle[8] in der Lage, nicht nur die Frage einer Unabhängigkeit von mehreren nominal skalierten Variablen zu klären, sondern auch Schätzer für die Stärke der Einzeleffekte zu bestimmen. Bei der Betrachtung der Einflüsse von mehreren „unabhängigen“ Variablen auf eine dichotome „abhängige“ Variable können sog. Logit-Modelle[9] zur Analyse herangezogen werden (vgl. hierzu auch Kapitel 5 „Logistische Regressionsanalyse“ in diesem Buch). Eine weitere Methode zur Analyse von mehrdimensionalen Kontingenztafeln bietet die Korrespondenzanalyse (vgl. hierzu Kapitel 7 in Backhaus/Erichson/Weiber (2015), Fortgeschrittene Multivariate Analysemethoden), die eine graphische Darstellung der Zusammenhänge zwischen den Variablen ermöglicht.

[7] Siehe Everitt (1992), S. 60 ff.
[8] Siehe hierzu z. B. Agresti (2007), Kapitel 6; Fahrmeier, L./Hamerle, A./Tutz, G. (Hrsg.) (1996); Everitt (1992), S. 73 ff.; Bishop/Fienberg/Holland (2007). Vgl. auch die SPSS-Prozeduren HILOGLINEAR; LOGLINEAR.
[9] Eine ausführliche Darstellung bietet Haberman (1978), S. 292-353.

6.2.1 Erstellung der Kreuztabelle

Zur Untersuchung zweier nominalskalierter Variabler mit jeweils mehreren Ausprägungen wird zunächst eine zweidimensionale *Kreuztabelle* gebildet. Es wird die Gesamtzahl n_{ij} an Beobachtungen einer bestimmten Merkmalskombination (i-te Ausprägung der ersten Variablen ($i = 1, .., I$) und j-te Ausprägung der zweiten Variablen ($j = 1, .., J$)) bestimmt und in eine Tabelle eingetragen. Dabei bilden die I möglichen Merkmalsausprägungen der einen Variablen die verschiedenen Zeilen der Tabelle, die Ausprägungen der anderen Variablen die J verschiedenen Spalten. Aus der Anzahl der möglichen Merkmalskombinationen ergibt sich auch die Bezeichnung „IxJ-Kreuztabelle“ (Abbildung 6.6).

Randsummen

Die Randsummen (Zeilen- oder Spaltensummen) geben jeweils die Gesamtzahl der Beobachtungen einer bestimmten Merkmalsausprägung. n bezeichnet die Gesamtzahl aller Beobachtungen. Um die Kreuztabelle besser analysieren und interpretieren zu können, werden häufig statt absoluter Werte Prozentwerte, bezogen auf verschiedene Basen, in die Tabelle eingetragen.

4-Felder-Tafel

Betrachten wir zunächst den einfachen Fall zweier Merkmale, die jeweils nur zwei Ausprägungen annehmen können (binäre Variable). Nehmen wir an, dass eine Handelskette für die Planung ihrer Logistik wissen will, ob die Wohnlage im Zusammenhang mit der Verwendung von Butter bzw. Margarine als bevorzugtem Brotaufstrich steht. Zur Klärung der Frage werden zufällig 181 Personen ausgewählt und nach ihrem bevorzugten Brotaufstrich und ihrem Wohnort gefragt. Zur Untersuchung und Darstellung des Befragungsergebnisses verwenden wir die einfachste aller Kreuztabellen, die 2x2- oder auch 4-Felder-Tafel (Vgl. Abbildung 6.7).

IxJ Kreuztabelle	Merkmal 2					
	Ausprägung					Zeilen- oder Randsumme
Merkmal 1	1	2	..	..	J	
Ausprägung 1	n_{11}	n_{12}				$n_{1.}$
Ausprägung 2	n_{21}	n_{22}				$n_{2.}$
Ausprägung 3				...		
Ausprägung I	n_{I1}				n_{IJ}	$n_{I.}$
Spalten- oder Randsumme	$n_{.1}$	$n_{.2}$			$n_{.J}$	n

Abbildung 6.6: IxJ-Kreuztabelle

	Bevorzugter Brotaufstrich		
Wohnort	Margarine	Butter	$\sum$
ländlich	23	45	68
städtisch	83	30	113
$\sum$	106	75	181

Abbildung 6.7: Analyse der Produktpräferenzen (181 Einkaufsvorgänge)

6.2.2 Ergebnisinterpretation

Tabellierung

Zur besseren Übersichtlichkeit des obigen Sachverhaltes werden die absoluten Werte in Prozentzahlen transformiert. Üblicherweise finden drei verschiedene Darstellungen Verwendung. Die Wahl einer geeigneten Tabellierung kann in der Regel erst durch die konkrete Fragestellung entschieden werden. Zu unterscheiden sind

a) Zeilenprozente (andere literaturübliche Bezeichnung: Quer- oder auch Horizontalprozentuierung)

b) Spaltenprozente (Längs- oder Vertikalprozentuierung)

c) Totalprozente.

Die jeweiligen Ergebnisse zeigen die Abbildungen 6.8 bis 6.10. In Abbildung 6.8 beziehen sich die Prozentangaben auf die Spaltensumme, d. h. auf die Beobachtungsgesamtzahl einer Merkmalsausprägung der zweiten Variablen „Bevorzugter Brotaufstrich“ (Spaltenprozente). In Abbildung 6.9 bildet die Zeilensumme die Basis für die Prozentberechnung (Zeilenprozente) und in Abbildung 6.10 ist es die Gesamtzahl aller Beobachtungen (Totalprozente).

	Bevorzugter Brotaufstrich	
Wohnort	Margarine	Butter
ländlich	21,7 %	60 %
städtisch	78,3 %	40 %
$\sum$	100 %	100 %

Abbildung 6.8: Analyse der Produktpräferenzen (181 Einkaufsvorgänge) Darstellung mit Spaltenprozenten

	Bevorzugter Brotaufstrich		
Wohnort	Margarine	Butter	$\sum$
ländlich	33,8 %	66,2 %	100 %
städtisch	73,5 %	26,5 %	100 %

Abbildung 6.9: Analyse der Produktpräferenzen (181 Einkaufsvorgänge) Darstellung mit Zeilenprozenten

Fragestellungs-Bezug

Jede dieser Darstellungen liefert andere Informationen. Daher ist die Auswahl der geeigneten Tabellierung immer abhängig von der konkreten Fragestellung.

Für unser Beispiel heißt das: Will die Handelskette verstärkt Margarine vertreiben und daher Filialen gezielt beliefern, so ist es für sie wichtig, welche Filialen überproportional viele Margarinekäufer haben, damit sie ihre Absatzbemühungen auf diese Filialen konzentrieren kann. Daher ist die Darstellung in Abbildung 6.8 aussagekräftig. Dort kann man erkennen, dass der weitaus größere Teil der Verwender von Margarine in der Stadt lebt.

	Bevorzugter Brotaufstrich		
Wohnort	Margarine	Butter	$\sum$
ländlich	12,7 %	24,9 %	37,6 %
städtisch	45,9 %	16,6 %	62,4 %
$\sum$	58,6 %	41,4 %	100 %

Abbildung 6.10: Analyse der Produktpräferenzen (181 Einkaufsvorgänge) Darstellung mit Totalprozenten

Für den Filialleiter eines Supermarktes ist hingegen die Fragestellung eine andere. Ihn dürfte interessieren, ob seine Kunden Unterschiede hinsichtlich der Nachfrage nach Butter bzw. Margarine aufweisen, um so für seinen Standort die entsprechende Sortimentspolitik zu planen. Daher wäre hier die Darstellung in Abbildung 6.9 interessant. Dort ist zu erkennen, dass z. B. Bewohner aus ländlichen Gegenden überwiegend Butter nachfragen. Die Darstellungsform in Abbildung 6.10 gibt einen generellen Überblick darüber, wie Wohnlage und Sortenpräferenz in der Stichprobe zusammenhängen.

Aus der Rohdatenbasis in Abbildung 6.7 ergeben sich auf den ersten Blick deutliche Hinweise darauf, dass die Wohngegend und die Bevorzugung eines Brotaufstrichs nicht voneinander unabhängig sind. So wohnt fast jede dritte befragte Person der Stichprobe auf dem Lande, während es bei den Butterliebhabern mindestens jede zweite war und bei den Margarineverwendern nur etwa jeder fünfte. Wären die Variablen unabhängig, würden wir eine ungefähr gleiche Verteilung der Merkmalsausprägungen in allen Spalten erwarten und ebenso eine gleiche Verteilung der Merkmalsausprägungen in allen Zeilen.

An dieser Stelle ist allerdings darauf hinzuweisen, dass ungleiche Verteilungen allein nicht ausreichen, um hieraus einen Zusammenhang zu folgern. So ist es durchaus möglich, dass durch die Einbeziehung einer dritten Variablen in die Untersuchung eine getroffene Beurteilung revidiert werden muss. Im Falle eines vermuteten Zusammenhangs kann dieser in seiner ursprünglichen Art bestätigt oder als andersartig erkannt, er kann allerdings auch als scheinbarer Zusammenhang aufgedeckt werden. Umgekehrt kann ein fehlender (nicht erkennbarer) Zusammenhang zweier Variabler durch Berücksichtigung einer dritten Variablen als nicht existierend bestätigt oder aber als lediglich bisher verdeckt entlarvt werden.

Beispiel

Beispiel: In Abbildung 6.11 ist das Untersuchungsergebnis einer Erhebung zur Auswirkung des Familienstandes auf den Kauf von Diätprodukten dargestellt. Bei der Untersuchung wurden 132 verheiratete und 158 ledige Personen danach befragt, ob sie Diätprodukte verwenden.

	Verwendet Diätprodukte		Gesamt
Familienstand	ja	nein	
verheiratet	30 (23 %)	102 (77 %)	132 (100 %)
ledig	100 (63 %)	58 (37 %)	158 (100 %)

Abbildung 6.11: Zusammenhang zwischen Familienstand und der Verwendung von Diätmargarine (n = 290)

Erkennbare Zusammenhänge

Die Darstellung erfolgt in absoluten Werten und zusätzlich in Zeilenprozenten (in Klammern), weil die vorrangige Fragestellung die nach Unterschieden im Kaufverhalten von ledigen und verheirateten Personen ist. Es ist ein Zusammenhang zwischen dem Kauf von Diätprodukten und dem Familienstand erkennbar; die Verhältnisse innerhalb der Merkmalsausprägungen „verheiratet“ und „ledig“ sind deutlich unterschiedlich. Verheiratete Personen scheinen Diätprodukten gegenüber skeptischer zu sein.

Wird die Stichprobe allerdings nach dem Alter in zwei Untergruppen aufgeteilt, entsteht das in den Abbildungen 6.12 und 6.13 aufgeführte Bild:

	Verwendet Diätprodukte		Gesamt
Familienstand	ja	nein	
verheiratet	10 (83 %)	2 (17 %)	12 (100 %)
ledig	90 (81 %)	21 (19 %)	111 (100 %)

Abbildung 6.12: Untergruppe der unter 35-jährigen (n = 123)

	Verwendet Diätprodukte		Gesamt
Familienstand	ja	nein	
verheiratet	20 (13 %)	100 (83 %)	120 (100 %)
ledig	10 (21 %)	37 (79 %)	47 (100 %)

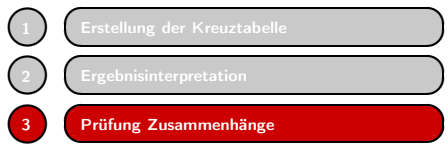
Abbildung 6.13: Untergruppe der über 35-jährigen (n = 167)

Jetzt ist jeweils in jeder Ausprägung der Variablen Familienstand das Verhältnis zwischen Verwendern und Nichtverwendern von Diätprodukten in etwa gleich. In unserer Experimentgruppe ist also das Alter eine Variable, die einen Einfluss auf die Diätproduktnachfrage hat. Jüngere Leute fragen offenbar verstärkt Diät-Produkte nach, ältere deutlich weniger. Da der Familienstand aber in der Regel mit dem Alter zusammenhängt, ist auf den ersten Blick der Eindruck entstanden, als ob der Familienstand ein Indikator dafür wäre, ob Personen Diätprodukte verwenden oder nicht.

Durch die Einbeziehung der dritten Variablen ist der erkannte Zusammenhang nicht widerlegt worden, sondern er wurde modifiziert und konnte so besser verstanden und erklärt werden. Grundsätzlich ist es möglich, durch die Einbeziehung einer dritten Variablen die bisherige Schlussfolgerung, sei es nun die der Existenz oder die des Nichtvorhandensein eines Zusammenhangs, zu bestätigen oder zu ändern.[10] Daher ist es für die verantwortungsvolle Interpretation einer Untersuchung äußerst wichtig, schon bei der Variablenauswahl darauf zu achten, mögliche weitere Einflussfaktoren zu berücksichtigen.

[10] Eine ausführliche Darstellung aller Möglichkeiten mit Beispielen findet sich bei Iacobucci/Churchill Jr. (2015) oder Böhler/Fürst (2014).

6.2.3 Prüfung der Zusammenhänge



Nachdem die Vermutung eines Zusammenhanges durch die Kreuztabellierung gestützt wird, kann mit Hilfe statistischer Verfahren (Tests) geprüft werden, ob dieser Tatbestand nur zufällig in der Stichprobe auftrat oder sich auf die Grundgesamtheit übertragen lässt. Die Methode, die dazu herangezogen wird, ist der χ^2-Test, der im Folgenden einer genaueren Betrachtung unterzogen wird. Allerdings liefert der χ^2-Unabhängigkeitstest keine Anhaltspunkte zur *Stärke* des Zusammenhangs zwischen den Variablen. Hierzu können weitere statistische Maße herangezogen werden, die im zweiten Unterabschnitt behandelt werden.

χ^2-Test

6.2.3.1 Prüfung der statistischen Unabhängigkeit

Betrachten wir wieder das in Abbildung 6.7 dargestellte Untersuchungsergebnis und unterstellen, dass die Erkenntnisse durch die Einbeziehung weiterer Variablen der Vermutung eines Zusammenhangs zwischen Wohngegend und dem Kaufverhalten bzgl. Butter/Margarine nicht widersprechen. Folgende heuristische Überlegung kann eine erste Antwort auf die Frage bieten, ob die Bevorzugung von Butter oder Margarine unabhängig oder abhängig von der Wohngegend des Käufers ist. Aus Abbildung 6.7 ist zu entnehmen, dass 106 von 181 Befragten Margarine bevorzugen. Ebenso ist abzulesen, dass 68 Probanden in ländlicher Wohngegend leben. Gemäß der Annahme, dass beide Merkmale unabhängig voneinander sind, muss man erwarten, dass das Verhältnis von Landbewohner zu Stadtbewohner in der Gesamtstichprobe dem Verhältnis in den Untergruppen der Käufer bzw. Nichtkäufer von Margarine entspricht.

Erwartung und Realität

Überprüfen wir das: Aus Abbildung 6.10 wissen wir, dass 37,6 % aller Befragten auf dem Lande leben. Bei unterstellter Unabhängigkeit[11] müssten also auch 37,6 % der Margarineverwender, also etwa (0,376 · 106 =) 40 Personen, bzw. 37,6 % der Butterverwender, also 28 (= 0,376 · 75) Personen auf dem Lande leben. In unserer Untersuchung haben wir jedoch völlig andere Zahlen erhalten. Dort beobachteten wir lediglich 23 anstatt nach obiger Überlegung erwarteten 40 Personen, die auf dem Lande leben und Margarine kaufen. Anstatt der erwarteten 28 Personen in unserer Stichprobe mit den Merkmalsausprägungen Butterverwender und Landbewohner beobachteten wir insgesamt 45. Analoge Berechnungen und Vergleiche kann man nun für alle auftretenden Kombinationen von Merkmalsausprägungen durchführen. Als Faustformel für die Berechnung der erwarteten absoluten Werte gilt dabei jeweils:

$$\text{Erwarteter Wert} = \frac{\text{Zeilensumme} \cdot \text{Spaltensumme}}{\text{Gesamtsumme}}$$

Insgesamt ergeben sich jeweils mehr oder minder große Abweichungen. Diese können wir als ein Maß zur Überprüfung der eingangs unterstellten Hypothese der Unabhängigkeit der Merkmale auffassen. Nach diesem Prinzip arbeitet auch der χ^2-Test. Der χ^2-Test ist ein Test zur Überprüfung der Unabhängigkeit zweier Merkmale bzw. der Homogenität eines Merkmals in zwei Stichproben. Die statistischen Hypothesen lauten:

[11] Bei der stochastischen Unabhängigkeit zweier Ereignisse A und B gilt für die Bestimmung der Wahrscheinlichkeit des gemeinsamen Eintretens von A und B: $p(A \cap B) = p(A) \cdot p(B)$.

H_0: X und Y sind voneinander unabhängig.

bzw. im Fall der Überprüfung der Verteilung eines Merkmals in zwei unabhängigen Stichproben:

H_0: Der Anteil jeder Merkmalsausprägung der Variablen X ist in beiden Stichproben gleich.[12]

Wir können uns hier allerdings auf den Fall der Kontingenzanalyse beschränken, da die methodische Vorgehensweise zur Homogenitätsprüfung identisch ist. Die Testgröße des χ^2-Tests leiten wir wie folgt ab. (Wir verwenden die Bezeichnungen aus der Abbildung 6.6).

Nullhypothese

Bei insgesamt beobachteten $n_{i.}$ Probanden mit i-ter Merkmalsausprägung der ersten Variablen und $n_{.j}$ Beobachtungen der Ausprägung j der zweiten Variablen erwarten wir unter der Nullhypothese, dass in unserer Stichprobe $e_{ij} = n_{i.} \cdot n_{.j}/n$ Personen gleichzeitig die j-te Ausprägung in der zweiten Variablen und i-te Ausprägung beim ersten Merkmal aufweisen. Die Differenz zwischen der erwarteten Anzahl e_{ij} und der beobachteten Anzahl n_{ij} ist ein erster Hinweis darauf, ob die Merkmale unabhängig sind oder nicht. Je kleiner die Differenz, desto mehr spricht für die Unabhängigkeit; bzw. je größer die Differenz, desto eher scheint die Nullhypothese der Unabhängigkeit der Merkmale nicht zu stimmen. Die Testgröße des χ^2-Tests berücksichtigt alle Abweichungen, indem sie die Gesamtsumme bildet. Um zu verhindern, dass Abweichungen nach oben und unten sich gegenseitig aufheben, wird allerdings jedes Mal das Quadrat der Differenz verwendet. Die Division jedes Summanden durch die erwartete Anzahl hat zur Folge, dass gleiche Abweichungen in Abhängigkeit von der absoluten Größe der erwarteten Werte unterschiedlich gewichtet (normiert) werden. Die Teststatistik des χ^2-Test lautet demnach:

Teststatistik

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(n_{ij} - e_{ij})^2}{e_{ij}} \tag{6.1}$$

Führen wir den χ^2-Test für unser Beispiel durch. Wir überprüfen die Nullhypothese

H_0: Die bevorzugte Verwendung von Butter/Margarine und die Wohnlage sind unabhängig.

Als Testniveau wählen wir 5 %. Mit Hilfe der Abbildung 6.7 ergeben sich folgende Rechenschritte:

$$\begin{aligned}
e_{11} &= n_{1.} \cdot n_{.1}/n = 68 \cdot 106/181 = 39,8 \\
e_{12} &= n_{1.} \cdot n_{.2}/n = 68 \cdot 75/181 = 28,2 \\
e_{21} &= n_{2.} \cdot n_{.1}/n = 113 \cdot 106/181 = 66,2 \\
e_{22} &= n_{2.} \cdot n_{.2}/n = 113 \cdot 75/181 = 46,8
\end{aligned}$$

Für die Testgröße erhält man damit:

$$\begin{aligned}
\chi^2 = &(23 - 39,8)^2/39,8 + (45 - 28,2)^2/28,2 + (83 - 66,2)^2/66,2 \\
&+ (30 - 46,8)^2/46,8 = 27,4
\end{aligned}$$

[12] Es ist bei 4-Felder-Tafeln auch möglich, einen Test durchzuführen, um zu überprüfen, ob der Anteil der Merkmalsträger in der einen Stichprobe größer (oder kleiner) als in der anderen Stichprobe ist. Zur Vorgehensweise bei solchen einseitigen Test vgl. Fleiss/Levin/Paik (2003), S. 58-60.

Wie an den Beispielzahlen zu erkennen ist, gilt bei 4-Felder-Tafeln immer, dass die Differenz zwischen beobachteten und erwarteten Werten gleich ist und daher nur einmal berechnet werden muss. Durch einige Umformungen ist daher eine einfachere Möglichkeit zur Bestimmung der Größe χ^2 im 4-Felder-Fall (und nur dort) möglich:

$$\chi^2 = n \cdot (n_{11} \cdot n_{22} - n_{12} \cdot n_{21})^2 / (n_{1.} \cdot n_{.1} \cdot n_{2.} \cdot n_{.2})$$

Durch Einsetzen ergibt sich (bis auf Rundungsfehler) das gleiche Ergebnis wie oben:

$$\chi^2 = 181 \cdot 9272025 / 61087800 = 27,47$$

Die Statistik χ^2 ist unter der Nullhypothese (approximativ) χ^2-verteilt mit $(I-1) \cdot (J-1)$ Freiheitsgraden. Überschreitet die Teststatistik einen dem Signifikanzniveau entsprechenden Wert der χ^2-Tabelle (vgl. Anhang A.4), so ist die Nullhypothese, die Annahme der Unabhängigkeit der Merkmale, mit der vorher festgelegten Irrtumswahrscheinlichkeit zu verwerfen.

Anhand der χ^2-Tabelle im Anhang A.4 bestimmt sich bei einem vorgegebenen Signifikanzniveau von 5 % und $(I-1) \cdot (J-1) = 1$ Freiheitsgraden der Vergleichswert als 3,84. Der Vergleich ergibt:

$$\chi^2 = 27,47 > 3,84$$

Daher kann die Nullhypothese mit einer Irrtumswahrscheinlichkeit von 5 % abgelehnt werden.

Die Testgröße χ^2 ist unter H_0 streng genommen nur approximativ χ^2-verteilt.[13] Bei kleinen Stichprobenumfängen ist diese Approximation nicht befriedigend. Zur Verbesserung bietet sich zum einen die korrigierte Teststatistik nach Yates oder der exakte Fisher-Test an. Die Yates-Korrektur lautet:

Yates-Korrektur

$$\chi^2_{korr} = \frac{n \cdot (|n_{11} \cdot n_{22} - n_{12} \cdot n_{21}| - n/2)^2}{n_{1.} \cdot n_{.1} \cdot n_{2.} \cdot n_{.2}}$$

Der Wert der Teststatistik χ^2_{korr} ist ebenfalls mit dem kritischen Wert aus der χ^2-Verteilung zu vergleichen. Für unser Beispiel ergibt sich:

$$\chi^2_{korr} = \frac{181 \cdot (|690 - 3735| - 90,5)^2}{68 \cdot 113 \cdot 75 \cdot 106} = 25,86$$

und daher ebenfalls die Ablehnung der Nullhypothese. Die Ablehnung der Nullhypothese heißt, dass die Variablen *nicht* unabhängig sind (mit einer Irrtumswahrscheinlichkeit von kleiner als 5 %), und daher *nehmen wir an*, dass sie abhängig sind. Ein Beweis für die Abhängigkeit ist damit nicht erbracht.

Die Anwendung der Yates-Korrekturformel soll die Approximation für kleinere Stichproben verbessern und wird i.a. für Stichprobenumfänge zwischen 20 und 60 Einheiten empfohlen. Manche Autoren empfehlen ihre Anwendung generell. Mit zunehmendem Stichprobenumfang ergeben sich immer kleinere Unterschiede, da der Korrekturterm immer unbedeutender wird.[14]

[13] Sie besitzt eigentlich eine diskrete Verteilung (Multinomialverteilung).

[14] Vgl. Hartung (2009), S. 414; Fleiss/Levin/Paik (2003), S. 57-58 (dort auch ein Überblick über die strittige Diskussion); Everitt (1992), S. 13 f.; Büning/Trenkler (1994), S. 228 empfehlen die Verwendung des exakten Fisher-Tests für Stichprobenumfänge kleiner als 40.

Fisher-Test

Für Tests der Hypothese mit Stichprobenumfängen kleiner als 20 oder bei stark asymmetrischen Randverteilungen (starker Asymmetrie der Zeilen- und Spaltensumme) wird allgemein die Anwendung des exakten Fisher-Tests empfohlen.[15] Die Bezeichnung „exakt“ resultiert aus der Tatsache, dass für die dort verwendete Teststatistik die Verteilung bekannt und für kleine Stichproben berechnet und tabelliert ist.

Wir halten fest: Der χ^2-Test für unser Beispiel hat zum Ergebnis, dass wir eine Abhängigkeit zwischen der bevorzugten Verwendung von Butter bzw. Margarine und der Wohngegend annehmen können.

6.2.3.2 Prüfung der Stärke des Zusammenhangs

Nachdem ein χ^2-Test eine Abhängigkeit der Variablen anzeigt, wird nun versucht, weitere Informationen über die Art des Zusammenhanges, wie Stärke oder Richtung, zu bestimmen. Da χ^2 u.a. eine Funktion des Stichprobenumfanges ist, ist diese Größe als Indikator für die *Stärke* des Zusammenhangs nicht brauchbar. Der Leser kann dies selber überprüfen: So führt eine Verdoppelung aller Stichprobenwerte zur Verdoppelung der χ^2-Werte, obwohl die Stärke des Zusammenhangs davon nicht berührt wird.[16] Noch weniger Anhaltspunkte liefert die χ^2-Testgröße für eine Interpretation der *Richtung* der Abhängigkeit. Bei der Berechnung dieser Größe werden Abweichungen von den erwarteten Größen nach oben und unten durch die Quadrierung gleich bewertet.

Es gibt zwei Gruppen von Indikatoren für die Stärke des Zusammenhanges. Die erste Gruppe basiert trotz der Interpretationsschwierigkeiten auf der χ^2-Teststatistik.

Phi-Koeffizient

Das einfachste Maß ist der *Phi-Koeffizient (ϕ)*:

$$\phi = \sqrt{\frac{\chi^2}{n}} \tag{6.2}$$

Je größer der Wert von ϕ ist, desto stärker ist der Zusammenhang. Als Faustformel wird angegeben, dass ein Wert größer als 0,3 eine Stärke der Abhängigkeit anzeigt, die mehr als trivial ist.[17] Der ϕ-Koeffizient besitzt allerdings eine Reihe von Nachteilen. Insbesondere ist zu beachten, dass die ϕ-Koeffizienten aus verschiedenen Untersuchungen sich nicht vergleichen lassen. Ebenso ist zu beachten, dass bei der Einteilung von stetigen Variablen in zwei Klassen, etwa bei der Transformation einer intervallskalierten Variablen auf eine nominalskalierte Größe, die Wahl der Schnittlegung einen starken Einfluss auf die Testgröße ϕ besitzt.[18]

In unserem Beispiel erhalten wir als Maßgröße:

$$\phi = \sqrt{\frac{27,4}{181}} = 0,389$$

Wir können also nicht nur von der Tatsache eines Zusammenhangs zwischen den Variablen unseres Beispiels ausgehen, sondern gemäß obiger Faustformel auch unterstellen, dass dieser Zusammenhang von Bedeutung ist.

[15] Vgl. Lienert (1973), S. 171 oder Hartung (2009), S. 414-416.
[16] Der Leser mache sich aber bewusst, dass die damit verbundene höhere Signifikanz des Testergebnisses als Folge des höheren Informationsgehaltes durch die „verdoppelte“ Stichprobe sinnvoll ist.
[17] Vgl. Fleiss/Levin/Paik (2003), S. 99.
[18] Vgl. Fleiss/Levin/Paik (2003), S. 99.

Kontingenzkoeffizient

Bei der Untersuchung von Kreuztabellen mit Variablen mit mehr als zwei Ausprägungen kann ϕ Werte über 1 annehmen. In solchen Fällen wird die Verwendung des *Kontingenzkoeffizienten* empfohlen, der eine Modifikation von ϕ darstellt:

$$CC = \sqrt{\frac{\chi^2}{\chi^2 + n}} \tag{6.3}$$

Dieser Koeffizient nimmt nur Werte zwischen 0 und 1 an, kann allerdings nur selten den Maximalwert von 1 erreichen. Die obere Grenze ist eine Funktion der Anzahl der Spalten und Zeilen der Tabelle. Zur Beurteilung sollte daher der jeweilige theoretische Maximalwert mitbetrachtet werden. Die Tatsache unterschiedlicher Obergrenzen lässt auch einen Vergleich zweier Koeffizienten i.a. nicht zu. Die Obergrenze der anzunehmenden Werte von CC wird berechnet nach:

$$CC = \sqrt{(R-1)/R} \quad \text{mit R} = \min(\text{I,J}) \tag{6.4}$$

Für unser Beispiel erhalten wir:

$$CC = 0,362 \quad \text{und} \quad CC_{max} = \sqrt{1/2} = 0,707$$

Cramer's V

Ein anderes Maß, welches ebenfalls Werte zwischen 0 und 1 und auch unabhängig von der Anzahl der Dimensionen den Maximalwert 1 annehmen kann, ist Cramer's V:

$$Cramer's\ V = \sqrt{\frac{\chi^2}{n(R-1)}} \quad \text{(mit R wie oben)} \tag{6.5}$$

Falls eine der untersuchten Variablen binär ist, sind ϕ und Cramer's V identisch.

Die Assoziationsmaße der ersten Gruppe, die sämtlich auf der χ^2-Statistik basieren, nehmen den Wert 0 an, falls keine Assoziation vorliegt und den maximalen Wert bei vollständiger Abhängigkeit. Probleme entstehen bei der Interpretation von Zwischenwerten und bei der Beurteilung, welche Art von Zusammenhang eigentlich vorliegt.

Tau-Koeffizient

Neben den Assoziationsmaßen der ersten Gruppe gibt es Koeffizienten, die Aufschluss über die Stärke einer Assoziation zweier Variabler liefern, indem sie messen, inwieweit die Kenntnis der Ausprägung einer Variablen bei der Prognose der anderen Variablen hilft. Diese Koeffizienten sind die sogenannten tau-(τ-) und lambda-(λ-) Maße von Goodmann und Kruskal.

Lambda-Koeffizient

Die λ-Maße vergleichen die Wahrscheinlichkeit einer falschen Vorhersage der Ausprägung der ersten (abhängigen) Variablen bei Unkenntnis der Ausprägung der zweiten (unabhängigen) Variablen mit der Wahrscheinlichkeit einer falschen Vorhersage der Ausprägung der ersten Variablen bei Kenntnis der Ausprägung der zweiten Variablen.

Je nachdem, welche Variable als erste und welche als zweite betrachtet wird, ergeben sich unterschiedliche Resultate. In unserem Beispiel sei zunächst die Fehlerreduktion betrachtet, die sich bei der Prognose des Wohnorts aus Kenntnis des bevorzugten Brotaufstrichs ergibt.

Ausgehend von den Werten in Abbildung 6.7 würden wir zur Prognose des Wohnortes eines beobachteten Kunden bei Unkenntnis des von der Person bevorzugten Brotaufstrichs am ehesten auf einen städtischen Wohnort tippen, da die meisten der Befragten aus dieser Kategorie stammen und wir damit die geringste Fehlerwahrscheinlichkeit haben. Damit würden wir jedoch, wie aus Abbildung 6.10 ersichtlich,

37,6 % der Personen falsch einschätzen. Sollte die Präferenz einer befragten Person für Margarine uns vor der Prognose ihres Wohnortes bekannt sein, würden wir in Anlehnung an Abbildung 6.7 wiederum auf einen städtischen Wohnort tippen, da unter der Gruppe der die Margarine bevorzugenden Personen die Städter in der Mehrheit sind. Lediglich 23 Personen oder 12,7 % aller Personen (vgl. Abbildung 6.10) würden unter diesen Umständen falsch eingeordnet. Anders wäre es, wenn wir von der Butter-Präferenz einer Person Kenntnis hätten. Aufgrund des höheren Anteils der Landbewohner in der entsprechenden Gruppe würden wir jetzt einen Landbewohner erwarten und in unserer Stichprobe auf diese Art und Weise 16,6 % der Befragten falsch einschätzen.

Insgesamt würden wir bei Kenntnis des jeweils bevorzugten Brotaufstrichs 12,7 % + 16,6 % = 29,3 % der Befragten falsch einschätzen. Im Vergleich zur Fehleinschätzung ohne diese Kenntnis (37,6 %) ergibt sich eine Reduktion um 8,3 Prozentpunkte.

Das $\lambda_{Wohnort}$-Maß bestimmt sich nun aus dem Verhältnis von Fehlerreduktion durch Kenntnis der zweiten Variablen (Brotaufstrich) zur Fehlprognosewahrscheinlichkeit bei Unkenntnis:

$$\lambda_{Wohnort} = \frac{8,3\,\%}{37,6\,\%} = 0,221$$

Analog kann man den Koeffizienten λ_{Sorte} bestimmen, welcher den Nutzen quantifiziert, der durch die Kenntnis des Wohnortes bei der Prognose der bevorzugten Sorte Brotaufstrich entsteht.

Allgemein bestimmen sich die Koeffizienten für die beiden Variablen 1 und 2 nach:

$$\lambda_1 = \frac{\sum_j \max_i n_{ij} - \max_i n_{i.}}{n - \max_i n_{i.}} \tag{6.6}$$

$$\lambda_2 = \frac{\sum_i \max_j n_{ij} - \max_j n_{.j}}{n - \max_j n_{.j}} \tag{6.7}$$

Die λ-Werte bewegen sich immer zwischen 0 und 1. Dabei bedeutet ein Wert nahe Null, dass die Kenntnis der zweiten Variablen für die Prognose der ersten keinen Nutzen stiftet, ein Wert bei Eins, dass die Kenntnis eine fehlerfreie Prognose ermöglicht. Bei der Interpretation ist allerdings zu beachten, dass ein Wert von Null nur bedeutet, dass ein möglicherweise vorhandener Zusammenhang nicht zur Vorhersage geeignet ist. Mit anderen Worten, die Koeffizienten messen nur eine bestimmte Art von Zusammenhang.

Für den Fall, dass die Bestimmung einer abhängigen (ersten) und unabhängigen (zweiten) Variablen aus dem Sachverhalt nicht einwandfrei möglich ist, kann das symmetrische λ verwendet werden. Dieses nimmt Werte zwischen den beiden obigen λ-Werten an und bestimmt sich nach:

$$\lambda_{sym} = \frac{\frac{1}{2}\left(\sum_i \max_j n_{ij} + \sum_j \max_i n_{ij}\right) - \frac{1}{2}\left(\max_j n_{.j} + \max_i n_{i.}\right)}{n - \frac{1}{2}\left(\max_j n_{.j} + \max_i n_{i.}\right)} \tag{6.8}$$

Während bei den λ-Maßen jeweils zur Prognose die Kategorie mit den meisten Beobachtungen gewählt wurde, bestimmen die τ-Maße ihre Prognose unter Berücksichtigung der gesamten Randverteilungen, d. h. unter Berücksichtigung der Häufigkeiten aller Ausprägungen der Variablen.[19]

6.3 Fallbeispiel

6.3.1 Problemstellung

Das in Abschnitt 6.2.1 dargestellte Beispiel soll nachfolgend mit dem Programm SPSS berechnet werden. Zur Durchführung einer Kontingenzanalyse dient die Prozedur CROSSTABS („Kreuztabellen“).

Zunächst sind die Daten in geeigneter Form einzugeben, die von der in Abbildung 6.7 gezeigten Kreuztabelle abweicht. SPSS erwartet standardmäßig, dass die Spal-

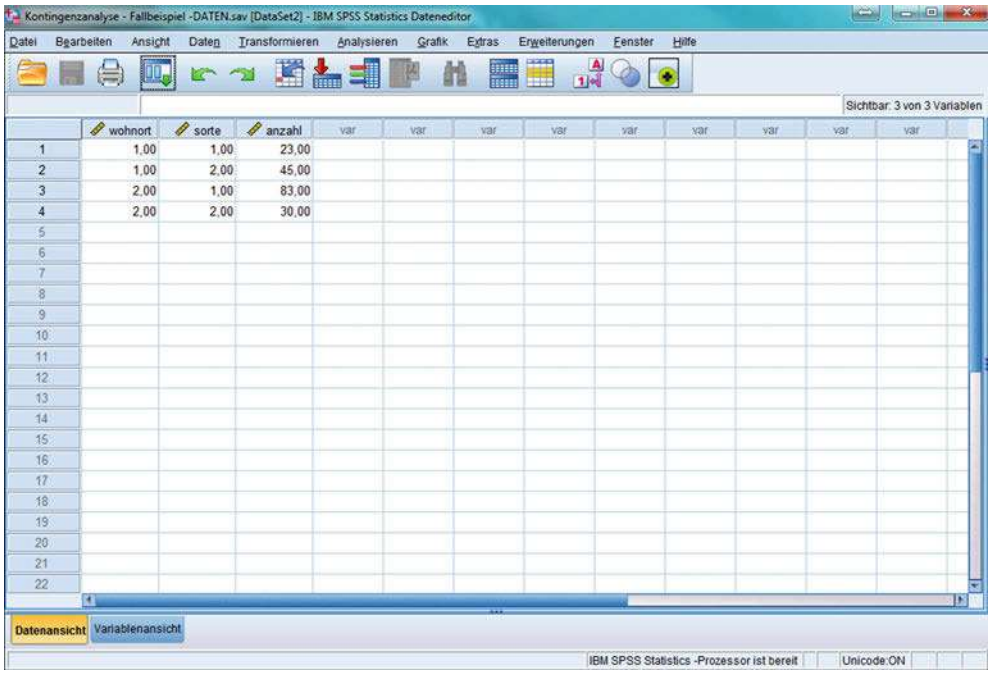


Abbildung 6.14: Daten-Editor (vergrößerte Darstellung)

[19] Zu einer genaueren Darstellung vgl. Hartung (2009), S. 456 ff. Untersuchungen von Variablen auf *ordinalem* Niveau können darüber hinaus mit Hilfe solcher Kennziffern wie Somers Dependenzmaß oder Kendalls tau-Statistik vorgenommen werden. Vgl. hierzu Everitt (1992), Kap. 3.7.3, S. 130.
Eine weitere Methode zur Untersuchung eines signifikanten χ^2-Wertes ist die Residual-Analyse. Hierbei werden die Abweichungen der beobachteten Häufigkeiten von den erwarteten Werten berechnet, um die Merkmalskombinationen zu bestimmen, die „aus dem Rahmen fallen“. Vgl. Everitt (1992), S. 46 f.; Lienert (1973), S. 538 oder Haberman (1978), S. 17-21.
In diesem Kapitel werden nur Kontingenzmaße für die Untersuchung eines signifikanten χ^2-Ergebnisses betrachtet. Für die Untersuchung eines signifikanten Ergebnisses des Test auf Homogenität gegen Heterogenität zweier Verteilungen gibt es ebenfalls spezielle Maße wie die Anteilsdifferenz, das relative Risiko oder den Kreuzproduktquotienten, die für die 4-Felder-Tafel auch im SPSS-Programm abrufbar sind. Vgl. Lienert (1973), S. 457-463.

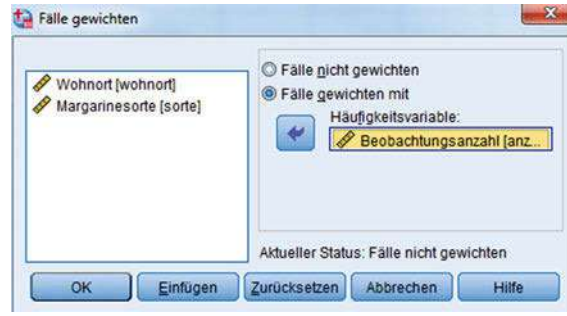


Abbildung 6.15: Dialogfeld „Fälle gewichten“

ten der Datenmatrix sich auf Variable und die Zeilen auf die Beobachtungen dieser Variablen beziehen. Neben den beiden kategorialen Variablen „Wohnort“ und „Brotaufstrichsorte“ ist daher eine dritte Variable zu definieren, die die Beobachtungszahlen in den Zellen der Kreuztabelle enthält (siehe Abbildung 6.14). Die Spalten (Variablen) „wohnort“ und „sorte“ definieren die vier Zellen der Kreuztabelle und die Spalte (Variable) „anzahl“ enthält die zugehörigen Beobachtungszahlen. Die Zuordnung der Beobachtungszahlen zu den Zellen erfolgt mit Hilfe des Menüpunktes *Daten* und der Option *Fälle gewichten*. Man gelangt damit zum Dialogfenster in Abbildung 6.15, wo die gewünschten Spezifikationen vorgenommen werden können. Über *OK* wird dieses Dialogfenster wieder verlassen.

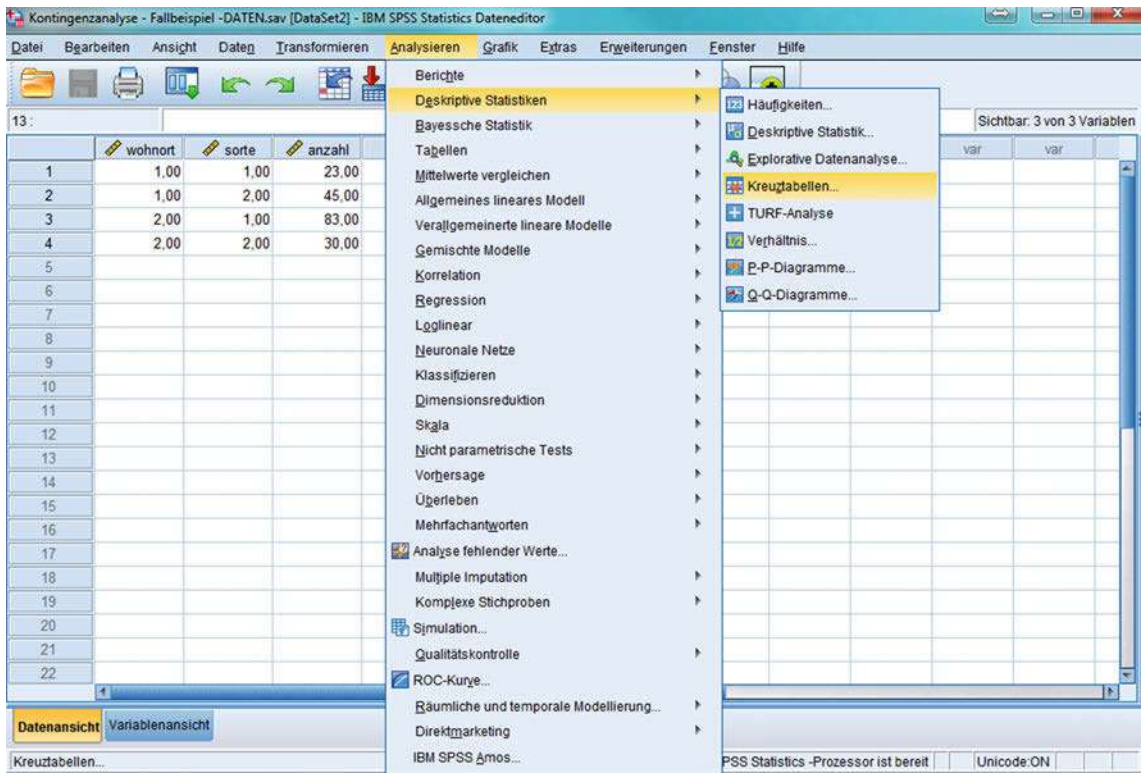


Abbildung 6.16: Daten-Editor mit Auswahl „Kreuztabellen“

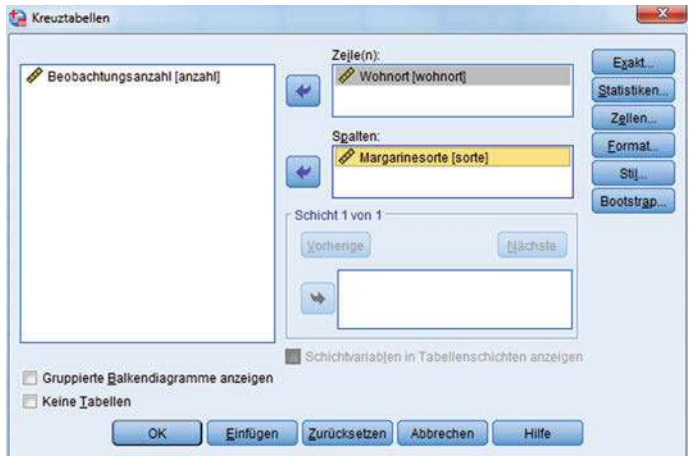


Abbildung 6.17: Dialogfeld „Kreuztabellen“

Im Anschluss daran erfolgt die Kontingenzanalyse mit dem Menüpunkt „Analysieren“ und den Unterpunkten „Deskriptive Statistiken“ und „Kreuztabellen“ (vgl. Abbildung 6.16).

Im geöffneten Dialogfeld „Kreuztabellen“ werden die Variablen für die Zeilen (hier: Wohnort) und für die Spalten (hier: Brotaufstrichsorte) festgelegt und in die entsprechenden Felder übertragen (vgl. Abbildung 6.17).

Durch Anklicken des Button „Statistik“ öffnet sich ein Dialogfeld, in dem verschiedene Teststatistiken ausgewählt werden können, wobei sich im Folgenden auf die oben erklärten Tests beschränkt werden soll (vgl. Abbildung 6.18). Zunächst wird der Chi-Quadrat-Test zur Überprüfung der Unabhängigkeit der Merkmale markiert. Da es sich bei den beiden Variablen „Wohnort“ und „Brotaufstrichsorte“ um nominal skalierte Merkmale handelt, werden desweiteren die entsprechenden Statistiken, „Kontingenzkoeffizient“, „Phi und Cramer's V“ sowie „Lambda“ mit einem Häkchen versehen. Durch Anklicken von „Weiter“ gelangt man zurück zum Dialogfeld „Kreuztabellen“.

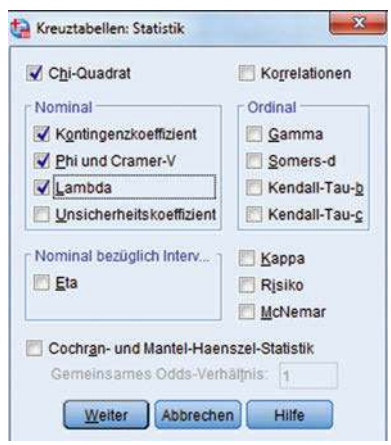


Abbildung 6.18: Dialogfeld „Statistiken“

Abbildung 6.19: Dialogfeld „Zellen anzeigen“

Der Button „Zellen“ führt in ein entsprechendes Dialogfeld, indem die darzustellenden Parameter für die Vierfelder-Tafel einzustellen sind. Die damit zu generierende Tabelle dient zur Veranschaulichung der Ergebnisse der Kontingenzanalyse (vgl. Abbildung 6.19).

Dort wird festgelegt welche Häufigkeiten (Beobachtete, Erwartete) angezeigt werden sollen, welche Prozentwerte (Zeilenweise, Spaltenweise, Gesamt) sowie in welcher Form die Residuen zu berechnen sind. Letztere stellen die Differenz zwischen „beobachteten“ und „erwarteten“ Häufigkeiten dar. Durch Anklicken des „Weiter“ Buttons gelangt man wiederum zurück zum Dialogfeld „Kreuztabellen“ und startet durch „OK“ die Prozedur. Hierdurch ergeben sich die Berechnungen, welche bereits aus dem Abschnitt 6.2 bekannt sind (vgl. Abbildung 6.22). Neu sind seit dem Erscheinen von IBM SPSS 22 die beiden Button „Stil“ und „Bootstrap“. Mit „Stil“ wird der Tabellenstil festgelegt (vgl. Abb. 6.20).

Bootstrapping

Abbildung 6.20: Dialogfeld „Tabellenstil“

Der Button „Bootstrap“ ermöglicht ein Resampling, was insbesondere bei kleinen Stichproben zur Anwendung kommt. Beim Bootstrapping (Vgl. Abb. 6.21) werden aus einer Stichprobe viele (z. B. 1000 als Voreinstellung) Stichproben gezogen, was durch zurücklegen ermöglicht wird. Bootstrapping ist ein Verfahren, bei dem sich der Anwender „an den eigenen Haaren aus dem Sumpf zieht“ und auf diese Weise Informationen über eine ansonsten „ungekannte“ Grundgesamtheit erhält.[20]

Abbildung 6.21: Dialogfeld „Bootstrap“

6.3.2 Ergebnisse

Zunächst wird im Ausdruck die Vierfelder-Tafel dargestellt. Neben der Anzahl an Beobachtungen jeder Kombination von Merkmalsausprägungen (ANZAHL) in jeder Zelle, werden die Zeilen- (WOHNORT), die Spalten- (BROTAUFSTRICHSORTE) und Totalprozente (GESAMTZAHL) ausgedruckt. Diese Angaben sind identisch zu den Informationen in den Abbildungen 6.7 bis 6.10. Ebenfalls aufgelistet wird in der Darstellung die erwartete Anzahl jeder Merkmalskombination e_{ij} (ERWARTETE ANZAHL) sowie die Differenz zwischen beobachtetem und erwartetem Wert (RESIDUEN).

Im unteren Teil der Abbildung sind die Statistiken χ^2 (CHI-QUADRAT NACH PEARSON) und die Yates-Korrektur χ^2_{korr} (KONTINUITÄTSKORREKTUR) bestimmt. Die uns bereits bekannten Werte führen unter Berücksichtigung der Freiheitsgrade (DF) aufgrund der Vergleichsgröße (SIGNIFIKANZ) bei einem Testniveau von 5 % in beiden Teststatistiken wieder zu einer Ablehnung der Nullhypothese.

Zusätzlich automatisch ausgedruckt werden die Mantel-Haenszel-Statistik (ZUSAMMENHANG LINEAR MIT LINEAR) und die Likelihood-Statistik (LIKELIHOOD- QUOTIENT). Der Mantel-Haenzel-Test ist allerdings für Fragestellungen mit nominalskalierten Variablen nicht anwendbar und wird von uns daher

[20] Vgl. Für die Grundlagen des Bootstrapping siehe Efron (2003); Efron/Tibshirani (1994).

Wohnort * Margarinesorte Kreuztabelle

			Margarinesorte		
			MARGARINE	BUTTER	Gesamt
Wohnort	LÄNDLICH	Anzahl	23	45	68
		Erwartete Anzahl	39,8	28,2	68,0
		% innerhalb von Wohnort	33,8%	66,2%	100,0%
		% innerhalb von Margarinesorte	21,7%	60,0%	37,6%
		% der Gesamtzahl	12,7%	24,9%	37,6%
		Residuen	-16,8	16,8	
	STÄDTISCH	Anzahl	83	30	113
		Erwartete Anzahl	66,2	46,8	113,0
		% innerhalb von Wohnort	73,5%	26,5%	100,0%
		% innerhalb von Margarinesorte	78,3%	40,0%	62,4%
		% der Gesamtzahl	45,9%	16,6%	62,4%
		Residuen	16,8	-16,8	
Gesamt		Anzahl	106	75	181
		Erwartete Anzahl	106,0	75,0	181,0
		% innerhalb von Wohnort	58,6%	41,4%	100,0%
		% innerhalb von Margarinesorte	100,0%	100,0%	100,0%
		% der Gesamtzahl	58,6%	41,4%	100,0%

Abbildung 6.22: Ergebnisse der Kontingenzanalyse (1. Teil)

Likelihood-Statistik

nicht weiter beachtet.[21] Der auf der Likelihood-Statistik beruhende Test basiert auf dem Testprinzip der Maximum-Likelihood-Schätzung und führt bei großen Stichproben zu ähnlichen Ergebnissen wie der χ^2-Test.[22] Der Auszug in Abbildung 6.23 endet mit der Angabe der kleinsten erwarteten Anzahl pro Zelle (MINIMALE ERWARTETE HÄUFIGKEIT). Ist diese kleiner als fünf, so bestimmt SPSS anstelle der χ^2-Statistik den exakten Fisher-Test.

In den Abbildung 6.24 und Abbildung 6.25 sind die verschiedenen Assoziationsmaße aufgelistet:

Die auf der χ^2-Statistik beruhenden Maße sind zunächst angegeben: Der ϕ-Koeffizient (PHI), Cramer's V (CRAMER'S V) und der Kontingenzkoeffizient CC (KONTINGENZKOEFFIZIENT). Da die betrachteten Variablen binär sind, sollten ϕ und Cramer's V identisch sein. SPSS berechnet jedoch an dieser Stelle für 2x2-Tafeln den Korrelationskoeffizienten mit Vorzeichen. In der jeweiligen Zeile ist unter SIGNIFIKANZ auch die Größe angegeben, mit der eine Testentscheidung bzgl. der Nullhypothese, dass das betrachtete Maß gleich Null sei, möglich ist. In unserem Fall

[21] Vgl. zu einer ausführlichen Darstellung Bishop/Fienberg/Holland (2007) oder Fleiss/Levin/Paik (2003), S. 250 ff.

[22] Eine Darstellung des zugrundeliegenden Modells und der Teststatistik sowie einem genauen Vergleich mit der Chi-Quadrat-Statistik findet der Leser bei Hartung (2009), S. 435-439.

Chi-Quadrat-Tests

	Wert	df	Asymptotische Signifikanz (zweiseitig)	Exakte Signifikanz (2-seitig)	Exakte Signifikanz (1-seitig)
Chi-Quadrat nach Pearson	27,473[a]	1	,000		
Kontinuitätskorrektur[b]	25,864	1	,000		
Likelihood-Quotient	27,773	1	,000		
Exakter Test nach Fisher				,000	,000
Zusammenhang linear-mit-linear	27,321	1	,000		
Anzahl der gültigen Fälle	181				

a. 0 Zellen (,0%) haben eine erwartete Häufigkeit kleiner 5. Die minimale erwartete Häufigkeit ist 28,18.

b. Wird nur für eine 2x2-Tabelle berechnet

Abbildung 6.23: Ergebnisse der Kontingenzanalyse (2. Teil)

Symmetrische Maße

		Wert	Näherungsweise Signifikanz
Nominal- bzgl. Nominalmaß	Phi	-,390	,000
	Cramer-V	,390	,000
	Kontingenzkoeffizient	,363	,000
Anzahl der gültigen Fälle		181	

Abbildung 6.24: Ergebnisse der Kontingenzanalyse (3. Teil)

können wir uns bei einem Testniveau von 5 % in allen drei Fällen gegen die Nullhypothese entscheiden.

In der Abbildung 6.25 finden sich die Assoziationsmaße, welche die Stärke des Zusammenhangs über die Reduktion von Prognosefehlern messen. Zunächst sind dies die λ-Maße (LAMBDA), beginnend mit dem symmetrischen λ (SYMMETRISCH), danach das für den Fall der Variablen WOHNORT als zu prognostizierender (WOHNORT ABHÄNGIG) und abschließend das für den Fall der Variablen BROTAUFSTRICHSORTE als zu prognostizierender Variable (BROTAUFSTRICHSORTE ABHÄNGIG). Aus der Angabe des Standardfehlers der Statistik (ASYMPTOTISCHER STANDARDFEHLER) kann man ein Konfidenzintervall für die Statistik bilden.[23]

In den nächsten Zeilen befinden sich die Angaben zu den τ-Maßen (GOODMAN UND KRUSKAL TAU) mit jeweils einer der beiden Variablen als Prognosevariable. Hier wird zusätzlich zum Standardfehler der Statistik ein Test berechnet für die Nullhypothese, dass das betrachtete τ gleich Null ist.[24]

[23] Zur Bildung vgl. Hartung (2009), S. 457-458.
[24] Zur Verteilung der Teststatistik vgl. Hartung (2009), S. 461.

Abbildung 6.25: Ergebnisse der Kontingenzanalyse (4. Teil)

Richtungsmaße

			Wert	Asymptotischer Standardfehler[a]	Näherungsweises t[b]	Näherungsweise Signifikanz
Nominal- bzgl. Nominalmaß	Lambda	Symmetrisch	,259	,095	2,464	,014
		Wohnort abhängig	,221	,112	1,747	,081
		Margarinesorte abhängig	,293	,092	2,722	,006
	Goodman-und-Kruskal-Tau	Wohnort abhängig	,152	,054		,000[c]
		Margarinesorte abhängig	,152	,054		,000[c]

a. Die Null-Hyphothese wird nicht angenommen.

b. Unter Annahme der Null-Hyphothese wird der asymptotische Standardfehler verwendet.

c. Basierend auf Chi-Quadrat-Näherung

6.3.3 SPSS-Kommandos

Abbildung 6.26 gibt die Kommandodatei zum Fallbeispiel wieder.

```
* MVA: Fallbeispiel Kontingenzanalyse.
* DATENDEFINITION.
DATA LIST FREE / wohnort sorte anzahl.
VARIABLE LABELS wohnort "Wohnort"
  /sorte "Brotaufstrichsorte"
  /anzahl "Beobachtungsanzahl".
VALUE LABELS
  /wohnort 1 "ländlich" 2 "städtisch"
  /sorte 1 "Margarine" 2 "Butter".

BEGIN DATA
1 1 23
1 2 45
2 1 83
2 2 30
END DATA.

* PROZEDUR.
WEIGHT BY anzahl.
* Kontingenzanalyse für den Margarinemarkt.
CROSSTABS
  /TABLES = wohnort BY sorte
  /FORMAT = AVALUE TABLES
  /STATISTIC = CHISQ CC PHI LAMBDA
  /CELLS = COUNT EXPECTED ROW COLUMN TOTAL RESID.
```

Abbildung 6.26: SPSS-Job zur Kontingenzanalyse

6.4 Anwendungsempfehlungen

Aus der Diskussion im Abschnitt 6.2.1 ergibt sich, dass schon im Planungsprozess mit dem gebotenen Sachverstand zu klären ist, welche Art von Untersuchung angemessen ist und welche Variablen zu erheben sind.

Jeglicher auf der Grundlage der Kontingenzanalyse ermittelte Zusammenhang kann nur ein statistischer Zusammenhang sein. Hieraus z. B. eine Kausalität zu begründen, kann zu erheblichen Irrtümern und Fehlschlüssen führen.

Im Folgenden werden die wichtigsten Voraussetzungen des χ^2-Tests zusammengestellt:

1. Die einzelnen Beobachtungen müssen voneinander unabhängig sein.[25]

2. Jede Beobachtung muss eindeutig einer Kombination von Merkmalsausprägungen zugeordnet werden können.

3. Der Anteil der Zellen mit erwarteten Häufigkeiten, die kleiner als fünf sind, darf 20 % nicht überschreiten (Faustformel). Keine dieser Häufigkeiten darf kleiner als eins sein.[26] Ein Zusammenfassen mehrerer Merkmalsklassen zu einer, um hierdurch größere zu erwartende Werte zu erreichen, ist nur unter ganz bestimmten Bedingungen zulässig und sollte sorgfältig überlegt sein.[27]

[25] Dies ist z. B. dann nicht gegeben, wenn die Merkmale zu unterschiedlichen Zeitpunkten an denselben Personen erhoben wurden. Bei diesen sog. verbundenen Stichproben muss auf den McNemar-Test oder Cochran-Test zurückgegriffen werden. Vgl. Bortz (2010).

[26] Zu alternativen Auswertungsmöglichkeiten in den Fällen, in denen diese Voraussetzung nicht gegeben ist, vergleiche Lienert (1973), S. 398 ff. Everitt (1992), S. 39 ff. zitiert Arbeiten, nach denen obige Voraussetzungen zu restriktiv seien.

[27] Zu weiteren Information vgl. Everitt (1992), S. 39 ff. und Lienert (1973), S. 398.

4. Im 4-Felder-Fall bei Stichproben mit einem Umfang von weniger als 60 Einheiten sollte der χ^2-Test nicht angewandt werden. Bei Stichprobenumfängen zwischen 20 und 60 bietet sich die Yates-Korrektur an, bei noch kleineren Umfängen sollte im 4-Felder-Fall auf den exakten Fisher-Test ausgewichen werden.[28]

Literaturhinweise

A. Basisliteratur zur Kreuztabellierung und Kontingenzanalyse

Bishop, Y./Fienberg, S./Holland, P. (2007), Discrete Multivariate Analysis. Theory and Practice, New York.

Bortz, J. (2010), Statistik für Human- und Sozialwissenschaftler, 7. Auflage, Berlin u. a.

Fienberg, S. (2007), The Analysis of Cross-Classified Categorical Data, 2. Auflage, New York.

SPSS Inc. (2017), SPSS Statistics Base 25, Chicago.

Wickens, T. (1989), Multiway Contingency Tables Analysis for the Social Sciences, Hillsdale.

B. Zitierte Literatur

Agresti, A. (2007), An Introduction to Categorical Data Analysis, 2. Auflage, Hoboken (N.J.).

Bishop, Y./Fienberg, S./Holland, P. (2007), Discrete Multivariate Analysis. Theory and Practice, New York.

Böhler, H./Fürst, A. (2014), Marktforschung, 4. Auflage, Stuttgart u. a.

Bortz, J. (2010), Statistik für Human- und Sozialwissenschaftler, 7. Auflage, Berlin u. a.

Büning, H./Trenkler, G. (1994), Nichtparametrische statistische Methoden, 2. Auflage, Berlin u. a.

Efron, B. (2003), Second Thoughts on the Bootstrap, in: *Statistical Science*, Vol. 18, Nr. 2, S. 135–140.

Efron, B./Tibshirani, R. (1994), An Introduction to the Bootstrap, Boca Raton (FL).

Everitt, B. (1992), The Analysis of Contingency Tables, 2. Auflage, New York.

[28] Zur Diskussion um die Empfehlung der ständigen Anwendung der Yates-Korrektur vgl. Fleiss/Levin/Paik (2003), S. 27 und Büning/Trenkler (1994), S. 228.

Fahrmeier, L./Hamerle, A./Tutz, G. (Hrsg.) (1996), Multivariate statistische Verfahren, 2. Auflage, Berlin u. a.

Fleiss, J./Levin, B./Paik, M. (2003), Statistical Methods for Rates and Proportions, 3. Auflage, New Jersey.

Haberman, S. (1978), Analysis of Qualitative Data: Vol. 1: Introductory Topics, New York u. a.

Hartung, J. (2009), Statistik: Lehr- und Handbuch der angewandten Statistik, 15. Auflage, München u. a.

Iacobucci, D. /Churchill Jr., G. (2015), Marketing Research: Methodological Foundations, 11. Auflage, Mason (OH).

Kendall, M./Stuart, A. (1979), The advanced theory of statistics, Vol. 2, 4. Auflage, London u. a.

Lienert, G. (1973), Verteilungsfreie Methoden in der Biostatistik, Band I, Meisenheim.

Zeisel, H. (1970), Die Sprache der Zahlen, Köln u. a.

7 Faktorenanalyse

K. Backhaus, B. Erichson, W. Plinke, R. Weiber, *Multivariate Analysemethoden*,
https://doi.org/10.1007/978-3-662-56655-8_8

7.1 Problemstellung

Die explorative Faktorenanalyse ist ein Verfahren der multivariaten Analyse, das darauf gerichtet ist, Strukturen in großen Variablensets erkennen zu können. Große Variablensets sind oftmals dadurch gekennzeichnet, dass mit steigender Zahl der Variablen davon auszugehen ist, dass sich mehr und mehr Variablen überlappen. Statistisch drückt sich dies in Korrelationen zwischen den Variablen aus. Die exploratorische Faktorenanalyse (EFA) versucht, die Beziehungszusammenhänge in einem großen Variablenset insofern zu strukturieren, als sie Gruppen von Variablen identifiziert, die hoch miteinander korreliert sind und diese von weniger korrelierten Gruppen trennt. Die Gruppen von jeweils hoch korrelierten Variablen bezeichnet man auch als Faktoren. Neben der Strukturierungsfunktion wird die Faktorenanalyse auch zur Datenreduktion eingesetzt. Wir sprechen von einer Datenreduktion, wenn zusätzlich zur Strukturierung Ausprägungen für die strukturierten Faktoren (Faktorwerte) ermittelt werden. Liegen solche Faktorwerte vor, dann lassen sie sich anstelle der Originalwerte der Variablen verwenden. In Abbildung 7.1 sind einige Anwendungsbeispiele der Faktorenanalyse zusammengestellt. Sie vermitteln einen Einblick in die Problemstellung, die Zahl und Art der Merkmale, die aus den Merkmalen extrahierten Faktoren sowie die jeweiligen Untersuchungseinheiten.

Datenstrukturierung

Datenreduktion

Problemstellung	Merkmale	Faktoren
Stadtanalyse[1]	Bevölkerungszahl, Beschäftigtenzahl, Dienstleistungsangebot, Schulbildung, Häuserwert.	Bevölkerungs- und Beschäftigtenfaktor, Ausbildungs- und Wirtschaftsfaktor.
Untersuchungen der kognitiven Fähigkeiten[2]	Streckenplanung, Gruppierung von Symbolen, Erkennung von Ähnlichkeiten, etc. Wortschatz, Schlussfolgerungseigenschaften, Satzbau, etc.	Bildliche Fähigkeit, Verbale Fähigkeit.
Kostenanalyse[3]	18 Kostenarten differenziert nach jeweils 5 Kosteneigenschaften.	Beeinflussbarkeit, Deckungsdringlichkeit.
Blutdruckmessung[4] (SBDM = Systolische Blutdruckmessung; DBDM = Diastolische Blutdruckmessung)	1a. SBDM, 1b. bis 12 SBDM, 2a. DBDM, 2b. bis 12 DBDM.	Systolischer Blutdruck, Diastolischer Blutdruck.

Abbildung 7.1: Anwendungsbeispiele der Faktorenanalyse

[1] Vgl. Harmann (1976), S. 13 ff.
[2] Vgl. Carroll (2004).
[3] Vgl. Plinke (1985), S. 118 ff.
[4] Vgl. Überla (1977), S. 264 ff.

Veranschaulichen wir uns die Problemstellung noch einmal anhand eines konkreten Beispiels. In einer Befragung seien Probanden nach ihrer Einschätzung von Emulsionsfetten (Butter, Margarine) befragt worden. Dabei seien die Marken Rama, Sanella, Becel, Du darfst, Holländische Markenbutter und Weihnachtsbutter anhand der Variablen Anteil ungesättigter Fettsäuren, Kaloriengehalt, Vitamingehalt, Haltbarkeit und Preis auf einer siebenstufigen Skala von hoch bis niedrig beurteilt worden. Beispiel

Die nachfolgende Abbildung 7.2 zeigt einen Ausschnitt aus dem entsprechenden Fragebogen.

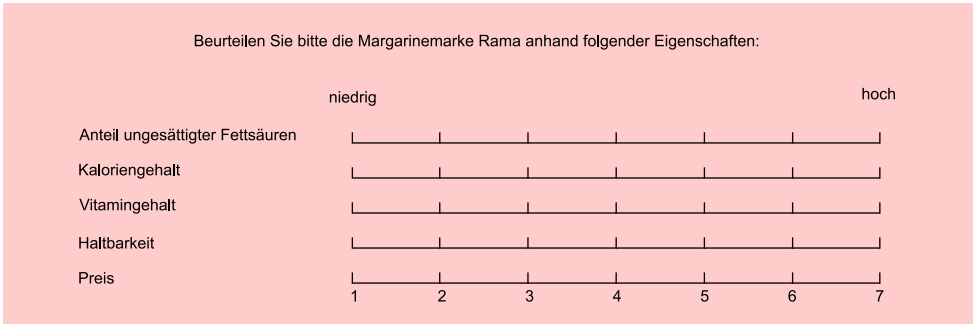


Abbildung 7.2: Fragebogenausschnitt

Die Beantwortung des obigen Fragebogenausschnitts durch die 30 befragten Probanden liefert subjektive Eigenschaftsurteile der fünf Variablen für die Margarinemarke Rama, sodass eine $(30x5)$-Matrix entsteht. Diese Matrix kann der weiteren Analyse zugrunde gelegt werden. Wir haben dann 5 Eigenschaften und 30 Fälle, wobei wir für unsere Analyse *unterstellen*, dass die Befragtenurteile *unabhängig* voneinander sind.

Will man jedoch die sechs Marken gleichzeitig analysieren, so werden häufig für jede Eigenschaft pro Marke Durchschnittswerte über alle 30 Befragten gebildet. Wir erhalten dann eine $(6x5)$-Matrix, wobei die Marken als Fälle interpretiert werden. Bei einer solchen Durchschnittsbildung muss man sich allerdings bewusst sein, dass man bestimmte Informationen (nämlich die über die Streuung der Ausprägungen zwischen den Personen) verliert. Wird die Struktur der Variablen untersucht, spricht man auch von R-Faktorenanalyse. Durchschnittsbildung R-Faktorenanalyse

Je größer die Streuung der Stichprobenwerte über die Zahl der Probanden ist, um so problematischer ist der Aussagewert bei einer solchen Vorgehensweise. Man könnte in einem solchen Fall eine Faktorenanalyse auch über die befragten Probanden rechnen (Q-Faktorenanalyse). Das Ergebnis wäre einer Clusteranalyse vergleichbar. Der Unterschied zwischen beiden Verfahren bestünde im wesentlichen darin, dass bei der exploratorischen Faktorenanalyse zur Ähnlichkeitenbestimmung Korrelationen und bei der Clusteranalyse vorwiegend Distanzen herangezogen werden. Da in praktischen Fällen die R-Faktorenanalyse klar dominiert, beziehen sich auch die nachfolgenden Ausführungen auf die der $(6x5)$-Matrix zugrundeliegenden Durchschnittswerte über alle Personen und damit auf eine R-Faktorenanalyse. Im abschließenden Kapitel wird ein Lösungsvorschlag für eine Alternative zur Durchschnittsbildung vorgestellt. Q-Faktorenanalyse

Es sei unterstellt, dass die ausgewählten Eigenschaften für die Beurteilung von Emulsionsfetten auch als relevant angesehen werden können. Für die folgenden Betrachtungen verdichten wir nun die Werte aus Abbildung 7.3 durch Bildung der arithmetischen Mittel für jede Objekt/Variablen-Kombination über alle 30 Befragten. Als

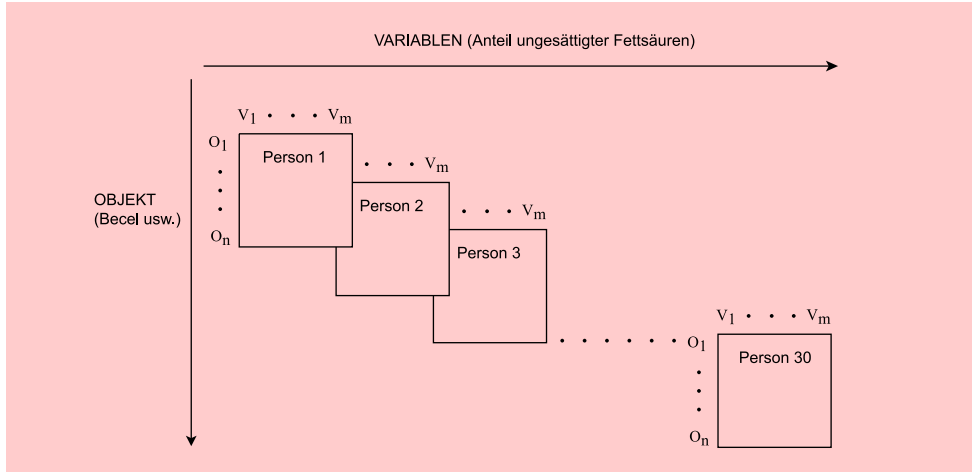


Abbildung 7.3: Ausgangsdaten im Beispiel

Durchschnittswert der 30 befragten Probanden mögen sich bei dieser Befragung über alle Probanden folgende Werte ergeben haben (Abbildung 7.4).

	Eigenschaften				
Marken	X_1	X_2	X_3	X_4	X_5
Rama	1	1	2	1	2
Sanella	2	6	3	3	4
Becel	4	5	4	4	5
Du darfst	5	6	6	2	3
Holländische Butter	2	3	3	5	7
Weihnachtsbutter	3	4	4	6	7

Abbildung 7.4: Mittelwertmatrix für das 6-Produkte-Beispiel

wobei:

X_1 = Anteil ungesättigter Fettsäuren
X_2 = Kaloriengehalt
X_3 = Vitamingehalt
X_4 = Haltbarkeit
X_5 = Preis

Gruppierung

Ein erster Blick auf die Ausgangsdatenmatrix macht bereits deutlich, dass die Eigenschaften (Variablen) X_1 bis X_3 bei den Margarinemarken (Sanella, Becel und Du darfst; Ausnahme: Rama) tendenziell höher bewertet wurden als bei den Buttersorten (Holländische Butter und Weihnachtsbutter), während die Eigenschaften X_4 und X_5 primär bei den Buttersorten höher ausgeprägt sind. Die Ausgangsdaten geben damit in diesem Beispiel bereits einen Hinweis darauf, dass zwei Gruppen (X_1, X_2, X_3 und X_4, X_5) ähnlich beurteilter Variablen existieren, die sich in der Beurteilung

untereinander aber unterscheiden. Damit lässt sich in diesem Beispiel auf Grund der *Datenstruktur* ein Beziehungszusammenhang vermuten. Will man diese Vermutung genauer überprüfen, so ist es erforderlich, auf eine statistische Maßgröße zurückzugreifen, die die Quantifizierung von Beziehungen zwischen Variablen erlaubt. Ein solches statistisches Maß stellt der *Korrelationskoeffizient* dar. Durch die Berechnung von *Korrelationen* zwischen allen Variablen lässt sich die Stärke der Beziehungszusammenhänge zwischen allen Variablen berechnen.

Korrelation

Ausgehend von den fünf Eigenschaften, die in der Befragung verwendet wurden, wird aufgrund der sich in den Daten manifestierenden Beziehungen zwischen X_1 bis X_3 bzw. X_4 und X_5 vermutet, dass eigentlich nur zwei unabhängige Beschreibungsdimensionen für die Aufstrichfette existieren (die die Varianten in den Variablen bedingen). X_1 bis X_3 könnten z. B. Ausdruck *eines* Faktors sein, den man etwa mit „Gesundheit“ bezeichnen könnte, denn sowohl der Anteil ungesättigter Fettsäuren als auch Kaloriengehalt und Vitamingehalt haben „etwas mit der Gesundheit zu tun“. Ebenso können die Variablen X_4 und X_5 (Haltbarkeit und Preis) Ausdruck für Wirtschaftlichkeitsüberlegungen sein. Man könnte also vermuten, dass sich die Variablen X_1 bis X_5 in diesem konkreten Fall auf zwei komplexere Variablenbündel verdichten lassen. Diese „Variablenbündel“ bezeichnen wir im Folgenden als *Faktoren*. Der Zusammenhang zwischen Variablen und Faktoren ist in Abbildung 7.5 graphisch veranschaulicht.

1. Faktor

2. Faktor

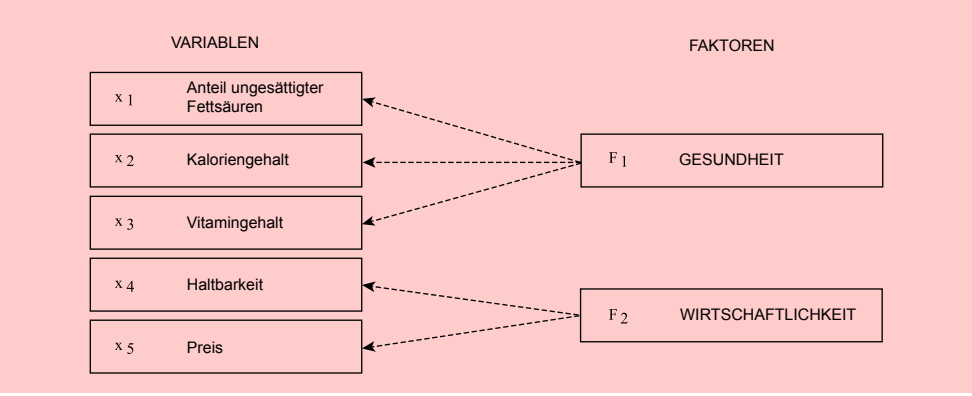


Abbildung 7.5: Grundgedanke der Faktorenanalyse im Beispiel

Werden die betrachteten Eigenschaften zu Faktoren zusammengefasst, so ist unmittelbar einsichtig, dass gegenüber der Mittelwertmatrix in Abbildung 7.4 ein weiterer Informationsverlust entsteht, da i. d. R. weniger Faktoren als ursprüngliche Eigenschaften betrachtet werden. Dieser Informationsverlust ist darin zu sehen, dass zum einen die Faktoren in der Summe i. d. R. nur weniger Varianz erklären können als die fünf Ausgangsvariablen besitzen und zum anderen die Varianz einer jeden Ausgangsgröße in der Erhebungsgesamtheit ebenfalls durch die Faktoren i. d. R. nicht vollständig erklärt werden kann. Der Verlust an erklärter Varianz wird im Rahmen der Faktorenanalyse zugunsten der Variablenverdichtung bewusst in Kauf genommen. Allerdings muss sich der Anwender vorab überlegen, in welchem Ausmaß dieser Erklärungsverlust (im Sinne eines Varianzerklärungsverlustes) bei den einzelnen Ausgangsvariablen toleriert bzw. wieviel Varianz durch die Faktoren bei einer bestimmten Variablen erklärt werden soll. Den Umfang an Varianzerklärung, den die Faktoren ge-

Informationsverlust

meinsam für eine Ausgangsvariable liefern, wird als *Kommunalität* bezeichnet. Die Art und Weise, mit der die Kommunalitäten bestimmt werden, ist unmittelbar an die Methode der Faktorenermittlung gekoppelt. Je nachdem, welche Überlegungen der Kommunalitätenbestimmung zugrunde liegen, werden unterschiedliche Faktorenanalyseverfahren relevant.

Kommunalität

Ist eine Entscheidung über die Höhe der Kommunalitäten der einzelnen Ausgangsvariablen getroffen, so muss weiterhin über die *Anzahl der zu extrahierenden Faktoren* entschieden werden, da das Ziel der Faktorenanalyse gerade darin zu sehen ist, weniger Faktoren als ursprüngliche Variable zu erhalten. Hier steht der Anwender vor dem Zielkonflikt, dass mit einer geringen Faktorenzahl tendenziell ein großer Informationsverlust (im Sinne von nicht erklärter Varianz) verbunden ist und umgekehrt. In unserem Beispiel hatten wir uns aufgrund einer Plausibilitätsbetrachtung für zwei Faktoren entschieden.

Faktorextraktion

Ist schließlich die Anzahl der Faktoren bestimmt, so ist es von besonderem Interesse, die Beziehungen zwischen den Ausgangsvariablen und den Faktoren zu kennen. Zu diesem Zweck werden „Korrelationen“ berechnet, die ein Maß für die Stärke und die Richtung der Zusammenhänge zwischen Faktoren und ursprünglichen Variablen angeben. Diese Korrelationen werden als Faktorladungen bezeichnet und in der sog. *Faktorladungsmatrix* zusammengefasst. Damit endet die Strukturierungsanalyse.

Faktorladungen

Wird zusätzlich eine Datenreduktion angestrebt, ist es von Interesse, wie die befragten Personen die Marken Rama, Sanella, Becel, Du darfst, Holländische Markenbutter und Weihnachtsbutter im Hinblick auf die beiden „latenten“ Faktoren „Gesundheit“ und „Wirtschaftlichkeit“ beurteilen würden. Gesucht ist also die entsprechende Matrix zu Abbildung 7.4, die die Einschätzung der Marken bezüglich der beiden Faktoren „Gesundheit“ und „Wirtschaftlichkeit“ enthält. Diese „Einschätzungen“ werden als *Faktorwerte* (im Programm SPSS als *factor scores*) bezeichnet. Abbildung 7.6 zeigt die entsprechende Faktorwerte-Matrix für unser kleines Ausgangsbeispiel. Die Darstellung enthält standardisierte Werte, wobei die Ausprägungen als Abweichungen vom Mittelwert dargestellt sind.

Faktorwerte

	Faktor 1	Faktor 2
Rama	$-1,21136$	$-1,25027$
Sanella	$-0,48288$	$-0,26891$
Becel	$0,57050$	$0,19027$
Du darfst	$1,56374$	$-0,88742$
Holl. Butter	$-0,63529$	$0,94719$
Weihnachtsbutter	$0,19530$	$1,26914$

Abbildung 7.6: Faktorwerte-Matrix

Die Faktorwerte liefern nicht nur einen Anhaltspunkt für die Einschätzung der Margarinesorten bezüglich der gefundenen Faktoren, sondern erlauben darüber hinaus (im Fall einer 2- oder 3-Faktorlösung) eine *graphische Darstellung* der Faktorenergebnisse. Durch solche „Mappings“ lassen sich besonders gut die Positionen von Objekten (hier: Margarinemarken) im Hinblick auf die gefundenen Faktoren visualisieren (vgl. Abbildung 7.7).

Visualisierung

Dabei wird deutlich, dass es sich bei diesem „mapping“ um eine „relative“ Darstellung handelt: Die Faktorwerte werden als Abweichung von dem auf Null normierten

Abbildung 7.7: „Mapping“ der Faktorwerte

Mittelwert dargestellt, sodass hohe positive Faktorwerte stark überdurchschnittliche und hohe negative Faktorwerte stark unterdurchschnittliche Ausprägungen kennzeichnen. Über das absolute Niveau lässt sich nichts sagen.

Das in Abbildung 7.8 dargestellte Ablaufdiagramm enthält die wesentlichen Teilschritte bei der Durchführung einer Faktorenanalyse. Entsprechend diesem Ablaufdiagramm sind die nachfolgenden Betrachtungen aufgebaut. Allerdings ist zu beachten, dass sich bei konkreten Anwendungen der Faktorenanalyse insbesondere die Schritte (2) und (3) gegenseitig bedingen und nur schwer voneinander trennen lassen. Aus didaktischen Gründen wird hier aber eine Trennung vorgenommen.

Ablaufdiagramm

7.2 Vorgehensweise

Abbildung 7.8: Ablauf der Faktorenanalyse

7.2.1 Variablenauswahl und Korrelationsmatrix

Datenqualität

Die Güte der Ergebnisse einer Faktorenanalyse ist von der Qualität der Ausgangsdaten abhängig. Es muss deshalb besondere Sorgfalt auf die Wahl der Untersuchungsmerkmale verwendet werden. Insbesondere ist darauf zu achten, dass die erhobenen Merkmale auch für den Untersuchungsgegenstand relevant sind. Irrelevante Merkmale sind vorab auszusortieren und als ähnlich erachtete Kriterien müssen zusammengefasst werden. Insbesondere bei der Formulierung von Befragungsitems ist darauf zu achten, dass bereits die Wortwahl der Fragestellungen das Antwortverhalten der Befragten und damit die Streuung der Daten beeinflusst. Weiterhin sollten die Befragten einer möglichst homogenen Stichprobe entstammen, da die Höhe der Korrelationen zwischen den Untersuchungsmerkmalen (Variablen) durch den Homogenitätsgrad der Befragungsstichprobe beeinflusst wird.

Die oben aufgezeigten Sachverhalte schlagen sich insgesamt in den Korrelationen nieder, die als Maß für den Zusammenhang zwischen Variablen verwendet werden. Es wurden deshalb Prüfkriterien entwickelt, die es erlauben, Variablenzusammenhänge auf ihre Eignung für eine Faktorenanalyse zu überprüfen. Wir werden deshalb im Folgenden zunächst auf die Ermittlung von Korrelationen näher eingehen und sodann ausgewählte (statistische) Prüfkriterien erläutern.

7.2.1.1 Korrelationsanalyse zur Aufdeckung der Variablenzusammenhänge

Faktoren, die als „hinter den Variablen“ stehende latente Größen angesehen werden, repräsentieren den Zusammenhang zwischen verschiedenen Ausgangsvariablen. Bevor solche Faktoren ermittelt werden können, ist es zunächst erforderlich, die Zusammenhänge zwischen den Ausgangsvariablen messbar zu machen. Als methodisches Hilfsmittel wird hierzu die *Korrelationsrechnung* herangezogen.

Korrelationen

Bereits anhand der Korrelationen lässt sich erkennen, ob Zusammenhänge zwischen Paaren von Variablen bestehen, sodass Variablen als voneinander abhängig und damit als „bündelungsfähig“ angesehen werden können.

Für die Mittelwertmatrix (Abbildung 7.4) im obigen Beispiel lässt sich z. B. die Korrelation zwischen X_1 (Anteil ungesättigter Fettsäuren) und X_2 (Kaloriengehalt) wie folgt berechnen:

Korrelationskoeffizient:

Korrelationskoeffizient

$$r_{x_1,\,x_2} = \frac{\sum_{k=1}^{K}(x_{k1} - \overline{x}_1) \cdot (x_{k2} - \overline{x}_2)}{\sqrt{\sum_{k=1}^{K}(x_{k1} - \overline{x}_1)^2 \cdot \sum_{k=1}^{K}(x_{k2} - \overline{x}_2)^2}} \tag{7.1}$$

mit:

x_{k1} = Ausprägung der Variablen 1 bei Objekt k (in unserem Beispiel läuft k von 1 bis 6 (6 Marken))

$\overline{x}_1$ = Mittelwert der Ausprägung von Variable 1 über alle Objekte k

x_{k2} = Ausprägung der Variablen 2 bei Objekt k

$\overline{x}_2$ = Mittelwert der Ausprägung von Variable 2 über alle Objekte k

Setzt man in Formel (7.1) die entsprechenden Werte der Ausgangsdatenmatrix ein, so ergibt sich ein Korrelationskoeffizient von $r_{x_1,\,x_2} = 0,71176$. Um die im Einzelnen notwendigen Rechenschritte zu erleichtern, bedient man sich zur Ermittlung der Korrelationskoeffizienten am besten einer Hilfstabelle (Abbildung 7.9). Dabei stellt $\overline{x}_1$ den Mittelwert über alle Marken für die Eigenschaft „Ungesättigte Fettsäuren“ $((1+2+4+5+2+3) \;:\; 6 = 2,83)$ und $\overline{x}_2$ für die Eigenschaft „Kaloriengehalt“ $((1+6+5+6+3+4) \;:\; 6 = 4,17)$ dar.

	$(\mathbf{x_{k1} - \overline{x}_1})$	$(\mathbf{x_{k2} - \overline{x}_2})$	$\mathbf{(x_{k1} - \overline{x}_1)^2}$	$\mathbf{(x_{k2} - \overline{x}_2)^2}$	$\mathbf{(x_{k1} - \overline{x}_1) \cdot (x_{k2} - \overline{x}_2)}$
Rama	$-1,83333$	$-3,16667$	$3,36110$	$10,0278$	$5,80555$
Sanella	$-0,83333$	$1,83333$	$0,69444$	$3,36110$	$-1,52777$
Becel	$1,16667$	$0,83333$	$1,36112$	$0,69444$	$0,97222$
Du darfst	$2,16667$	$1,83333$	$4,69446$	$3,36110$	$3,97222$
Holl. Butter	$-0,83333$	$-1,16667$	$0,69444$	$1,36111$	$0,97222$
WB	$0,16667$	$-0,16667$	$0,02778$	$0,02778$	$-0,02778$
			$10,83334$	$18,83333$	$10,16666$
			$\sum_{k=1}^{6} (x_{k_1} - \overline{x}_1)^2$	$\sum_{k=1}^{6} (x_{k_2} - \overline{x}_2)^2$	$\sum_{k=1}^{6} (x_{k_1} - \overline{x}_1) \cdot (x_{k_2} - \overline{x}_2)$
	$r_{x_1,\,x_2} = \frac{10,16664}{\sqrt{10,83334 \cdot 18,83333}} = 0,71176$				

Abbildung 7.9: Hilfstabelle zur Berechnung eines Korrelationskoeffizienten

Berechnet man die Korrelationskoeffizienten über alle Eigenschaften, ergibt sich für die Mittelwertmatrix die in Abbildung 7.10 abgebildete Korrelationsmatrix.

	UNGEFETT	KALORIEN	VITAMIN	HALTBARK	PREIS
UNGEFETT	1.00000				
KALORIEN	**0.71176**	1.00000			
VITAMIN	0.96134	0.70397	1.00000		
HALTBARK	0.10894	0.13771	0.07825	1.00000	
PREIS	0.04385	0.06652	0.02362	0.98334	1.00000

Abbildung 7.10: Korrelationsmatrix für das 6-Produkt-Beispiel

In der Regel empfiehlt es sich, die Ausgangsdatenmatrix vorab zu standardisieren, da dadurch

Datenstandardisierung

- die Korrelationsrechnung und die im Rahmen der Faktorenanalyse erforderlichen Rechenschritte erleichtert werden;
- Interpretationserleichterungen erzielt werden;
- eine Vergleichbarkeit der Variablen ermöglicht wird, die in unterschiedlichen Maßeinheiten erhoben wurden (z. B. Einkommen gemessen in Euro und Verkauf von Gütern in Stck.).

Standardisierte Werte

Eine Standardisierung der Datenmatrix erfolgt durch die Bildung der Differenz zwischen dem Mittelwert und dem jeweiligen Beobachtungswert einer Variablen sowie der anschließenden Division durch die Standardabweichung. Dadurch wird sichergestellt, dass der neue Mittelwert gleich Null und die Standardabweichung einer Variablen gleich Eins ist. Die Werte einer standardisierten Datenmatrix bezeichnen wir im Folgenden nicht mehr mit x, sondern mit z.

Standardisierte Variable

$$z_{k_j} = \frac{x_{kj} - \overline{x}_j}{s_j}$$

mit:

x_{kj} = Beobachtungswert der j-ten Variablen bei Objekt k
$\overline{x}_j$ = Durchschnitt aller Beobachtungswerte der j-ten Variablen über alle Objekte
s_j = Standardabweichung der j-ten Variablen
z_{k_j} = Standardisierter Beobachtungswert der j-ten Variablen bei Objekt k

Korrelation von standardisierten Werten

Aus der standardisierten Datenmatrix ergibt sich auch eine einfachere Berechnung der Korrelationsmatrix $\mathbf{R}$ nach folgender Formel:

$$\mathbf{R} = \frac{1}{K-1} \cdot \mathbf{Z}' \cdot \mathbf{Z} \tag{7.2}$$

wobei $\mathbf{Z}'$ die transponierte Matrix der standardisierten Ausgangsdatenmatrix $\mathbf{Z}$ darstellt.

Der Leser möge selbst anhand des Beispiels die Gültigkeit der Formel überprüfen. Dabei wird klar werden, dass die Korrelationsmatrix auf *Basis der Ausgangsdaten identisch* ist mit der Korrelationsmatrix auf *Basis der standardisierten Daten*. Wird die Korrelationsmatrix aus *standardisierten* Daten errechnet, so sind in diesem Falle Varianz-Kovarianzmatrix und Korrelationsmatrix *identisch* . Für den Korrelationskoeffizienten lässt sich auch schreiben:

Korrelation und Kovarianz

$$r_{x_1,\,x_2} = \frac{S_{x_1,\,x_2}}{S_{x_1} S_{x_2}} \qquad \text{mit:} \qquad S_{x_1,\,x_2} = \frac{1}{K-1} \sum_k (x_{k1} - \overline{x}_1)(x_{k2} - \overline{x}_2)$$

Da wegen der Standardisierung die beiden Varianzen im Nenner 1 sind, folgt, dass Korrelationskoeffizient und Kovarianz ($S_{x_1,\,x_2}$) identisch sind.

Die Korrelationsmatrix zeigt dem Anwender auf, welche Variablen der Ausgangsbefragung offenbar mit welchen anderen Variablen dieser Befragung „irgendwie zusammenhängen“. Sie zeigt ihm jedoch *nicht*, ob

1. die Variablen sich gegenseitig bedingen

   oder

2. das Zustandekommen der Korrelationswerte durch einen oder mehrere hinter den zusammenhängenden Variablen stehenden Faktoren bestimmt wird.

Angesichts der beiden klar trennbaren Blöcke der Korrelationsmatrix (vgl. die abgegrenzten Vierecke in Abb. 7.10) lässt sich vermuten, dass die Variablen x_1 bis x_3 und x_4/x_5 durch zwei Faktoren „erklärt“ werden könnten.

Eigenwert

Ausgehend von dieser *Hypothese* stellt sich unmittelbar die Frage, mit welchem Gewicht denn die beiden Faktoren an der Beschreibung der beobachteten Zusammenhänge beteiligt sind. Es ist ja denkbar, dass der Faktor „Gesundheit“ als alleiniger Beschreibungsfaktor für die Variablen x_1 bis x_3 fast für die gesamten Unterschiede in der Ausgangsbefragung verantwortlich ist. Es kann aber auch sein, dass er nur einen Teil der unterschiedlichen Beurteilungen in der Ausgangsbefragung erklärt. Die größere oder geringere Bedeutung beider Faktoren lässt sich in einer Gewichtszahl ausdrücken, die im Rahmen einer Faktorenanalyse auch als *Eigenwert* bezeichnet wird.

7.2.1.2 Eignung der Korrelationsmatrix

Zu Beginn des Abschnittes 7.2.1 hatten wir bereits darauf hingewiesen, dass sich die Eignung der Ausgangsdaten für faktoranalytische Zwecke in der Korrelationsmatrix (K-M) widerspiegelt. Dabei liefern bereits die *Ausgangsdaten* selbst einen Anhaltspunkt zur Eignungsbeurteilung der Daten zum Zwecke der Faktorenanalyse, da die Höhe der Korrelationskoeffizienten durch die Verteilung der Variablen in der Erhebungsgesamtheit (Symmetrie, Schiefe und Wölbung der Verteilung) beeinflusst wird.

Eignung der K-M für eine Faktorenanalyse

Liegt einer Erhebung eine heterogene Datenstruktur zugrunde, so macht sich dies durch viele kleine Werte in der Korrelationsmatrix bemerkbar, womit eine sinnvolle Anwendung der Faktorenanalyse in Frage gestellt ist. Es ist deshalb *vorab* eine Prüfung der Variablen auf Normalverteilung, zumindest aber auf Gleichartigkeit der Verteilungen empfehlenswert, obwohl die Faktorenanalyse selbst keine Verteilungsannahmen setzt.

Bezogen auf unser 6-Produkte-Beispiel treten neben sehr hohen Werten ($> 0,7$) insbesondere im unteren Teil der Matrix kleine Korrelationen auf (vgl. Abbildung 7.10), sodass die Korrelationsmatrix selbst kein eindeutiges Urteil über die Eignung der Daten zur Faktorenanalyse zulässt.

Es ist deshalb zweckmäßig, weitere Kriterien zur Prüfung heranzuziehen. Hierzu bieten sich insbesondere statistische Prüfkriterien an, die eine Überprüfung der Korrelationskoeffizienten auf Eignung zur Faktorenanalyse ermöglichen. Es ist durchaus empfehlenswert, mehr als ein Kriterium zur faktoranalytischen Eignung der Datenmatrix anzuwenden, da die verschiedenen Kriterien unterschiedliche Vor- und Nachteile haben. Im Einzelnen werden durch SPSS folgende Kriterien bereitgestellt:

Signifikanzprüfung der Korrelationen

Signifikanzniveaus beschreiben die Wahrscheinlichkeit, mit der eine zuvor formulierte Hypothese zutrifft oder nicht. Für alle Korrelationskoeffizienten lassen sich die Signifikanzniveaus anführen. Zuvor wird die sogenannte *Nullhypothese* (H_0) formuliert, die aussagt, dass kein Zusammenhang zwischen den Variablen besteht. Das Signifikanzniveau des Korrelationskoeffizienten beschreibt anschließend, mit welcher *Irrtumswahrscheinlichkeit* eben diese Nullhypothese abgelehnt werden kann. Ein beispielhaftes Signifikanzniveau von $0,01$ bedeutet, dass mit dieser *Irrtumswahrscheinlichkeit* die Nullhypothese abgelehnt werden kann, sprich zu 1 % wird sich der Anwender täuschen, wenn er von einem Zusammenhang ungleich Null zwischen den Variablen ausgeht. Anders ausgedrückt: Mit einer Wahrscheinlichkeit von 99 % wird sich die Korrelation von Null unterscheiden.

Irrtumswahrscheinlichkeit

Korrelationsmatrix

		Anteil ungesättigter Fettsäuren	Kalorien-gehalt	Vitamin-gehalt	Haltbarkeit	Preis
Sig. (1-seitig)	Anteil ungesättigter Fettsäuren		,05632	,00111	,41862	,46713
	Kaloriengehalt	,05632		,05924	,39737	,45018
	Vitamingehalt	,00111	,05924		,44144	,48229
	Haltbarkeit	,41862	,39737	,44144		,00021
	Preis	,46713	,45018	,48229	,00021	

Abbildung 7.11: Signifikanzniveaus der Korrelationskoeffizienten im 6-Produkte-Beispiel

Für unser Beispiel zeigt Abbildung 7.11, dass sich tendenziell diejenigen Korrelationskoeffizienten signifikant von Null unterscheiden (niedrige Werte in Abbildung 7.11), die in Abbildung 7.10 hohe Werte ($> 0,7$) aufweisen, während die Korrelationskoeffizienten mit geringen Werten auch eine geringere Signifikanz (Werte $> 0,4$) besitzen. Das bedeutet, dass sich z. B. die Korrelation zwischen den Variablen „Vitamingehalt“ und „Haltbarkeit“ nur mit einer Wahrscheinlichkeit von $(1 - 0,44 =)$ 56 % von Null unterscheidet. Anzumerken ist hier, dass es sich um **p-Werte der Korrelationen** handelt, die von SPSS als „Signifikanz“ ausgewiesen werden. Der p-Wert beschreibt die Wahrscheinlichkeit, unter Gültigkeit der Nullhypothese das erhaltene Ergebnis oder ein extremeres zu erhalten, und ist eng verbunden mit dem Signifikanzniveau des klassischen Hypothesentests (kritische Irrtumswahrscheinlichkeit α). Je niedriger der p-Wert, desto eher ist die Nullhypothese abzulehnen.

Inverse der Korrelationsmatrix

Die Eignung einer Korrelationsmatrix für die Faktorenanalyse lässt sich weiterhin an der Struktur der Inversen der Korrelationsmatrix erkennen. Dabei wird davon ausgegangen, dass eine Eignung dann gegeben ist, wenn die *Inverse eine Diagonalmatrix* darstellt, d. h. die Nicht-diagonal-Elemente der inversen Korrelationsmatrix möglichst nahe bei Null liegen. Für das 6-Produkte-Beispiel zeigt Abbildung 7.12, dass insbesondere für die Werte der Variablen „Ungesättigte Fettsäuren“ und „Vitamingehalt“ sowie „Haltbarkeit“ und „Preis“ hohe Werte auftreten, während alle anderen Werte *relativ* nahe bei Null liegen. Es existiert allerdings kein allgemeingültiges Kriterium dafür, wie stark und wie häufig die Nicht-diagonal-Elemente von Null abweichen dürfen.

Merkmale der Inversen

Bartlett-Test (test of sphericity)

Der Bartlett-Test überprüft die Hypothese, dass die Stichprobe aus einer Grundgesamtheit entstammt, in der die Variablen unkorreliert sind.[5]

Gleichbedeutend mit dieser Aussage ist die Frage, ob die Korrelationsmatrix nur zufällig von einer Einheitsmatrix abweicht, da im Falle der Einheitsmatrix alle Nicht-

[5] Vgl. Dziuban/Shirkey (1974), S. 358 ff.

Inverse Korrelationsmatrix

	Anteil ungesättigter Fettsäuren	Kalorien -gehalt	Vitamin-gehalt	Haltbarkeit	Preis
Anteil ungesättigter Fettsäuren	14,499	-,607	-13,198	-5,580	5,204
Kaloriengehalt	-,607	2,179	-,829	-2,181	2,046
Vitamingehalt	-13,198	-,829	14,000	4,793	-4,410
Haltbarkeit	-5,580	-2,181	4,793	38,179	-37,266
Preis	5,204	2,046	-4,410	-37,266	37,385

Abbildung 7.12: Inverse der Korrelationsmatrix im 6-Produkte-Beispiel

diagonal-Elemente Null sind, d. h. keine Korrelationen zwischen den Variablen vorliegen. Es werden folgende Hypothesen formuliert:

H_0: Die Variablen in der Erhebungsgesamtheit sind unkorreliert.

H_1: Variablen in der Erhebungsgesamtheit sind korreliert.

Der Bartlett-Test setzt voraus, dass die Variablen in der Erhebungsgesamtheit einer *Normalverteilung* folgen und die entsprechende Prüfgröße annähernd Chi-Quadrat-verteilt ist. Letzteres aber bedeutet, dass der Wert der Prüfgröße in hohem Maße durch die Größe der Stichprobe beeinflusst wird. Für unser Beispiel erbrachte der Bartlett-Test eine Prüfgröße von $17,371$ bei einem Signifikanzniveau von $0,0665$. Das bedeutet, dass mit einer Wahrscheinlichkeit von $(1 - 0,0665 =)$ $93,35$ % davon auszugehen ist, dass die Variablen der Erhebungsgesamtheit korreliert sind. Setzt man als kritische Irrtumswahrscheinlichkeit einen Wert von $0,05$ fest, so wäre für unser Beispiel die Nullhypothese anzunehmen. Es ist folglich davon auszugehen, dass kein systematischer Unterschied zwischen der Korrelationsmatrix und der Einheitsmatrix besteht, oder mit anderen Worten, dass sich die Korrelationsmatrix nur zufällig von der Einheitsmatrix unterscheidet. Das lässt den Schluss zu, dass die Ausgangsvariablen in unserem Fall unkorreliert sind.

Voraussetzungen

Allerdings sei an dieser Stelle nochmals darauf hingewiesen, dass die Anwendung des Bartlett-Tests eine Prüfung der Ausgangsdaten auf Normalverteilung voraussetzt, die in unserem Fall noch erfolgen müsste.

Anti-Image-Kovarianz-Matrix

Der Begriff Anti-Image stammt aus der Image-Analyse von Guttmann.[6] Guttmann geht davon aus, dass sich die Varianz einer Variablen in zwei Teile zerlegen lässt: das Image und das Anti-Image.

Das *Image* beschreibt dabei den Anteil der Varianz, der durch die verbleibenden Variablen mit Hilfe einer multiplen Regressionsanalyse (vgl. Kapitel 1) erklärt werden kann, während das *Anti-Image* denjenigen Teil darstellt, der von den übrigen

Multiple Regression

[6] Vgl. Guttmann (1953), S. 277 ff.

Variablen unabhängig ist. Da die Faktorenanalyse unterstellt, dass den Variablen gemeinsame Faktoren zugrunde liegen, ist es unmittelbar einsichtig, dass Variablen nur dann für eine Faktorenanalyse geeignet sind, wenn das Anti-Image der Variablen möglichst gering ausfällt. Das aber bedeutet, dass die Nicht-diagonal-Elemente der Anti-Image-Kovarianz-Matrix möglichst nahe bei Null liegen müssen bzw. diese Matrix eine Diagonalmatrix darstellen sollte. Für das 6-Produkte-Beispiel zeigt Abbildung 7.13, dass die Forderung nach einer entsprechenden Diagonalmatrix erfüllt ist.

Anti-Image-Matrizen

		Anteil ungesättigter Fettsäuren	Kalorien-gehalt	Vitamin-gehalt	Haltbarkeit	Preis
Anti-Image-Kovarianz	Anteil ungesättigter Fettsäuren	,06897	-,01920	-,06502	-,01008	,00960
	Kaloriengehalt	-,01920	,45883	-,02716	-,02621	,02511
	Vitamingehalt	-,06502	-,02716	,07143	,00897	-,00842
	Haltbarkeit	-,01008	-,02621	,00897	,02619	-,02611
	Preis	,00960	,02511	-,00842	-,02611	,02675

Abbildung 7.13: Anti-Image-Kovarianz-Matrix im 6-Produkte-Beispiel

Kritische Werte

Als Kriterium dafür, wann die Forderung nach einer solchen Diagonalmatrix erfüllt ist, schlagen Dziuban und Shirkey vor, die Korrelationsmatrix dann als für die Faktorenanalyse ungeeignet anzusehen, wenn der Anteil der Nicht-diagonal-Elemente, die ungleich Null sind ($> 0,09$), in der Anti-Image-Kovarianzmatrix (AIC) 25 % oder mehr beträgt.[7] Das trifft in unserem Fall für keines der Nicht-diagonal-Elemente der AIC-Matrix zu, womit nach diesem Kriterium die Korrelationsmatrix für faktoranalytische Auswertungen geeignet ist.

Measure of Sampling Adequacy und Kaiser-Meyer-Olkin-Kriterium

Während die Überlegungen von Dziuban und Shirkey auf Plausibilität beruhen, haben Kaiser, Meyer und Olkin versucht, eine geeignete Prüfgröße zu entwickeln und diese zur Entscheidungsfindung heranzuziehen. Sie berechnen ihre Prüfgröße, die als *„measure of sampling adequacy (MSA)“* bezeichnet wird, auf Basis der Anti-Image-Korrelationsmatrix. Das MSA-Kriterium zeigt an, in welchem Umfang die Ausgangsvariablen zusammengehören und dient somit als Indikator dafür, ob eine Faktorenanalyse sinnvoll erscheint oder nicht.

MSA-Kriterium

Das MSA-Kriterium erlaubt sowohl eine Beurteilung der Korrelationsmatrix insgesamt als auch einzelner Variablen; sein Wertebereich liegt zwischen 0 und 1. Kaiser und Rice schlagen folgende Beurteilungen vor:[8]

[7] Vgl. Dziuban/Shirkey (1974), S. 359.
[8] Vgl. Kaiser/Rice (1974), S. 111 ff.

MSA $\geq 0,9$	marvelous	(„wunderbar“)
MSA $\geq 0,8$	meritorious	(„verdienstvoll“)
MSA $\geq 0,7$	middling	(„ziemlich gut“)
MSA $\geq 0,6$	mediocre	(„mittelmäßig“)
MSA $\geq 0,5$	miserable	(„kläglich“)
MSA $< 0,5$	unacceptable	(„untragbar“)

Sie vertreten die Meinung, dass sich eine Korrelationsmatrix mit MSA $< 0,5$ nicht für eine Faktorenanalyse eignet.[9] Als wünschenswert sehen sie einen Wert von MSA $\geq 0,8$ an.[10] Durch Aggregation der variablenspezifischen MSA-Werte ergibt sich das Kaiser-Meyer-Olkin-(KMO-)Kriterium, das – ebenso wie der Bartlett-Test – Auskunft über die Zusammengehörigkeit der Variablen insgesamt gibt.[11] Das KMO-Kriterium wird als das beste zur Verfügung stehende Verfahren zur Prüfung der Korrelationsmatrix angesehen, weshalb seine Anwendung vor der Durchführung einer Faktorenanalyse auf jeden Fall zu empfehlen ist.[12]

KMO-Kriterium

Bezogen auf unser 6-Produkte-Beispiel ergab sich für die Korrelationsmatrix insgesamt ein MSA-Wert von $0,576$, womit sich für unser Beispiel ein nur „klägliches“ Ergebnis ergibt. Darüber hinaus gibt SPSS in der Diagonalen der Anti-Image-Korrelationsmatrix aber auch das MSA-Kriterium für die einzelnen Variablen an.

Anti-Image-Matrizen

		Anteil ungesättigter Fettsäuren	Kalorien-gehalt	Vitamin-gehalt	Haltbarkeit	Preis
Anti-Image-Korrelation	Anteil ungesättigter Fettsäuren	,59680[a]	-,10794	-,92633	-,23717	,22350
	Kaloriengehalt	-,10794	,87789[a]	-,15001	-,23906	,22661
	Vitamingehalt	-,92633	-,15001	,59755[a]	,20730	-,19274
	Haltbarkeit	-,23717	-,23906	,20730	,47060[a]	-,98640
	Preis	,22350	,22661	-,19274	-,98640	,46701[a]

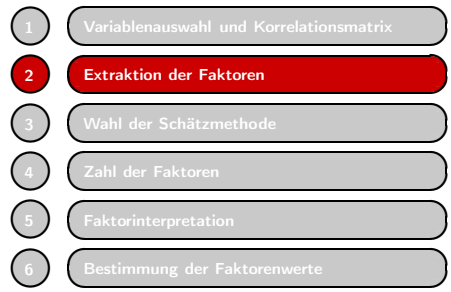
a. Maß der Stichprobeneignung

Abbildung 7.14: Anti-Image-Korrelations-Matrix im 6-Produkte-Beispiel

Abbildung 7.14 macht deutlich, dass lediglich die Variable „Kaloriengehalt“ mit einem MSA-Wert von $0,87789$ als „verdienstvoll“ anzusehen ist, während alle übrigen Variablen eher „klägliche“ oder „untragbare“ Ergebnisse aufweisen. Die variablenspezifischen MSA-Werte liefern damit für den Anwender einen Anhaltspunkt dafür, welche Variablen aus der Analyse auszuschließen wären, wobei sich ein sukzessiver Ausschluss von Variablen mit jeweiliger Prüfung der vorgestellten Kriterien empfiehlt.

[9] Vgl. Cureton/D'Agostino (1993), S. 389 f.
[10] Vgl. Kaiser (1970), S. 405.
[11] Vgl. Weiber/Mühlhaus (2014), S. 132 f.
[12] Vgl. Stewart (1981), S. 57 f.; Dziuban/Shirkey (1974), S. 360 f.

7.2.2 Extraktion der Faktoren



Die bisherigen Ausführungen haben verdeutlicht, dass bei Faktorenanalysen große Sorgfalt auf die Wahl der Untersuchungsmerkmale und -einheiten zu verwenden ist, da durch die Güte der Ausgangsdaten, die den Startpunkt der Faktorenanalyse darstellen, alle Ergebnisse der Faktorenanalyse beeinflusst werden. Im Folgenden ist nun zu fragen, wie denn nun die Faktoren rein rechnerisch aus den Korrelationen ermittelt werden können. Wir werden zunächst das *Fundamentaltheorem der Faktorenanalyse* darstellen und anschließend die Extraktion auf graphischem Wege plausibel machen. Auf die Unterschiede zwischen konkreten (rechnerischen) Faktorextraktionsverfahren gehen wir dann im Zusammenhang mit der Wahl der Schätzmethode (Abschnitt 7.2.3) ein.

Fundamentaltheorem

7.2.2.1 Das Fundamentaltheorem

Während die bisherigen Überlegungen die Ausgangsdaten und ihre Eignung für faktoranalytische Zwecke betrafen, stellt sich nun die Frage, wie sich die Faktoren rechnerisch aus den Variablen ermitteln lassen. Zu diesem Zweck geht die Faktorenanalyse von der grundlegenden Annahme aus, dass jeder Beobachtungswert einer Ausgangsvariablen x_j oder der standardisierten Variablen z_j sich als eine Linearkombination mehrerer (hypothetischer) Faktoren beschreiben lässt.

Linearitätsannahme

Mathematisch lässt sich dieser Zusammenhang wie folgt formulieren:

$$x_{kj} = a_{j1} \cdot p_{k1} + a_{j2} \cdot p_{k2} + ... + a_{jQ} \cdot p_{kQ} \tag{7.3}$$

bzw. für standardisierte x-Werte

$$z_{kj} = a_{j1} \cdot p_{k1} + a_{j2} \cdot p_{k2} + ... + a_{jQ} \cdot p_{kQ} = \sum_{q=1}^{Q} a_{jq} \cdot p_{kq} \tag{7.4}$$

Die obige Formel (7.4) besagt für das 2-Faktorenbeispiel nichts anderes, als dass z. B. die standardisierten Beobachtungswerte für „Anteil ungesättigter Fettsäuren“ und „Vitamingehalt“ beschrieben werden durch die Faktoren p_1 und p_2, so wie sie im Hinblick auf Marke k gesehen wurden (p_{k1} bzw. p_{k2}), jeweils multipliziert mit ihren Gewichten bzw. Faktorenladungen beim Merkmal j, also für Faktor 1 a_{j1} und für Faktor 2 a_{j2}.

Formelinterpretation

Die Faktorladung gibt dabei an, *wieviel* ein Faktor mit einer Ausgangsvariablen zu tun hat. Im mathematisch-statistischen Sinne sind Faktorladungen nichts anderes als eine *Maßgröße für den Zusammenhang zwischen Variablen und Faktor*, und das ist wiederum nichts anderes als ein *Korrelationskoeffizient zwischen Faktor und Variablen.*

Um die Notation zu verkürzen, schreibt man häufig den Ausdruck (7.4) auch in Matrixschreibweise. Identisch mit Formel (7.4) ist daher auch folgende Matrixschreibweise, die die *Grundgleichung der Faktorenanalyse* darstellt. Dabei wird unterstellt, dass in der Matrix der Ausgangsdaten **Z** die Fälle in den Zeilen und die Variablen in

Matrixschreibweise

den Spalten aufgeführt sind:

$$\mathbf{Z} = \mathbf{P} \cdot \mathbf{A}' \tag{7.5}$$

Aufbauend auf diesem Grundzusammenhang lässt sich dann auch eine *Rechenvorschrift* ableiten, die aufzeigt, wie aus den erhobenen Daten die vermuteten Faktoren mathematisch ermittelt werden können.

Wir hatten gezeigt, dass die Korrelationsmatrix $\mathbf{R}$ sich bei standardisierten Daten wie folgt aus der Datenmatrix $\mathbf{Z}$ ermitteln lässt:

$$\mathbf{R} = \frac{1}{K-1} \cdot \mathbf{Z}' \cdot \mathbf{Z} \tag{7.6}$$

Da $\mathbf{Z}$ aber im Rahmen der Faktorenanalyse durch $\mathbf{P} \cdot \mathbf{A}'$ beschrieben wird ($\mathbf{Z} = \mathbf{P} \cdot \mathbf{A}'$), ist in (7.6) $\mathbf{Z}$ durch Formel (7.5) zu ersetzen, sodass sich folgender Zusammenhang ergibt:

$$\mathbf{R} = \frac{1}{K-1} \cdot (\mathbf{P} \cdot \mathbf{A}')' \cdot (\mathbf{P} \cdot \mathbf{A}') \tag{7.7}$$

Nach Auflösung der Klammern ergibt sich nach den Regeln der Matrixmultiplikation:

$$\mathbf{R} = \frac{1}{K-1} \cdot \mathbf{A} \cdot \mathbf{P}' \cdot \mathbf{P} \cdot \mathbf{A}' = \mathbf{A} \cdot \overbrace{\frac{1}{K-1} \cdot \mathbf{P}' \cdot \mathbf{P}}^{} \cdot \mathbf{A}' \tag{7.8}$$

Da alle Daten standardisiert sind, lässt sich der $\overbrace{\frac{1}{K-1} \cdot \mathbf{P}' \cdot \mathbf{P}}$ Ausdruck in Formel (7.8) auch als *Korrelationsmatrix der Faktoren* ($\mathbf{C}$) bezeichnen (vgl. Formel (7.6)), sodass sich schreiben lässt:

$$\mathbf{R} = \mathbf{A} \cdot \mathbf{C} \cdot \mathbf{A}' \tag{7.9}$$

Da die Faktoren als unkorreliert angenommen werden, entspricht $\mathbf{C}$ einer Einheitsmatrix (einer Matrix, die auf der Hauptdiagonalen nur Einsen und sonst Nullen enthält). Da die Multiplikation einer Matrix mit einer Einheitsmatrix aber wieder die Ausgangsmatrix ergibt, vereinfacht sich die Formel (7.9) zu:

$$\mathbf{R} = \mathbf{A} \cdot \mathbf{A}' \tag{7.10}$$

Die Beziehungen (7.9) und (7.10) werden von Thurstone – neben Spearmann einer der Begründer der Faktorenanalyse – als das *Fundamentaltheorem der Faktorenanalyse* bezeichnet, da sie den Zusammenhang zwischen Korrelationsmatrix und Faktorladungsmatrix beschreiben.

Fundamentaltheorem

Das Fundamentaltheorem der Faktorenanalyse besagt nicht anderes, als dass sich die Korrelationsmatrix durch die Faktorladungen (Matrix $\mathbf{A}$) und die Korrelationen zwischen den Faktoren (Matrix $\mathbf{C}$) reproduzieren lässt. Für den Fall, dass man von unabhängigen (orthogonalen) Faktoren ausgeht, reduziert sich das Fundamentaltheorem auf Formel (7.10). Dabei muss sich der Anwender allerdings bewusst sein, dass das Fundamentaltheorem der Faktorenanalyse nach Formel (7.10) stets nur unter der Prämisse einer Linearverknüpfung und Unabhängigkeit der Faktoren Gültigkeit besitzt.

7.2.2.2 Graphische Interpretation von Faktoren

Der Informationsgehalt einer Korrelationsmatrix lässt sich auch graphisch in einem Vektor-Diagramm darstellen, in dem die jeweiligen Korrelationskoeffizienten als Winkel zwischen zwei Vektoren dargestellt werden. Zwei Vektoren werden dann als linear unabhängig bezeichnet, wenn sie senkrecht (orthogonal) aufeinander stehen. Sind die beiden betrachteten Vektoren (Variablen) jedoch korreliert, ist der Korrelationskoeffizient also $\neq 0$, z. B. $0,5$, dann wird dies graphisch durch einen Winkel von 60° zwischen den beiden Vektoren dargestellt.

Korrelation und Winkel

Es stellt sich die Frage: Warum entspricht ein Korrelationskoeffizient von $0,5$ genau einem Winkel von 60°? Die Verbindung wird über den Cosinus des jeweiligen Winkels hergestellt.

Vektorbetrachtung

Verdeutlichen wir uns dies anhand des Ausgangsbeispiels (Abbildung 7.15): In Abbildung 7.15 repräsentieren die Vektoren $\overline{AC}$ und z. B. $\overline{AB}$ die beiden Variablen *Kaloriengehalt* und *Vitamingehalt*. Zwischen den beiden Variablen möge eine Korrelation von $0,5$ gemessen worden sein. Der Vektor $\overline{AC}$, der den Kaloriengehalt repräsentiert und der genau wie $\overline{AB}$ aufgrund der Standardisierung eine Länge von 1 hat, weist zu $\overline{AB}$ einen Winkel von 60° auf. Der Cosinus des Winkels 60°, der die Stellung der beiden Variablen zueinander (ihre Richtung) angibt, ist definiert als Quotient aus Ankathete und Hypothenuse, also als $\overline{AD}/\overline{AC}$. Da $\overline{AC}$ aber gleich 1 ist, ist der Korrelationskoeffizient identisch mit der Strecke $\overline{AD}$.

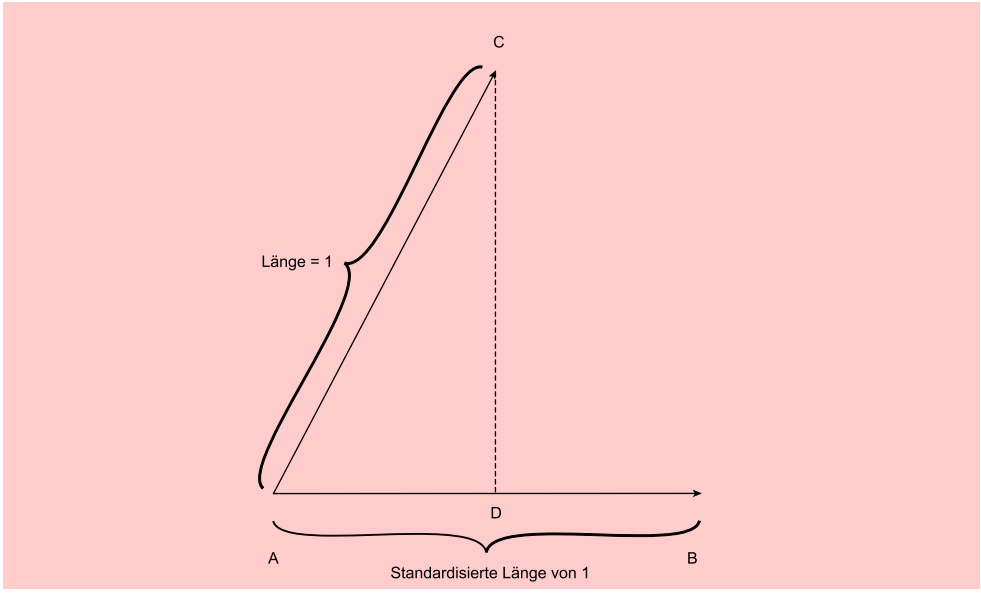


Abbildung 7.15: Vektordarstellung einer Korrelation zwischen zwei Variablen

Wie Abbildung 7.16 ausschnitthaft zeigt, ist z. B. der Cosinus eines 60°-Winkels gleich $0,5$. Entsprechend lässt sich jeder beliebige Korrelationskoeffizient zwischen zwei Variablen auch durch zwei Vektoren mit einem genau definierten Winkel zueinander darstellen. Verdeutlichen wir uns dies noch einmal anhand einer Korrelations-

Grad	cos	Grad	cos
45	0,7071	**90**	**0,0000**
44	7193	89	0175
43	7314	88	0349
42	7431	87	0523
41	7547	86	0698
40	**0,7660**	85	0872
39	7771	84	1045
38	7880	83	1219
37	7986	82	1392
36	8090	81	1564
35	8192	**80**	**0,1736**
34	8290	79	1908
33	8387	78	2079
32	8480	77	2250
31	8572	76	2419
30	**0,8660**	75	2588
29	8746	74	2756
28	8829	73	2924
27	8910	72	3090
26	8988	71	3256
25	9063	**70**	**0,3420**
24	9135	69	3584
23	9205	68	3746
22	9272	67	3907
21	9336	66	4067
20	**0,9397**	65	4226
19	9455	64	4384
18	9511	63	4540
17	9563	62	4695
16	9613	61	4848
15	9659	**60**	**0,5000**
14	9703	59	5150
13	9744	58	5299
12	9781	57	5446
11	9816	56	5592
10	**0,9848**	55	5736
9	9877	54	5878
8	9903	53	6018
7	9925	52	6157
6	9945	51	6293
5	9962	**50**	**0,6428**
4	9976	49	6561
3	9986	48	6691
2	9994	47	6820
1	9998	46	6947
0	**1,0000**	45	7071

Abbildung 7.16: Werte für den Cosinus (entommen aus: Gellert et al. (1977), S. 786)

matrix mit drei Variablen (Formel 7.11).

$$R = \begin{pmatrix} 1 & & \\ 0,8660 & 1 & \\ 0,1736 & 0,6428 & 1 \end{pmatrix} \tag{7.11}$$

R lässt sich auch anders schreiben (vgl. 7.12).

$$R = \begin{pmatrix} 0° & & \\ 30° & 0° & \\ 80° & 50° & 0° \end{pmatrix} \tag{7.12}$$

Der Leser möge die entsprechenden Werte selbst in einer Cosinus-Tabelle überprüfen.

Die der oben gezeigten Korrelationsmatrix zugrundeliegenden drei Variablen und ihre Beziehungen zueinander lassen sich relativ leicht in einem zweidimensionalen Raum darstellen (Abbildung 7.17).

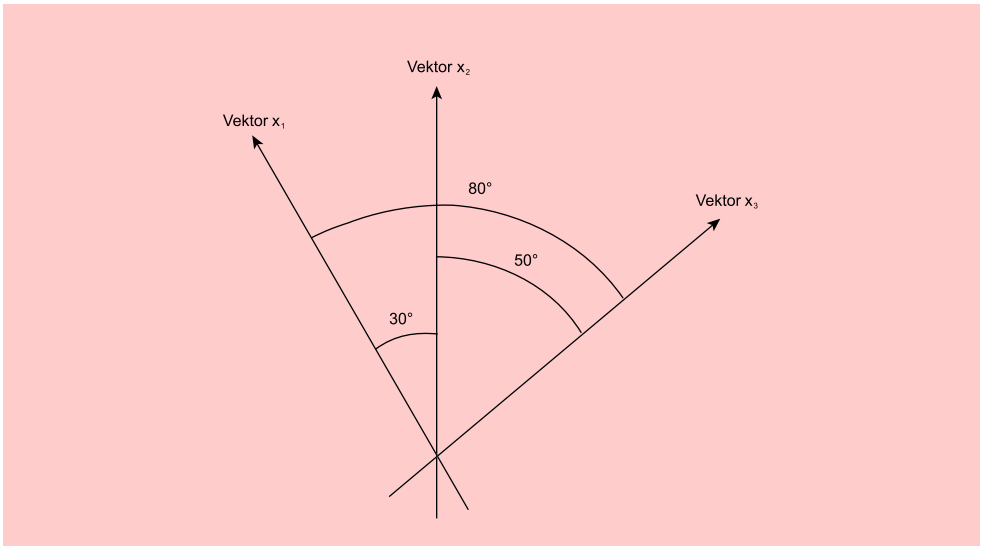


Abbildung 7.17: Graphische Darstellung des 3-Variablen-Beispiels

Dimensionsreduktion

Je mehr Variablen jedoch zu berücksichtigen sind, desto mehr Dimensionen werden benötigt, um die Vektoren in ihren entsprechenden Winkeln zueinander zu positionieren. Die Faktorenanalyse trachtet danach, das über die Korrelationskoeffizienten gemessene Verhältnis der Variablen zueinander *in einem möglichst gering dimensionierten Raum* zu reproduzieren. Die Zahl der benötigten Achsen gibt dann die entsprechende Zahl der Faktoren an.

Wenn man die Achsen als Faktoren ansieht, dann stellt sich unmittelbar die Frage: Wie werden diese Achsen (Faktoren) in ihrer Lage zu den jeweiligen Vektoren (Variablen) bestimmt?

Faktorvektor

Dazu vergegenwärtigt man sich am besten das Bild eines halboffenen Schirmes. Die Zacken des Schirmgestänges, die alle in eine bestimmte Richtung weisend die Variablen repräsentieren, lassen sich näherungsweise auch durch den Schirmstock darstellen. Vereinfacht man diese Überlegung aus Darstellungsgründen noch weiter auf den

2-Variablen-Fall wie in Abbildung 7.18, die einen Korrelationskoeffizienten von $0,5$ für die durch die Vektoren $\overline{OA}$ und $\overline{OB}$ dargestellten Variablen repräsentiert, dann gibt der Vektor $\overline{OC}$ eine zusammenfassende (faktorielle) Beschreibung wieder. Die beiden Winkel von 30° zwischen Vektor I bzw. Vektor II und Faktor-Vektor geben wiederum an, inwieweit der gefundene Faktor mit Vektor (Variable) I bzw. II zusammenhängt. Sie repräsentieren ebenfalls Korrelationskoeffizienten, und zwar die zwischen den jeweiligen Variablen und dem Faktor. Diese Korrelationskoeffizienten hatten wir oben als Faktorladungen bezeichnet. Die Faktorladungen des 1. Faktors betragen also in Bezug auf Variable I und Variable II: $\cos 30° = 0,8660$.

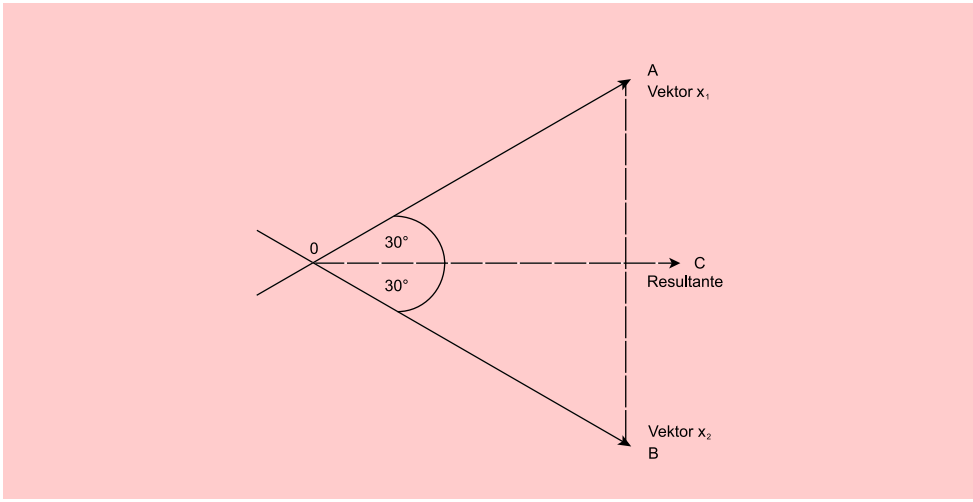


Abbildung 7.18: Faktorlösung bei 2 Variablen

7.2.2.3 Das Problem der Faktorextraktion

Nachdem wir nun wissen, was eine *Faktorladung* inhaltlich bedeutet, ist zu fragen: Wie findet man einen solchen Vektor (Faktor), der stellvertretend für mehrere zusammenhängende Variable fungieren kann? Erinnern wir uns noch einmal des Ausgangsbeispiels. Aufstrichfette waren nach den fünf Merkmalen

- Anteil ungesättigter Fettsäuren
- Kaloriengehalt
- Vitamingehalt
- Haltbarkeit
- Preis

bewertet worden.[13] Aus dieser Bewertung sei die Korrelationsmatrix in Abbildung 7.19 berechnet worden.

[13] Es werden hier andere Werte als im Ausgangsbeispiel verwendet, um zunächst eine eindeutige graphische Lösung zu ermöglichen.

	X_1	X_2	X_3	X_4	X_5
X_1		$10°$	$70°$	$90°$	$100°$
X_2	$0,9848$		$60°$	$80°$	$90°$
X_3	$0,3420$	$0,5000$		$20°$	$30°$
X_4	$0,0000$	$0,1736$	$0,9397$		$10°$
X_5	$-0,1736$	$0,0$	$0,8660$	$0,9848$	

Abbildung 7.19: Spiegelbildlich identische Korrelationsmatrix

Diese Korrelationsmatrix enthält in der unteren Dreiecks-Matrix die Korrelationskoeffizienten, in der oberen (spiegelbildlich identischen) Dreiecks-Matrix die entsprechenden Winkel. Graphisch ist der Inhalt dieser Matrix in Abbildung 7.20 dargestellt.

Graphische Darstellung einer Faktorextraktion

Das Beispiel wurde so gewählt, dass die Winkel zwischen den Faktoren in einer zweidimensionalen Darstellung abgebildet werden können – ein Fall, der in der Realität allerdings nur relativ selten reproduzierbar sein wird.

Wie findet man nun den 1. Faktor in dieser Vektordarstellung? Bleiben wir zunächst bei der graphischen Darstellung, dann sucht man den Schwerpunkt aus den fünf Vektoren. Der Leser möge sich dazu folgendes verdeutlichen:

In Abbildung 7.20 ist der Faktor nichts anderes als die Resultante der fünf Vektoren. Würden die fünf Vektoren fünf Seile darstellen mit einem Gewicht in O und jeweils eine Person würde mit gleicher Stärke an den Enden der Seile ziehen, dann würde sich das Gewicht in eine bestimmte Richtung bewegen (vgl. die gestrichelte Linie in Abbildung 7.21). Diesen Vektor bezeichnen wir als Resultante. Er ist die graphische Repräsentation des 1. Faktors.

Resultante

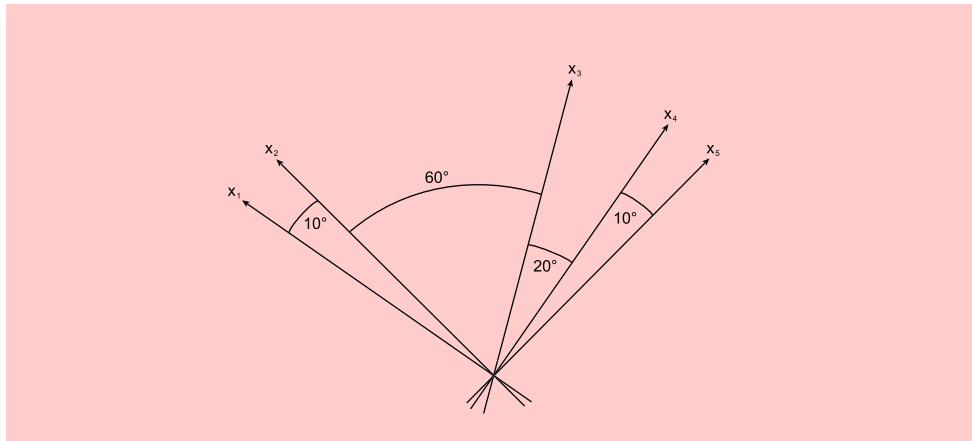


Abbildung 7.20: Graphische Darstellung des 5-Variablen-Beispiels

Betrachtet man nun die jetzt gebildeten Winkel zwischen dem 1. Faktor und den Ausgangsvektoren, dann hat man auch die gesuchten Faktorladungen gefunden.

Extraktion des 1. Faktors

Beispielsweise beträgt der Winkel zwischen 1. Faktor und 1. Variablen (Anteil ungesättigter Fettsäuren) $55°12'$. Dies entspricht einer Faktorladung von $0,5707$. Der Leser möge die übrigen Winkel selbst ausmessen.

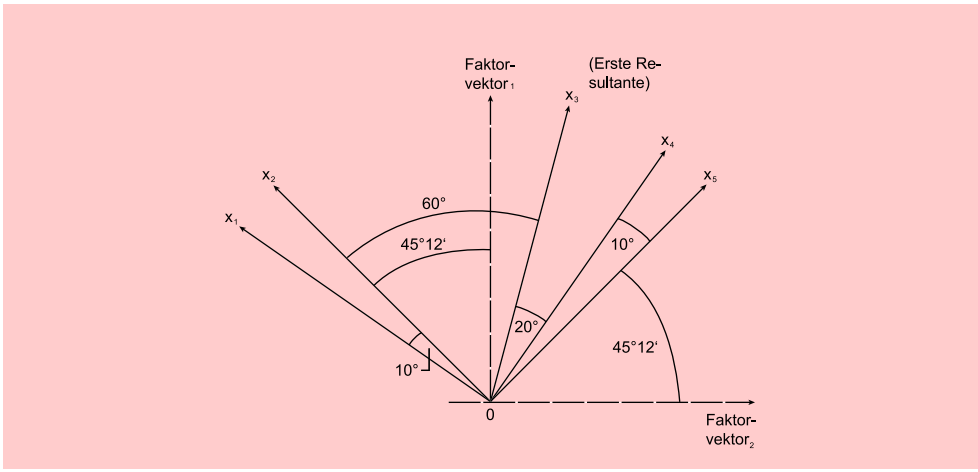


Abbildung 7.21: Graphische Darstellung des Schwerpunktes

Schlägt er die Werte für den Cosinus der jeweiligen Winkel in einer Cosinus-Tabelle nach, so wird er feststellen, dass sich die in Abbildung 7.22 gezeigten übrigen Faktorladungen ergeben.

	Faktor
x_1	$0,5707$
x_2	$0,7046$
x_3	$0,9668$
x_4	$0,8211$
x_5	$0,7096$

Abbildung 7.22: Einfaktorielle Ladungsmatrix

Extraktion des 2. Faktors

Ein zweiter Faktor, der ja vom 1. Faktor unabhängig sein soll, ergibt sich durch die Errichtung eines Vektors in O, der rechtwinklig zum 1. Faktor steht. Damit ergeben sich die in Abbildung 7.23 dargestellten Faktorladungen (der Leser möge die Werte selbst überprüfen).

	Faktor 1	Faktor 2
x_1	$0,5707$	$-0,8211$
x_2	$0,7046$	$-0,7096$
x_3	$0,9668$	$0,2554$
x_4	$0,8211$	$0,5707$
x_5	$0,7096$	$0,7046$

Abbildung 7.23: Zweifaktorielle Ladungsmatrix

Die negativen Faktorladungen zeigen an, dass der jeweilige Faktor negativ mit der entsprechenden Variablen verknüpft ist et vice versa.

Erklärung der Gesamtvarianz

In einem solchen Fall, wenn die ermittelten (extrahierten) Faktoren die Unterschiede in den Beobachtungsdaten restlos erklären, muss die Summe der Ladungsquadrate für jede Variable gleich 1 sein. Warum?

1. Durch die Standardisierung der Ausgangsvariablen haben wir einen Mittelwert von 0 und eine Standardabweichung von 1 erzeugt. Da die Varianz das Quadrat der Standardabweichung ist, ist auch die Varianz gleich 1:

$$s_j^2 = 1 \tag{7.13}$$

2. Die Varianz einer jeden Variablen j erscheint in der Korrelationsmatrix als Selbstkorrelation.

   Man kann diese Überlegung an der graphischen Darstellung in Abbildung 7.15 deutlich machen. Wir hatten gesagt, dass die Länge der Strecke $\overline{AD}$ den Korrelationskoeffizienten beschreibt, wenn $\overline{AC}$ standardisiert, also gleich 1 ist.

   Im Falle der Selbstkorrelation fallen $\overline{AC}$ und $\overline{AB}$ zusammen. Die Strecke $\overline{AB}$ bzw. $\overline{AC}$ mit der normierten Länge von 1 ergibt den (Selbst-) Korrelationskoeffizienten. Die Länge des Vektors $\overline{AB}$ bzw. $\overline{AC}$ gibt aber definitionsgemäß die Standardabweichung wieder. Wegen der Standardisierung ist diese jedoch mit dem Wert 1 gleich der Varianz, sodass tatsächlich gilt:

$$s_j^2 = 1 = r_{j_j} \tag{7.14}$$

Komplette Reproduktion

3. Es lässt sich zeigen, dass auch die Summe der Ladungsquadrate der Faktoren gleich 1 ist, wenn eine komplette Reproduktion der Ausgangsvariablen durch die Faktoren erfolgt.

   Schauen wir uns dazu ein Beispiel an, bei dem zwei Variablen durch zwei Faktoren reproduziert werden (Abbildung 7.24).

   Die Faktorladungen werden durch den Cosinus der Winkel zwischen Ausgangsvektoren und Faktoren beschrieben. Das bedeutet für Variable 1 z. B.:

   - Ladung des 1. Faktors: cos Winkel $COA = \overline{OC}/\overline{OA}$
   - Ladung des 2. Faktors: cos Winkel $DOA = \overline{OD}/\overline{OA}$

   Wenn obige Behauptung stimmt, müsste gelten:

$$(\frac{\overline{OC}}{\overline{OA}})^2 + (\frac{\overline{OD}}{\overline{OA}})^2 = 1 \tag{7.15}$$

   Überprüfung:

$$\frac{\overline{OC}^2}{\overline{OA}^2} + \frac{\overline{OD}^2}{\overline{OA}^2} = \frac{\overline{OC}^2 + \overline{OD}^2}{\overline{OA}^2} \tag{7.16}$$

   In Abbildung 7.24 in Verbindung mit dem Satz des Pythagoras gilt:

$$\overline{OA}^2 = \overline{OC}^2 + \overline{AC}^2 \tag{7.17}$$

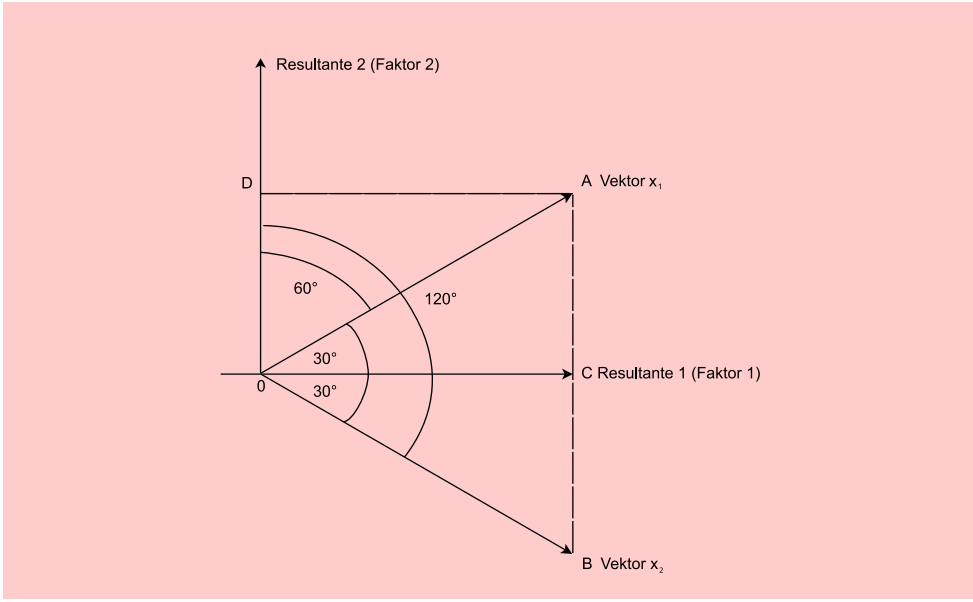


Abbildung 7.24: Zwei Variablen-Zwei Faktor-Lösung

Da nach Abbildung 7.24 $\overline{AC} = \overline{OD}$, gilt auch:

$$\overline{OA}^2 = \overline{OC}^2 + \overline{OD}^2 \tag{7.18}$$

7.18 eingesetzt in 7.16 ergibt dann:

$$\frac{\overline{OC}^2 + \overline{OD}^2}{\overline{OC}^2 + \overline{OD}^2} = 1 \tag{7.19}$$

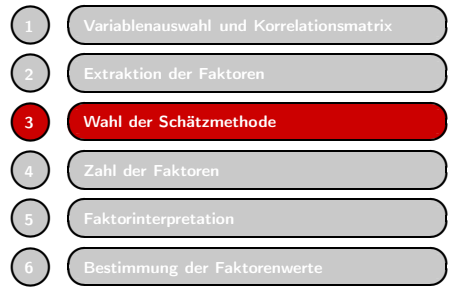
4. Als Fazit lässt sich somit folgende wichtige Beziehung ableiten:

$$s_j^2 = r_{j_j} = a_{j_1}^2 + a_{j_2}^2 + ... + a_{j_Q}^2 = 1, \tag{7.20}$$

wobei a_{j1} bis a_{jQ} die Ladungen der Faktoren 1 bis Q auf die Variable j angeben. Das bedeutet nichts anderes, als dass durch Quadrierung der Faktorladungen in Bezug auf eine Variable und deren anschließender Summation der durch die Faktoren wiedergegebene *Varianzerklärungsanteil der betrachteten Variablen* dargestellt wird: $\sum_q a_{jq}^2$ ist nichts anderes als das *Bestimmtheitsmaß* der Regressionsanalyse (vgl. Kapitel 1 in diesem Buch). Im Falle der Extraktion aller möglichen Faktoren ist der Wert des Bestimmtheitsmaßes gleich 1.

Varianzerklärungsanteil

7.2.3 Wahl der Schätzmethode



In einem konkreten Anwendungsfall, bei dem vor dem Hintergrund des Ziels der Faktorenanalyse die Zahl der Faktoren kleiner als die Zahl der Merkmale ist, kann es sein, dass die Summe der Ladungsquadrate (erklärte Varianz) kleiner als 1 ist. Dies ist dann der Fall, wenn aufgrund theoretischer Vorüberlegungen klar ist, dass nicht die gesamte Varianz durch die Faktoren bedingt ist. Das ist das sog. Kommunalitätenproblem, das erheblichen Einfluss auf die zu wählende Schätzmethode hat.

Restvarianz

Beispielsweise könnten die auf den Wert von 1 normierten Varianzen der Variablen „Kaloriengehalt“ und „Anteil ungesättigter Fettsäuren“ nur zu 70 % auf den Faktor „Gesundheit“ zurückzuführen sein. 30 % der Varianz sind nicht durch den gemeinsamen Faktor bedingt, sondern durch andere Faktoren oder durch Messfehler (Restvarianz). Abbildung 7.25 zeigt die Zusammenhänge noch einmal graphisch.

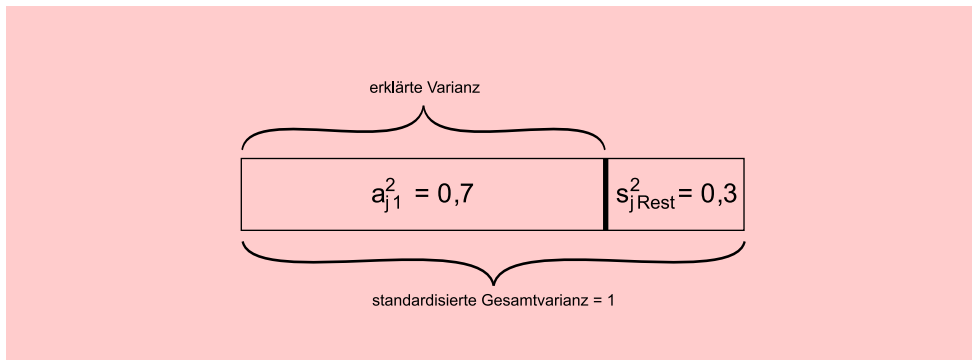


Abbildung 7.25: Die Komponenten der Gesamtvarianz bei der 1-Faktorlösung

Werden statt eines Faktors zwei Faktoren extrahiert, so lässt sich naturgemäß mehr Gesamtvarianz durch die gemeinsamen Faktoren erklären, z. B. 80 % wie in Abbildung 7.26. Den Teil der Gesamtvarianz einer Variablen, der durch die gemeinsamen Faktoren erklärt werden soll, bezeichnet man als *Kommunalität* h_j^2.

Kommunalität

Da i. d. R. die gemeinsamen Faktoren nicht die Gesamtvarianz erklären, sind die Kommunalitäten meist kleiner als Eins.

Messfehler und spezifische Varianz

Das heißt aber nichts anderes, als dass für die Faktorenanalyse das Fundamentaltheorem in Gleichung (7.10) durch eine nicht erklärte Komponente zu ergänzen ist. Wählt man für diesen Restterm, der potenzielle Messfehler und die spezifische Varianz beschreibt, das Symbol **U**, dann ergibt sich für (7.10)

$$\mathbf{R} = \mathbf{A} \cdot \mathbf{A}' + \mathbf{U} \tag{7.21}$$

Die Korrelationsmatrix **R** in (7.21) spiegelt ebenfalls in identischer Weise die aus den empirischen Daten errechneten Korrelationen wider, wobei im Gegensatz zu (7.10) hier eine explizite Unterscheidung zwischen *gemeinsamen Faktoren* (die sich in der Matrix **A** niederschlagen) und *spezifischen Faktoren* (die durch die Matrix **U** repräsentiert werden) vorgenommen wurde. Dabei umfassen die spezifischen Faktoren die

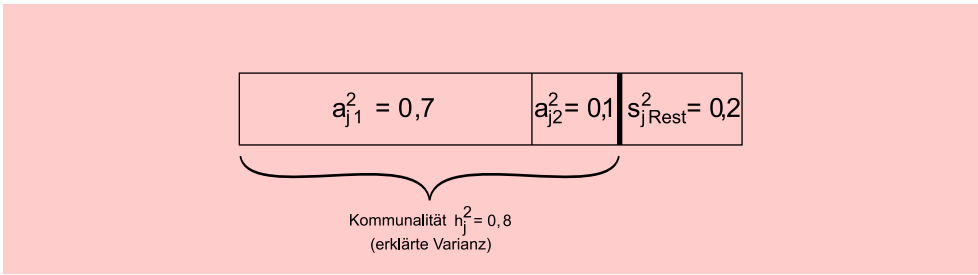


Abbildung 7.26: Die Komponenten der Gesamtvarianz bei einer 2-Faktorlösung

spezifische Varianz einer Variablen sowie die jeweiligen Messfehler. Spezifische Faktoren werden häufig auch als *Einzelrestfaktoren* bezeichnet.

Spezifische Faktoren

Ein wichtiges Problem bei der Faktorenanalyse besteht nun darin, die Kommunalitäten zu schätzen, deren Werte der Anwender ja nicht kennt – er hat nur die Korrelationsmatrix und sucht erst die Faktorladungen. Hierbei handelt es sich um ein subjektives Vorab-Urteil des Forschers, mit dem er einer Vermutung Ausdruck gibt. Setzt er die Kommunalität beispielsweise auf $0,8$, so legt er damit fest, dass *nach seiner Meinung* 80 % der Ausgangsvarianz durch gemeinsame Faktoren erklärbar sind. Um den Schätzcharakter deutlich zu machen, werden die Kommunalitäten häufig als Klammerwerte in die Hauptdiagonale der Korrelationsmatrix eingesetzt. Die so modifizierte Korrelationsmatrix fungiert dann als Ausgangsbasis für die oben beschriebene Faktorenextraktion.

Kommunalität

Hierbei lässt sich ein Zusammenhang zwischen der Anzahl verwendeter Variablen und der Bedeutung einer nahezu korrekten Einschätzung der Kommunalitäten aufstellen: Je größer die Zahl an Variablen ist, desto unwichtiger sind exakt geschätzte Kommunalitäten. Schließlich nimmt der prozentuale Anteil der diagonalen Elemente einer Matrix bei einer steigenden Anzahl an untersuchten Variablen ab. In einer $2x2$-Matrix bilden die diagonalen Elemente noch 50 % aller Elemente, bei einer $10x10$-Matrix sind dies nur noch 10 % (10 diagonale aus insgesamt 100 Elementen), bei einer $100x100$-Matrix gerade einmal noch 1 %. Eine fehlerhafte Eintragung in einem von 100 Elementen für eine Variable (im Falle einer $100x100$-Matrix) hat folglich eine deutlich geringer negative Auswirkung als im Falle einer $2x2$-Matrix.[14]

Zahl der Variablen und Kommunalitäten

Bei der Schätzung der Kommunalitäten ist der Anwender des Verfahrens nicht völlig frei. Vielmehr ergeben sich theoretische Ober- und Untergrenzen für die jeweiligen Werte, die aber hier im Einzelnen nicht dargestellt werden sollen.[15] Innerhalb dieser Grenzen existiert jedoch keine eindeutige Lösung.

Die Schätzung der Kommunalitäten ist eng verbunden mit der Extraktion von Faktoren und der Schätzung der Modellparameter, insb. der Faktorladungen. Hierzu existiert in der Literatur eine Reihe von Schätzverfahren, die aber jeweils zu unterschiedlichen Ergebnissen gelangen können. In SPSS sind insgesamt sieben Schätzverfahren implementiert, die in Abbildung 7.27 kurz charakterisiert sind.

Bei den Schätzverfahren ist zu beachten, dass das Verfahren der „Hauptkomponenten“ und die restlichen Extraktionsverfahren von theoretisch verschiedenen Grundüberlegungen ausgehen. Diese schlagen sich in einer unterschiedlichen sachlogischen Vorstellung des Anwenders im Hinblick auf die Schätzung der Kommunalitäten und

[14] Vgl. Loehlin (2004), S. 160 f.
[15] Vgl. Überla (1977), S. 155 ff.

die Extraktion von Faktoren nieder. Aus *theoretischer Sicht* können die Extraktionsverfahren deshalb in zwei Gruppen unterteilt werden, die wir – wie auch in der Literatur üblich – als „Hauptkomponentenanalyse“ und als „Faktorenanalyse“ bezeichnen:

(A) Faktorenextraktion nach der Hauptkomponentenanalyse

Das Extraktionsverfahren „Hauptkomponenten“, auch *Hauptkomponentenanalyse* genannt, geht von der Überlegung aus, dass die *gesamte Varianz* einer Ausgangsvariable durch die zu extrahierenden Faktoren (Hauptkomponenten) erklärt werden kann. Es wird also unterstellt, dass *keine* Einzelrestvarianzen existieren. Entsprechend wird auch *keine* Unterscheidung zwischen Kommunalitäten und Einzelrestvarianz vorgenommen. Damit ist folglich auch *keine* Schätzung der Kommunalitäten erforderlich, und im Ausgangspunkt der Analyse werden die Kommunalitäten deshalb auf den Wert 1 gesetzt. Eine Hauptkomponente stellt eine Linearkombination aus den betrachteten Ausgangsvariablen dar, wobei unterstellt wird, dass die Hauptkomponenten unkorreliert sind. Die Hauptkomponenten werden mathematisch durch Eigenwertzerlegung der Datenmatrix gefunden, und die Kommunalitäten ergeben sich dabei als Rest. Die Kommunalität einer Variablen in Höhe von 1 (= Varianz einer standardisierten Variablen) wird durch die Hauptkomponentenanalyse immer dann vollständig reproduziert, wenn ebenso viele Faktoren wie Variablen extrahiert werden. Werden weniger Faktoren als Variable extrahiert, so ergeben sich auch bei der Hauptkomponentenanalyse im Ergebnis Kommunalitätenwerte von kleiner 1. Dabei ist der „nicht erklärte“ Varianzanteil (1− Kommunalität) jedoch nicht als Einzelrestvarianz, sondern als durch die Faktoren nicht reproduzierter Varianzanteil und damit als (bewusst in Kauf genommener) Informationsverlust zu verstehen. Meist verfolgen Haupt-komponentenanalysen aber das Ziel, die *Varianz der Variablen* möglichst vollständig durch die Hauptkomponenten zu erklären. Auf diese Weise soll eine möglichst umfassende *Reproduktion* der Datenstruktur durch möglichst wenige, unkorrelierte Hauptkomponenten (Faktoren) erreicht werden.

(B) Extraktionsverfahren im Sinne der Faktorenanalyse

Im Gegensatz zur Hauptkomponentenanalyse geht die *Faktorenanalyse* – wie bereits oben erklärt – davon aus, dass die *Varianz* einer Variablen immer aus *zwei Teilen* besteht: der Kommunalität und der Einzelrestvarianz. Als „Startwert“ bei der Kommunalitätenschätzung sind deshalb immer Werte kleiner 1 vorzugeben. Zur Bestimmung dieser Startwerte liefern die übrigen in Abbildung 7.27 aufgeführten Schätzverfahren zwar unterschiedliche Ansatzpunkte, verfolgen aber als gemeinsames Ziel, die *Zusammenhänge* zwischen den Ausgangsvariablen, gemessen durch die *Korrelation*, möglichst vollständig durch die Faktoren zu erklären. Hingewiesen sei an dieser Stelle auch darauf, dass die Faktorenanalyse die sog. *Ausgangslösung*, bei der ebenso viele Faktoren wie Ausgangsvariablen extrahiert werden, immer mit Hilfe der *Hauptkomponentenanalyse* bestimmt. Auf Basis der Ausgangslösung wird dann eine Entscheidung über die *Anzahl* der zu extrahierenden Faktoren gefällt.

Die Unterschiede in den Grundüberlegungen von Hauptkomponenten- und Faktorenanalysen sind so elementar, dass in der Literatur häufig nur die zweite Gruppe von Schätzverfahren der Faktorenanalyse zugerechnet und die Hauptkomponentenanalyse als eigenständige Analysemethode (neben der Faktorenanalyse) behandelt wird. Die Unterschiede in den Grundideen beider Verfahrensgruppen müssen auch bei der Interpretation der Faktoren bzw. Hauptkomponenten Berücksichtigung finden:

Hauptkomponentenanalyse	Grundidee der Extraktionsmethodik
Hauptkomponenten (Principal components)	Es wird unterstellt, dass die Varianz der Ausgangsvariablen durch unkorrelierte Hauptkomponenten (Faktoren) vollständig reproduziert werden kann. Es existieren keine Einzelrestvarianzen, weshalb die Kommunalitäten der Ausgangsvariablen im Startpunkt der Schätzung auf 1 gesetzt werden.
Schätzverfahren der Faktorenanalyse	**Grundidee der Extraktionsmethodik**
Hauptachsen-Faktorenanalyse (Principal axis factoring)	Faktoren werden aus der beobachteten Korrelationsmatrix extrahiert. Als Anfangsschätzung der Kommunalitäten werden die quadrierten multiplen Korrelationskoeffizienten verwendet. Anschließend werden die Kommunalitäten durch die Faktorladungen so lange iterativ geschätzt, bis diese das Konvergenzkriterium der Extraktion erfüllen.
Maximum Likelihood	Es wird iterativ die Wahrscheinlichkeit (Likelihood) maximiert, dass mit Hilfe der Modellparameter der Faktorenanalyse (insb. Faktorladungen) die empirische Korrelationsmatrix erzeugt werden kann. Die Korrelationen werden dabei gewichtet, wobei Variablen mit hohen Einzelrestvarianzen ein geringeres Gewicht erhalten. Dabei wird eine Normalverteilung der Daten vorausgesetzt. Die Anpassungsgüte kann durch einen Chi-Quadrat-Test überprüft werden.
Ungewichtete kleinste Quadrate (Unweighted least squares)	Es wird iterativ die Summe der quadrierten Differenzen zwischen der beobachteten und der durch die Modellparameter reproduzierten Korrelationsmatrix minimiert. Dabei finden die Kommunalitäten (Hauptdiagonale der Korrelationsmatrix) keine Berücksichtigung.
Verallgemeinerte kleinste Quadrate (Generalized least squares)	Es wird iterativ die Summe der quadierten Differenzen zwischen der beobachteten und der durch die Modellparameter reproduzierten Korrelationsmatrix minimiert. Die Korrelationen werden dabei gewichtet, wobei Variable mit hohen Einzelrestvarianzen ein geringeres Gewicht erhalten. Die Anpassungsgüte kann durch einen Chi-Quadrat-Test überprüft werden.
Alpha-Faktorisierung (Alpha factoring)	Es wird unterstellt, dass die Variablen einer Analyse eine Stichprobe aus der Grundgesamtheit aller möglichen existierenden Variablen darstellen. Die Faktoren werden iterativ so extrahiert, dass sie den Wert von Cronbachs-Alpha der Faktoren maximieren. Dadurch soll die durchschnittliche Korrelation zwischen den Variablen verbessert werden.
Image-Faktorisierung (Image factoring)	Anstelle der Kommunalitäten wird das Image einer Variablen betrachtet, das nach Guttman den Teil der Varianz einer Variablen darstellt, der regressionsanalytisch durch alle anderen Variablen (und nicht durch hypothetische Faktoren) erzeugt werden kann. Durch diese Vorgehensweise wird das Kommunalitätenproblem umgangen.

Abbildung 7.27: Extraktionsverfahren der SPSS-Prozedur FACTOR

Bei der *Hauptkomponentenanalyse* wird letztendlich *keine* kausale Interpretation der „Faktoren“ vorgenommen und die Frage bei der Interpretation der Hauptkomponenten lautet:

> „Wie lassen sich die auf eine Hauptkomponente (Faktor) hoch ladenden Variablen durch einen *Sammelbegriff* zusammenfassen?“

Bei der *Faktorenanalyse* wird eine kausale Interpretation der Faktoren vorgenommen und die Frage bei der Interpretation der Faktoren lautet:

> „Wie lässt sich die *Ursache* bezeichnen, durch die die Korrelation der Variablen, die auf einen Faktor hoch laden, erzeugt wird?“

Im Gegensatz zu den obigen theoretischen Grundideen besteht jedoch in der „Rechentechnik“ von Hauptkomponentenanalyse und den übrigen Extraktionsverfahren kein wesentlicher Unterschied. Das ist auch der Grund, warum beide Verfahrensgruppen bei SPSS unter der Prozedur „FACTOR“ implementiert sind. Die Entscheidung darüber, ob bei der Faktorenanalyse eine Kommunalitätenschätzung unterbleibt (= Hauptkomponentenanalyse) oder mit einem der anderen in Abbildung 7.27 aufgeführten Verfahren vorgenommen wird, hat deshalb letztendlich auch *allein* aufgrund *sachlogischer* Überlegungen zu erfolgen.

Im Weiteren wird unterstellt, dass für unser Beispiel die Frage der „hypothetischen Erklärungsgrößen“ beim Margarinekauf von Interesse ist, weshalb nachfolgend die Vorgehensweise bei Verwendung einer **Hauptachsenanalyse** zur Kommunalitätenschätzung erläutert wird.

Kehren wir zu unserem Ausgangsbeispiel in Abbildung 7.4 und Abbildung 7.14 zurück, so zeigt Abbildung 7.28 die Anfangswerte der Kommunalitäten, die von SPSS im Rahmen der *Hauptachsenanalyse* (bei iterativer Kommunalitätenschätzung) als *Startwerte* vorgegeben werden.

Kommunalitäten

	Anfänglich
Anteil ungesättigter Fettsäuren	,93103
Kaloriengehalt	,54117
Vitamingehalt	,92857
Haltbarkeit	,97381
Preis	,97325

Extraktionsmethode: Hauptachsen-Faktorenanalyse.

Abbildung 7.28: Startwerte der Kommunalitäten im 6-Produkte-Beispiel

Startwerte für Kommunalitäten

SPSS verwendet als Startwerte für die iterative Bestimmung der Kommunalitäten das multiple Bestimmtheitsmaß, das den gemeinsamen Varianzanteil einer Variablen mit allen übrigen Variablen angibt. Setzt man diese Werte in die Korrelationsmatrix der Abbildung 7.10 anstelle der Einsen in die Hauptdiagonale ein und führt auf dieser Basis eine Faktorextraktion mit Hilfe der Hauptachsenanalyse durch (auf die Darstellung der einzelnen Iterationsschritte sei hier verzichtet), so ergibt sich bei (zunächst willkürlicher) Vorgabe von zwei zu extrahierenden Faktoren die in Abbildung 7.29 dargestellte *Faktorladungsmatrix*.

Reproduzierte Korrelationen

Multipliziert man die Faktorladungsmatrix mit ihrer Transponierten, so ergibt sich (gemäß dem Fundamentaltheorem der Faktorenanalyse in Formel (7.10)) die in Abb. 7.30 dargestellte (reproduzierte) Korrelationsmatrix. Abbildung 7.30 enthält im oberen Teil mit der Überschrift „Reproduzierte Korrelation“ in der Hauptdiagonalen

Faktorenmatrix[a]

	Faktor 1	Faktor 2
Anteil ungesättigter Fettsäuren	,94331	-,28039
Kaloriengehalt	,70669	-,16156
Vitamingehalt	,92825	-,30210
Haltbarkeit	,38926	,91599
Preis	,32320	,93608

Extraktionsmethode: Hauptachsenfaktorenanalyse.

a. 2 Faktoren extrahiert. Es werden 7 Iterationen benötigt.

Abbildung 7.29: Faktorladungen im 6-Produkte-Beispiel

die Endwerte der iterativ geschätzten Kommunalitäten bei zwei Faktoren. Die nicht-diagonal-Elemente geben die durch die Faktorenstruktur reproduzierten Korrelationen wieder. In der unteren Abbildung mit der Überschrift „Residuum“ werden die Differenzwerte zwischen den ursprünglichen (Abbildung 7.10) und den reproduzierten Korrelationen ausgewiesen. Dabei wird deutlich, dass in unserem Beispiel keiner der Differenzwerte größer als $0,05$ ist, sodass die auf der Basis der Faktorladungen ermittelte Korrelationsmatrix der ursprünglichen Korrelationsmatrix sehr ähnlich ist, sie also „sehr gut“ reproduziert.

Reproduzierte Korrelationen

		Anteil ungesättigter Fettsäuren	Kaloriengehalt	Vitamingehalt	Haltbarkeit	Preis
Reproduzierte Korrelation	Anteil ungesättigter Fettsäuren	,968[a]	,712	,960	,110	,042
	Kaloriengehalt	,712	,526[a]	,705	,127	,077
	Vitamingehalt	,960	,705	,953[a]	,085	,017
	Haltbarkeit	,110	,127	,085	,991[a]	,983
	Preis	,042	,077	,017	,983	,981[a]
Residuum[b]	Anteil ungesättigter Fettsäuren		,000	,001	-,001	,001
	Kaloriengehalt	,000		-,001	,011	-,011
	Vitamingehalt	,001	-,001		-,006	,006
	Haltbarkeit	-,001	,011	-,006		9,539E-5
	Preis	,001	-,011	,006	9,539E-5	

Extraktionsmethode: Hauptachsen-Faktorenanalyse.

a. Reproduzierte Kommunalitäten

b. Residuen werden zwischen beobachteten und reproduzierten Korrelationen berechnet. Es liegen 0 (0,0%) nicht redundante Residuen mit absoluten Werten größer 0,05 vor.

Abbildung 7.30: Die reproduzierte Korrelationsmatrix im 6-Produkte-Beispiel

Güte der Reproduktion

Das aber bedeutet nichts anderes, als dass sich die beiden gefundenen Faktoren ohne großen Informationsverlust zur Beschreibung der fünf Ausgangsvariablen eignen.

Wegen der unterstellten spezifischen Varianz und des damit verbundenen Problems der Kommunalitätenschätzung ist klar, dass durch die Rechenregel $\mathbf{R} = \mathbf{A} \cdot \mathbf{A}'$ die Ausgangs-Korrelationsmatrix $\mathbf{R}$ nicht identisch reproduziert werden kann. Dies gilt auch für die Kommunalitäten. Aus diesem Grunde kennzeichnen wir die reproduzierte Korrelationsmatrix als $\widehat{\mathbf{R}}$.

7.2.4 Zahl der Faktoren

Im vorangegangenen Abschnitt hatten wir uns willkürlich für zwei Faktoren entschieden. Generell ist zu bemerken, dass zur Bestimmung der Faktorenzahl keine eindeutigen Vorschriften existieren, sodass hier der *subjektive Eingriff* des Anwenders erforderlich ist. Allerdings lassen sich auch statistische Kriterien heranziehen, von denen insbesondere die folgenden als bedeutsam anzusehen sind:

Kaiser-(Eigenwert-) Kriterium

- *Kaiser-Kriterium.* Danach ist die Zahl der zu extrahierenden Faktoren gleich der Zahl der Faktoren mit Eigenwerten größer eins. Die Eigenwerte (Eigenvalues) werden berechnet als Summe der quadrierten Faktorladungen *eines* Faktors über alle Variablen. Sie sind ein Maßstab für die durch den jeweiligen Faktor erklärte Varianz der Beobachtungswerte. Der Begriff Eigenwert ist deutlich vom „erklärten Varianzanteil“ zu trennen. Letzterer beschreibt den Varianzerklärungsanteil, der durch die Summe der quadrierten Ladungen *aller Faktoren* im Hinblick auf *eine Variable* erreicht wird (theoretischer oberer Grenzwert Kommunalität $\sum_q a_{jq}^2$), während der Eigenwert den Varianzerklärungsbeitrag *eines Faktors* im Hinblick auf die Varianz *aller Variablen* beschreibt ($\sum_j a_{jq}^2$).

  Abbildung 7.31 zeigt nochmals die Faktorladungsmatrix aus Abbildung 7.29 auf, wobei in Klammern jeweils die *quadrierten Faktorladungen* stehen. Ad-

Faktorenmatrix

	Faktor 1	Faktor 2
Anteil ungesättigter Fettsäuren	,94331 (,8898)	-,28039 (,0786)
Kaloriengehalt	,70669 (,4994)	-,16156 (,0261)
Vitamingehalt	,92825 (,8616)	-,30210 (,0913)
Haltbarkeit	,38926 (,1515)	,91599 (,8390)
Preis	,32320 (,1045)	,93608 (,8762)
Eigenwerte	2,5068	1,9112

Abbildung 7.31: Bestimmung der Eigenwerte

diert man die Ladungsquadrate je Zeile, so ergeben sich die *Kommunalitäten* der Variablen (vgl. Abbildung 7.31). Von der Eigenschaft „Anteil ungesättigter Fettsäure“ werden folglich $(0,8898 + 0,0786 = 0,9684)$ 96,84 % der Varianz durch die zwei extrahierten Faktoren erklärt. Die spaltenweise Summation erbringt die Eigenwerte der Faktoren, die in Abbildung 7.31 in der untersten Zeile abgebildet werden. Die Begründung für die Verwendung des Kaiser-Kriteriums liegt darin, dass ein Faktor, dessen Varianzerklärungsanteil über alle Variablen kleiner als eins ist, weniger Varianz erklärt als eine einzelne Variable; denn die Varianz einer *standardisierten* Variablen beträgt ja gerade 1. In unserem Beispiel führt das Kaiser-Kriterium zu der Extraktion von zwei Faktoren, da bei der Extraktion eines dritten Faktors der entsprechende Eigenwert bereits kleiner 0,4 wäre.

Kommunalitäten

Eigenwerte

Begründung für Kaiser-Kriterium

Kommunalitäten

	Extraktion
Anteil ungesättigter Fettsäuren	,9685
Kaloriengehalt	,5255
Vitamingehalt	,9529
Haltbarkeit	,9906
Preis	,9807

Extraktionsmethode: Hauptachsen-Faktorenanalyse.

Abbildung 7.32: Kommunalitäten als erklärter Varianzanteil

- *Scree-Test.* Beim Scree-Test werden die Eigenwerte in einem Koordinatensystem nach abnehmender Wertefolge angeordnet. An der Stelle, an der die Differenz der Eigenwerte zwischen zwei Faktoren am größten ist, entsteht ein Knick. Der erste Punkt links von diesem Knick bestimmt die Anzahl der zu extrahierenden Faktoren: es sollen ja die wichtigsten Faktoren extrahiert werden. Der Hintergrund dieser Vorgehensweise ist darin zu sehen, dass die Faktoren mit den kleinsten Eigenwerten für Erklärungszwecke als unbrauchbar (Scree=Geröll) angesehen und deshalb auch nicht extrahiert werden. Das Verfahren liefert allerdings nicht immer eindeutige Lösungen, da Situationen denkbar sind, in denen sich aufgrund z. T. ähnlicher Differenzen der Eigenwerte kein eindeutiger Knick ermitteln lässt. Abbildung 7.33 zeigt den Scree-Test für das 6-Produkte-Beispiel, wonach hier, wie beim Kaiser-Kriterium, zwei Faktoren zu extrahieren wären.

  Scree-Test

  Obwohl es dem Forscher prinzipiell selbst überlassen bleibt, welches Kriterium er bei der Entscheidung über die Zahl zu extrahierender Faktoren zugrunde legt, kommt in empirischen Untersuchungen häufig das Kaiser-Kriterium zur Anwendung, obwohl dies umstritten ist.

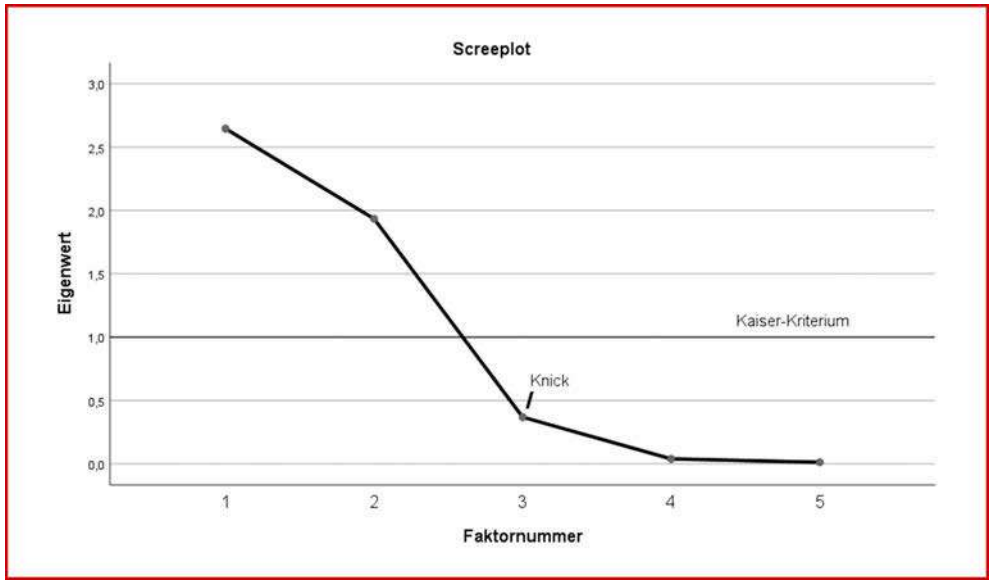
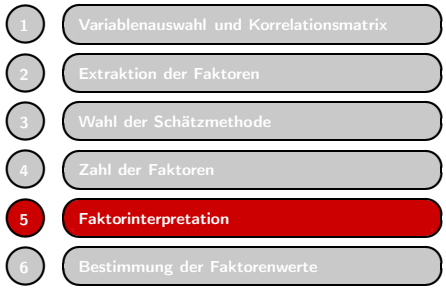


Abbildung 7.33: Scree-Test im 6-Produkte-Beispiel

7.2.5 Faktorinterpretation



Interpretationshilfe

Ist die Zahl der Faktoren bestimmt, so muss anschließend versucht werden, die Faktoren, die zunächst rein abstrakte Größen (Vektoren) darstellen, zu interpretieren. Dazu bedient man sich als Interpretationshilfe der Faktorladungen, die für unser Beispiel *nochmals* in Abbildung 7.34 wiedergegeben sind (vgl. auch Abbildung 7.29):

Es zeigt sich, dass der Faktor 1 besonders stark mit den Größen

- Anteil ungesättigter Fettsäuren
- Kaloriengehalt
- Vitamingehalt

korreliert. Man könnte zu dem Ergebnis kommen, dass letztendlich der Gesundheitsaspekt für das Beurteilungsverhalten der Befragten verantwortlich war und damit die Faktorladungen bestimmt hat. Wir bezeichnen den ersten Faktor deshalb als „Gesundheit“. Für die Variablen x_4 und x_5 sei unterstellt, dass der Wirtschaftlichkeitsaspekt bei der Beurteilung im Vordergrund stand.

Faktor: Gesundheit

Faktor: Wirtschaftlichkeit

Der Faktor wird deshalb als „Wirtschaftlichkeit“ charakterisiert. An dieser Stelle wird besonders deutlich, dass die Interpretation der Faktoren eine hohe Sachkenntnis des Anwenders bezüglich des konkreten Untersuchungsobjektes erfordert.

Faktorenmatrix

	Faktor 1	Faktor 2
Anteil ungesättigter Fettsäuren	**,94331**	-,28039
Kaloriengehalt	**,70669**	-,16156
Vitamingehalt	**,92825**	-,30210
Haltbarkeit	,38926	**,91599**
Preis	,32320	**,93608**

Extraktionsmethode: Hauptachsenfaktorenanalyse.

Abbildung 7.34: Faktorladungen im 6-Produkte-Beispiel

Die Faktorladungsmatrix in Abbildung 7.34 weist eine sogenannte *Einfachstruktur* auf, d. h. die Variablen laden immer nur auf *einem* Faktor hoch und auf allen anderen Faktoren (in diesem 2-Faktorfall jeweils auf dem anderen Faktor) niedrig. Bei größeren Felduntersuchungen ist dies jedoch häufig nicht gegeben und es fällt dann nicht leicht, die jeweiligen Faktoren zu interpretieren. Hier besteht nur die Möglichkeit, das Faktormuster offenzulegen, sodass der jeweils interessierte Verwender der Analyseergebnisse Eigeninterpretationen vornehmen kann. Das bedeutet allerdings auch, dass gerade die Faktorinterpretation subjektive Beurteilungsspielräume offen lässt. Das gilt besonders dann, wenn eine Interpretation wegen der wenig konsistenten Ladungen schwierig ist.

Einfachstruktur

Der Anwender muss entscheiden, ab welcher Ladungshöhe er eine Variable einem Faktor zuordnet. Dazu sind gewisse Regeln (Konventionen) entwickelt worden, wobei in der praktischen Anwendung „hohe“ Ladungen ab $0,5$ angenommen werden. Dabei ist allerdings darauf zu achten, dass eine Variable, wenn sie auf mehreren Faktoren Ladungen $\geq 0,5$ aufweist, bei *jedem* dieser Faktoren zur Interpretation herangezogen werden muss.

Konventionen

Laden mehrere Variable auf mehrere Faktoren gleich hoch, dann ist es häufig unmöglich, unmittelbar eine sinnvolle Faktorinterpretation zu erreichen (Abbildung 7.35).[16]

Möglicherweise hilft bei der Interpretation die Drehung (Rotation) des Koordinatenkreuzes in seinem Ursprung. Dreht man das Koordinatenkreuz in Abbildung 7.35 in seinem Ursprung, so lässt sich beispielsweise die Konstellation aus Abbildung 7.36 erreichen. Jetzt lädt das obere Variablenbündel vor allem auf Faktor 2 und das untere auf Faktor 1. Damit wird die Interpretation erheblich erleichtert.

Rotation

SPSS unterstützt verschiedene Möglichkeiten zur Rotation des Koordinatenkreuzes, wobei grundsätzlich zwei Kategorien unterschieden werden können.

[16] Zur Beurteilung der Faktorladungsstruktur sowie der Faktorinterpretation vergleiche Litfin/Teichmann/Clement (2000), S. 285 f.

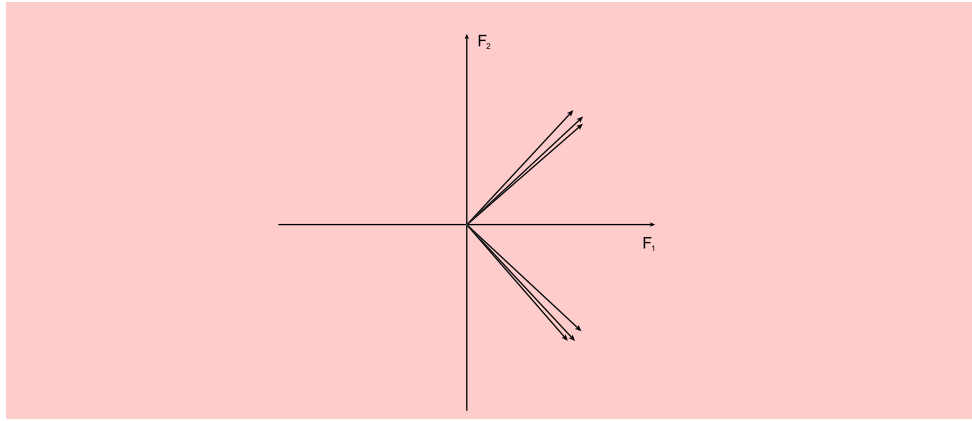


Abbildung 7.35: Unrotierte Faktorladungen

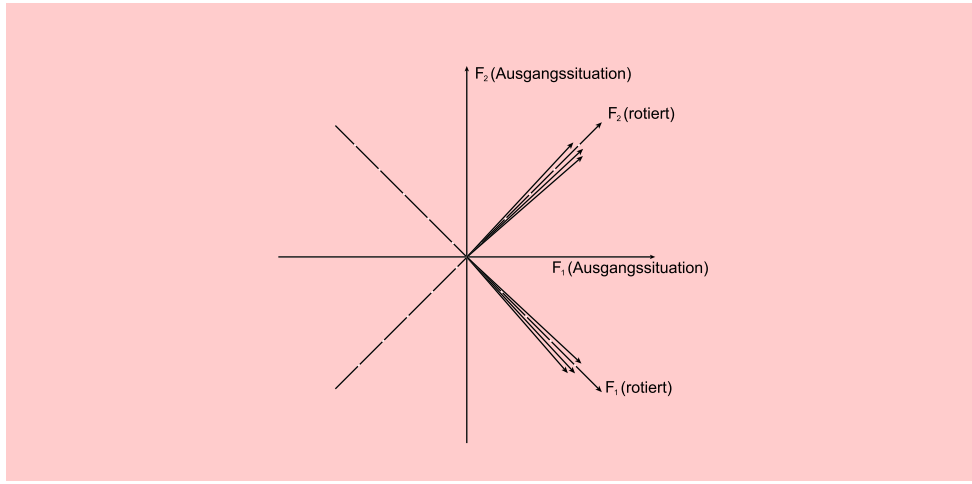


Abbildung 7.36: Rotierte Faktorladungen

Orthogonale Rotation

1. Sofern angenommen werden kann, dass die Faktoren untereinander nicht korrelieren, verbleiben die Faktorachsen während der Drehung in einem rechten Winkel zueinander. Es handelt sich hierbei um Methoden der *orthogonalen (rechtwinkligen) Rotation*.

Oblique Rotation

2. Die Achsen werden in einem schiefen Winkel zueinander rotiert, falls eine Korrelation zwischen den rotierten Achsen bzw. Faktoren angenommen wird. Hierbei spricht man von Methoden der *obliquen (schiefwinkligen) Rotation*.

Abbildung 7.37 zeigt das Ergebnis der rechtwinkligen Varimax-Rotation für unser Beispiel. Hierbei handelt es sich um eine sehr häufig angewendete Methode. Die Ergebnisse zeigen, dass die Faktorladungen auf die entsprechenden Faktoren jeweils noch höher geworden sind.

Rotierte Faktorenmatrix[a]

	Faktor 1	Faktor 2
Anteil ungesättigter Fettsäuren	,98357	,03229
Kaloriengehalt	,72152	,07020
Vitamingehalt	,97615	,00694
Haltbarkeit	,07962	,99208
Preis	,01060	,99025

Extraktionsmethode: Hauptachsenfaktorenanalyse.
Rotationsmethode: Varimax mit Kaiser-Normalisierung.

a. Die Rotation ist in 3 Iterationen konvergiert.

Abbildung 7.37: Rotierte Varimax-Faktorladungsmatrix im 6-Produkte-Beispiel

Um die Rotation nachvollziehen zu können, sollte sich der Leser die Formel für eine orthogonale Transformation um einen Winkel α *nach links* vergegenwärtigen:

$$\mathbf{A}^* = \mathbf{A} \cdot \mathbf{T}, \qquad \text{mit} \qquad \mathbf{T} = \begin{bmatrix} cos\alpha & -sin\alpha \\ sin\alpha & cos\alpha \end{bmatrix}$$

Im obigen Beispiel handelt es sich um eine Rotation um $18,43°$ nach rechts bzw. um $341,57°$ nach links:

$$\underset{\mathbf{A}}{\begin{bmatrix} 0,94331 & -0,28039 \\ 0,70669 & -0,16156 \\ 0,92825 & -0,30210 \\ 0,38926 & 0,91599 \\ 0,32320 & 0,93608 \end{bmatrix}} \cdot \underset{\mathbf{T}}{\begin{bmatrix} cos341,57° & -sin341,57° \\ sin341,57° & cos341,57° \end{bmatrix}} = \underset{\mathbf{A}^*}{\begin{bmatrix} 0,98357 & 0,03229 \\ 0,72152 & 0,07020 \\ 0,97615 & 0,00694 \\ 0,07962 & 0,99208 \\ 0,01060 & 0,99025 \end{bmatrix}}$$

Im SPSS-Output lässt sich der Rotationswinkel anhand der Faktor Transformationsmatrix bestimmen (vgl. Abbildung 7.38). Für das Bogenmaß $0,94868$ gilt für den Cosinus ein entsprechendes Winkelmaß von $\alpha = 18,43°$. Rotationswinkel

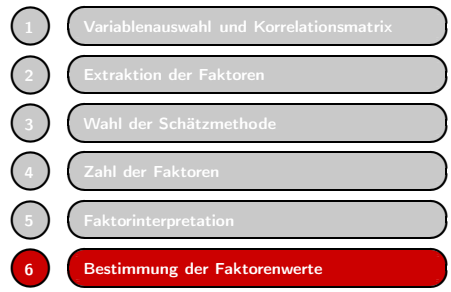
Faktor-Transformationsmatrix

Faktor	1	2
1	,94868	,31622
2	-,31622	,94868

Extraktionsmethode: Hauptachsenfaktorenanalyse.
Rotationsmethode: Varimax mit Kaiser-Normalisierung.

Abbildung 7.38: Faktor Transformationsmatrix

7.2.6 Bestimmung der Faktorenwerte (Factor Scores)



Faktorwerte

Für eine Vielzahl von Fragestellungen ist es von großem Interesse, nicht nur die Variablen auf eine geringere Anzahl von Faktoren zu verdichten, sondern danach zu erfahren, welche Werte die Objekte (Marken) nun hinsichtlich der extrahierten Faktoren annehmen. Man benötigt also nicht nur die Faktoren selbst, sondern auch die Ausprägung der Faktoren bei den Objekten bzw. Personen. Dieses bezeichnet man als das Problem der Bestimmung der *Faktorwerte* (in SPSS auch *factor scores* genannt). Das Ergebnis ist eine *Datenreduktion*.

Wie oben erläutert, ist es das Ziel der Faktorenanalyse, die standardisierte Ausgangsdatenmatrix $\mathbf{Z}$ als Linearkombination von Faktoren darzustellen. Es galt:

$$\mathbf{Z} = \mathbf{P} \cdot \mathbf{A}' \tag{7.5}$$

Wir haben uns bisher mit der Bestimmung von $\mathbf{A}$ (Faktorladungen) beschäftigt. Da $\mathbf{Z}$ gegeben ist, ist die Gleichung (7.5) nach den gesuchten Faktorwerten $\mathbf{P}$ aufzulösen. Bei Auflösung nach $\mathbf{P}$ ergibt sich durch Multiplikation von rechts mit der inversen Matrix $(\mathbf{A}')^{-1}$:

$$\mathbf{Z} \cdot (\mathbf{A}')^{-1} = \mathbf{P} \cdot \mathbf{A}' \cdot (\mathbf{A}')^{-1} \tag{7.22}$$

Probleme bei der Schätzung der Faktorwerte

Da $\mathbf{A}' \cdot (\mathbf{A}')^{-1}$ definitionsgemäß die Einheitsmatrix $\mathbf{E}$ ergibt, folgt:

$$\mathbf{Z} \cdot (\mathbf{A}')^{-1} = \mathbf{P} \cdot \mathbf{E} \tag{7.23}$$

Da $\mathbf{P} \cdot \mathbf{E} = \mathbf{P}$ ist, ergibt sich:

$$\mathbf{P} = \mathbf{Z} \cdot (\mathbf{A}')^{-1} \tag{7.24}$$

Für das in der Regel nicht quadratische Faktormuster $\mathbf{A}$ (es sollen ja gerade weniger Faktoren als Variable gefunden werden!) ist eine Inversion nicht möglich. Deshalb könnte in bestimmten Fällen folgende Vorgehensweise eine Lösung bieten:

Lösungsvorschlag

(7.5) wird von rechts mit $\mathbf{A}$ multipliziert:

$$\mathbf{Z} \cdot \mathbf{A} = \mathbf{P} \cdot \mathbf{A}' \cdot \mathbf{A} \tag{7.25}$$

Matrix $(\mathbf{A}' \cdot \mathbf{A})$ ist definitionsgemäß quadratisch und somit invertierbar:

$$\mathbf{Z} \cdot \mathbf{A} \cdot (\mathbf{A}' \cdot \mathbf{A})^{-1} = \mathbf{P} \cdot (\mathbf{A}' \cdot \mathbf{A}) \cdot (\mathbf{A}' \cdot \mathbf{A})^{-1} \tag{7.26}$$

Da $(\mathbf{A}' \cdot \mathbf{A}) \cdot (\mathbf{A}' \cdot \mathbf{A})^{-1}$ definitionsgemäß eine Einheitsmatrix ergibt, gilt:

$$\mathbf{P} = \mathbf{Z} \cdot \mathbf{A} \cdot (\mathbf{A}' \cdot \mathbf{A})^{-1} \tag{7.27}$$

In bestimmten Fällen können sich bei der Lösung dieser Gleichung aber ebenfalls Schwierigkeiten ergeben. Man benötigt dann Schätzverfahren zur Lösung dieses Problems. Je nach Wahl des Schätzverfahrens kann daher die Lösung variieren. In vielen Fällen wird zur Schätzung der Faktorwerte auf die Regressionsanalyse (vgl. Kapitel 1) zurückgegriffen. Für unser Beispiel ergab sich für den Term $\mathbf{A} \cdot (\mathbf{A}' \cdot \mathbf{A})^{-1}$ die in Abbildung 7.39 aufgeführte Koeffizientenmatrix der Faktorwerte.

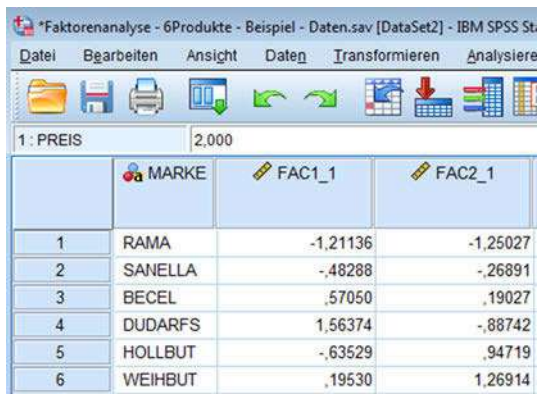
Koeffizientenmatrix der Faktorwerte

	Faktor 1	Faktor 2
Anteil ungesättigter Fettsäuren	,55098	-,04914
Kaloriengehalt	,01489	-,01014
Vitamingehalt	,42220	,00081
Haltbarkeit	,26113	,67344
Preis	-,28131	,33084

Extraktionsmethode: Hauptachsenfaktorenanalyse.
Rotationsmethode: Varimax mit Kaiser-Normalisierung.
Faktorscoremethode: Regression.

Abbildung 7.39: Koeffizientenmatrix der Faktorwerte

Die standardisierte Ausgangsdatenmatrix **Z** (Abbildung 7.41), multipliziert mit den Regressionskoeffizienten, ergibt dann die in Abbildung 7.40 aufgeführten Faktorwerte. Die standardisierte Datenmatrix **Z** errechnet sich gemäß der Formel $z_{kj} = \frac{x_{kj} - \overline{x}_j}{s_j}$ in Kapitel 7.2.1.1 zu der in Abbildung 7.40 dargestellten Matrix.



	MARKE	FAC1_1	FAC2_1
1	RAMA	-1,21136	-1,25027
2	SANELLA	-,48288	-,26891
3	BECEL	,57050	,19027
4	DUDARFS	1,56374	-,88742
5	HOLLBUT	-,63529	,94719
6	WEIHBUT	,19530	1,26914

Abbildung 7.40: Faktorwerte im 6-Produkte-Beispiel

Die Multiplikation der Matrix [6x5] in Abbildung 7.41 mit der Matrix [5x2] in Abbildung 7.39 ergibt die Faktorwerte-Matrix [6x2] in Abbildung 7.40.

*Faktorenanalyse - 6Produkte - Beispiel - Daten.sav [DataSet2] - IBM SPSS Statistics Dateneditor

Datei Bearbeiten Ansicht Daten Transformieren Analysieren Grafik Extras Er

13 :

	MARKE	ZUNGEFET	ZKALORIEN	ZVITAMIN	ZHALTBARK	ZPREIS
1	RAMA	-1,24550	-1,63164	-1,21988	-1,33631	-1,29099
2	SANELLA	-,56614	,94463	-,48795	-,26726	-,32275
3	BECEL	,79259	,42938	,24398	,26726	,16137
4	DUDARFS	1,47196	,94463	1,70783	-,80178	-,80687
5	HOLLBUT	-,56614	-,60113	-,48795	,80178	1,12962
6	WEIHBUT	,11323	-,08588	,24398	1,33631	1,12962

Abbildung 7.41: Standardisierte Ausgangsdatenmatrix **Z**

Positionierung

Die Faktorwerte lassen sich graphisch darstellen und liefern damit eine Visualisierung der beurteilten Margarinemarken im zweidimensionalen Faktorenraum (Abbildung 7.42). Gleichzeitig lassen sich in diese Darstellung, unter Rückgriff auf die (rotierte) Faktorladungsmatrix (Abbildung 7.37), auch die Positionen der Faktoren übertragen. Damit erhält der Anwender gleichzeitig einen optischen Anhaltspunkt dafür, wie stark die Achsen des Koordinatensystems (Faktoren) mit den Variablen in Verbindung stehen.

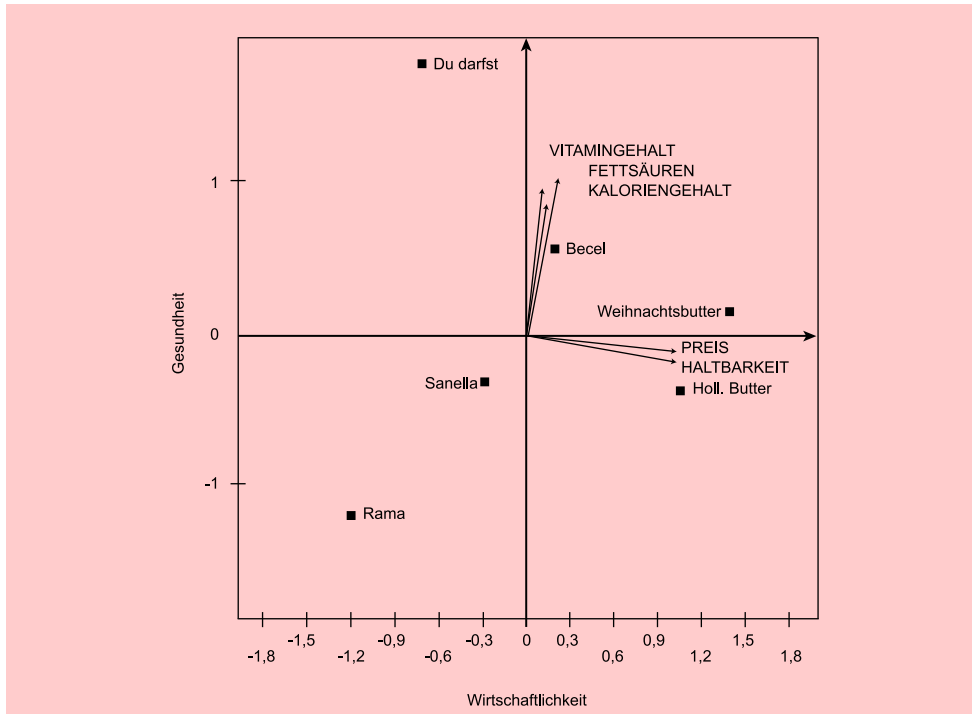


Abbildung 7.42: Faktorwerte-Plot und rotierte Faktorladungen im 6-Produkte-Beispiel

7.2.7 Zusammenfassende Darstellung der Faktorenanalyse

Wie im Einzelnen dargestellt, sind zur Durchführung einer Faktorenanalyse fünf grundlegende *Schritte* notwendig, um die Variablen einer Datenmatrix auf die den Daten zugrundeliegenden hypothetischen Faktoren zurückzuführen (Abbildung 7.43), wobei die Kantenlängen in Relation zueinander stehen: In der Ausgangsdatenmatrix

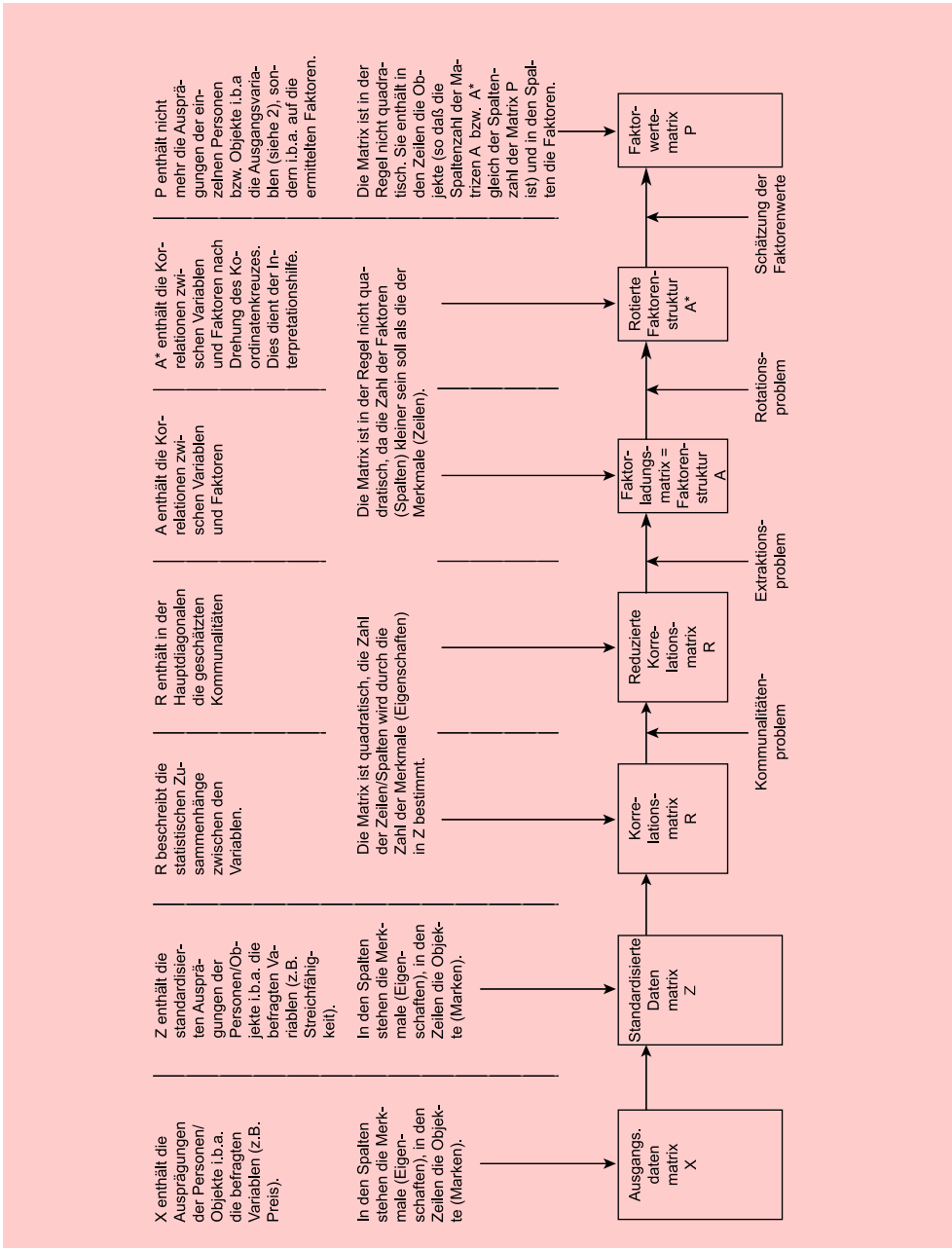


Abbildung 7.43: Die Rechenschritte der Faktorenanalyse

Abb. 7.4 wird analog zum Beispiel davon ausgegangen, dass die Zahl der Variablen (5) kleiner ist als die Zahl der Objekte (6). Die Korrelationsmatrix ist dagegen definitionsgemäß quadratisch. Aus der Darstellung wird noch einmal deutlich, welche Begriffe welchen Rechenoperationen bzw. Rechenergebnissen zuzuordnen sind.

Manipulationsmöglichkeiten

Zusammenfassend lässt sich festhalten: Bei der Ermittlung der Faktorenwerte aus den Ausgangsdaten sind zwei verschiedene Arten von Rechenschritten notwendig:

- Solche, die eindeutig festgelegt sind (die Entwicklung der standardisierten Datenmatrix und der Korrelationsmatrix aus der Datenmatrix),
- solche, wo der Verwender des Verfahrens subjektiv eingreifen kann und muss, wo das Ergebnis also von seinen Entscheidungen abhängt (z. B. die Kommunalitätenschätzung).

Geht man davon aus, dass die erhobenen Daten das für die Korrelationsanalyse notwendige Skalenniveau besitzen, d. h. sind sie mindestens intervallskaliert, dann sind lediglich die *ersten beiden Schritte* von **X** nach **Z** und **Z** nach **R** *manipulationsfrei*. Alle anderen notwendigen Rechenschritte, die in Abbildung 7.43 durch Pfeile gekennzeichnet sind, sind subjektiven Maßnahmen des Untersuchenden zugänglich und erfordern die Eingriffe.

Standardlösungen

In den gängigen Computerprogrammen für die Durchführung einer Faktorenanalyse wird dieses Problem i. d. R. so gelöst, dass dem Anwender des Verfahrens für die einzelnen Entscheidungsprobleme „Standardlösungen“ angeboten werden. Der Anwender muss nur eingreifen, wenn er eine andere Lösung anstrebt, beispielsweise statt des automatisch angewendeten Kaiser-Kriteriums eine bestimmte Anzahl an zu extrahierenden Faktoren vorgeben möchte.

Gerade diese Vorgehensweise ist jedoch immer dann höchst problematisch, wenn dem Anwender die Bedeutung der einzelnen Schritte im Verfahren nicht klar ist und er das ausgedruckte Ergebnis als „die“ Lösung ansieht.

Um diesen Fehler zu vermeiden und die Aussagekraft faktoranalytischer Untersuchungen beurteilen zu können, wird im Folgenden eine Faktoranalyse anhand eines komplexeren konkreten Beispiels vorgestellt. Um die einzelnen Rechenschritte nachprüfen zu können, kann der Datensatz von den Verfassern bezogen werden. Es werden verschiedene Lösungen bei den einzelnen Teilproblemen im Rechengang der Faktoranalyse vorgestellt und kommentiert, um so den möglichen Manipulationsspielraum bei der Verwendung des Verfahrens offenzulegen.

7.3 Fallbeispiel

7.3.1 Problemstellung

In einer empirischen Erhebung wurden elf Emulsionsfette (Butter und Margarine) im Hinblick auf bestimmte Eigenschaften beurteilt. Im Einzelnen handelte es sich um die in Abbildung 7.44 angeführten Marken und Eigenschaften.

Beispielbeschreibung

Es wurden 18 Personen bezüglich ihrer Einschätzung der 11 Marken befragt und anschließend die Einzelurteile auf Markenebene über Mittelwertbildung verdichtet. Die durchschnittlichen Eigenschaftsurteile gehen pro Marke im Folgenden als Datenmatrix in die Faktorenanalyse ein (vgl. die Datenmatrix in Abbildung 7.45).

Es soll auf Basis dieser Befragung geprüft werden, ob die zehn *Eigenschaften* alle *unabhängig voneinander* zur (subjektiven) Beurteilung der Marken notwendig sind

	Marken $M_k (k = 1, ..., 11)$		Eigenschaften $X_j (j = 1, ..., 10)$
1	Sanella	A	Streichfähigkeit
2	Homa	B	Preis
3	SB	C	Haltbarkeit
4	Delicado	D	Anteil ungesättigter Fettsäuren
5	Holl. Markenbutter	E	Back- und Brateignung
6	Weihnachtsbutter	F	Geschmack
7	Du darfst	G	Kaloriengehalt
8	Becel	H	Anteil tierischer Fette
9	Botteram	I	Vitamingehalt
10	Flora	K	Natürlichkeit
11	Rama		

Abbildung 7.44: Variable und Objekte des Beispiels

oder ob bestimmte komplexere Faktoren eine hinreichend genaue Beurteilung geben. In einem zweiten Schritt sollten die Marken entsprechend der Faktorenausprägung positioniert werden.Um mit Hilfe des Programmes SPSS eine Faktorenanalyse durch-

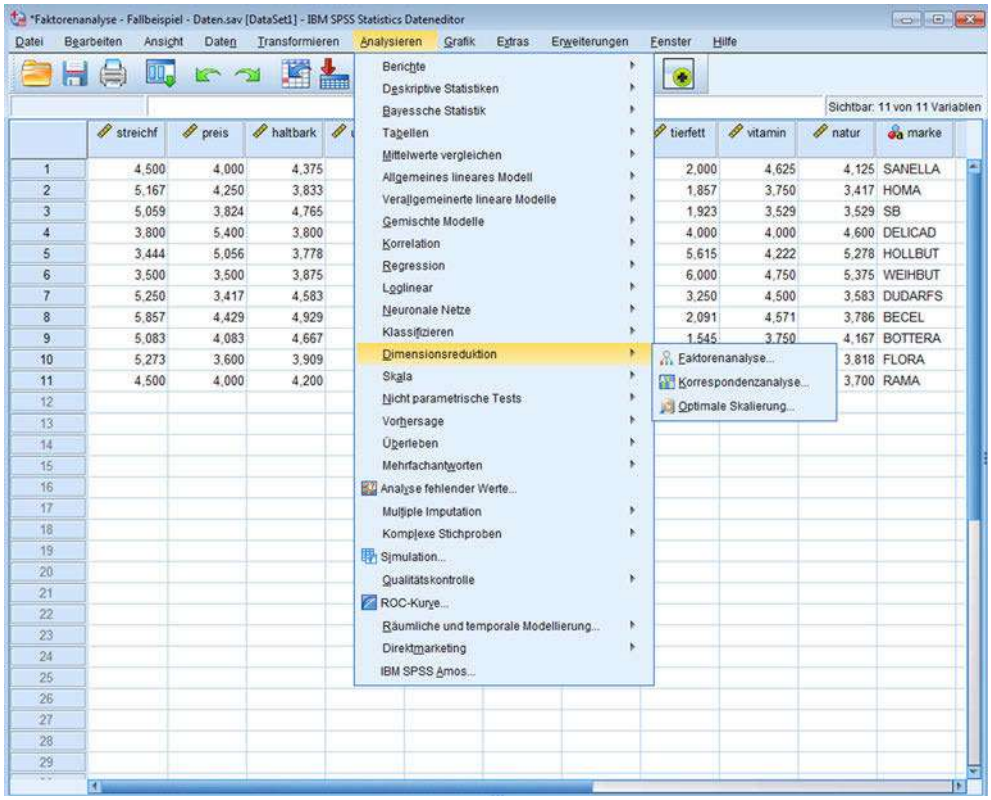


Abbildung 7.45: Dateneditor mit Auswahl des Analyseverfahrens „Faktorenanalyse“

führen zu können, wurde zunächst das Verfahren der Faktorenanalyse aus dem Menüpunkt Dimensionsreduktion ausgewählt (vgl. Abbildung 7.45).

Die untersuchten Variablen aus der Quellvariablenliste wurden danach in das Feld „Variablen“ übertragen (vgl. Abbildung 7.46) und ausgewählte Voreinstellungen des Programmes SPSS wurden verändert, um die Aussagekraft des Outputs zu steigern.

Der dabei anschließend erzeugte Ergebnisausdruck wird im Folgenden entsprechend den Analyseschritten des Kapitels 7.2 nachvollzogen und kommentiert.

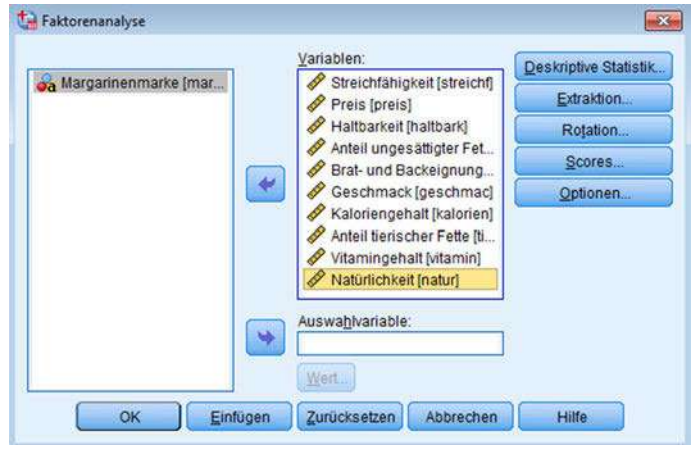


Abbildung 7.46: Dialogfeld „Faktorenanalyse“

7.3.2 Ergebnisse

1. Variablenauswahl und Errechnung der Korrelationsmatrix

Datenstandardisierung

Im ersten Schritt wird zunächst die Datenmatrix standardisiert und in eine Korrelationsmatrix überführt. Dieser Schritt erfolgt manipulationsfrei, d. h. es ist keine (subjektive) Entscheidung des Forschers erforderlich und er besitzt somit auch keine Eingriffsmöglichkeit. Die Korrelationsmatrix der Margarinestudie ist in Abbildung 7.47 wiedergegeben.

Eignung der Korrelationsmatrix

In Abschnitt 7.2.1.2 wurde ausführlich dargelegt, dass die Art und Weise der Befragung, die Struktur der Befragten sowie die Verteilungen der Variablen in der Erhebungsgesamtheit zu einer Verzerrung der Ergebnisse der Faktorenanalyse führen können. Diese Verzerrungen schlagen sich in der Korrelationsmatrix nieder, und folglich kann die Eignung der Daten anhand der Korrelationsmatrix überprüft werden. Bereits die Korrelationsmatrix macht deutlich, dass in der vorliegenden Studie relativ häufig geringe Korrelationswerte auftreten (vgl. z. B. die Variable „Ungesättigte Fettsäuren“) und manche Korrelationen sogar nah bei Null liegen (z. B. die Korrelation zwischen „Backeignung“ und „Preis“). Bereits daraus lässt sich schließen, dass diese Variablen für faktoranalytische Zwecke wenig geeignet sind. Die in Abschnitt 7.2.1.2 behandelten Prüfkriterien bestätigen in der Summe diese Vermutung.

KMO-Kriterium

Beispielhaft sei hier jedoch nur das Kaiser-Meyer-Olkin-Kriterium näher betrachtet, das für die Korrelationsmatrix insgesamt nur einen Wert von $0,437$ erbrachte und damit die Korrelationsmatrix als „untragbar“ für die Faktorenanalyse deklariert. Welche Variablen dabei für dieses Ergebnis verantwortlich sind, macht Abbildung 7.48 deutlich.

Abbildung 7.47: Die Korrelationskoeffizienten

Korrelationsmatrix

		Streichfähigkeit	Preis	Haltbarkeit	Anteil ungesättigter Fettsäuren	Brat- und Backeignung	Geschmack	Kalorien-gehalt	Anteil tierischer Fette	Vitamin-gehalt	Natürlichkeit
Korrelation	Streichfähigkeit	1,00000	-,3853	,67996	,33627	-,30914	-,47235	-,76286	-,79821	-,18589	-,86041
	Preis	-,38528	1,000	-,31859	-,69955	,08853	,40894	,04698	,26197	-,10843	,34815
	Haltbarkeit	,67996	-,3186	1,00000	,21163	,08494	-,11325	-,58603	-,50281	,03207	-,50802
	Anteil ungesättigter Fettsäuren	,33627	-,6996	,21163	1,00000	-,41754	-,30354	-,21686	-,20350	,18030	-,14506
	Brat- und Backeignung	-,30914	,0885	,08494	-,41754	1,00000	,66559	,59178	,45295	,21888	,44121
	Geschmack	-,47235	,4089	-,11325	-,30354	,66559	1,00000	,51782	,80040	,43373	,72025
	Kaloriengehalt	-,76286	,0470	-,58603	-,21686	,59178	,51782	1,00000	,81204	,36565	,76014
	Anteil tierischer Fette	-,79821	,2620	-,50281	-,20350	,45295	,80040	,81204	1,00000	,53309	,87546
	Vitamingehalt	-,18589	-,1084	,03207	,18030	,21888	,43373	,36565	,53309	1,00000	,45577
	Natürlichkeit	-,86041	,3481	-,50802	-,14506	,44121	,72025	,76014	,87546	,45577	1,00000

Abbildung 7.48: Anti-Image-Korrelations-Matrix der Margarinestudie

Anti-Image-Matrizen

		Streich-fähigkeit	Preis	Haltbarkeit	Anteil ungesättigter Fettsäuren	Brat- und Backeignung	Geschmack	Kalorien-gehalt	Anteil tierischer Fette	Vitamin-gehalt
Korrelation	Streichfähigkeit	,45777[a]	,33981	-,10512	-,11143	,61480	-,84225	-,62569	,83545	-,75416
	Preis	,33981	,36966[a]	,08188	,76538	,61519	-,58983	-,29977	,55726	-,26705
	Haltbarkeit	-,10512	,08188	,74145[a]	-,10522	-,41957	,13870	,49891	-,14367	-,14926
	Anteil ungesättigter Fettsäuren	-,11143	,76538	-,10522	,48550[a]	,46223	-,22180	-,15505	,22193	,01675
	Brat- und Backeignung	,61480	,61519	-,41957	,46223	,28670[a]	-,88320	-,88693	,86260	-,44878
	Geschmack	-,84225	-,58983	,13870	-,22180	-,88320	,36880[a]	,82444	-,97126	,66229
	Kaloriengehalt	-,62569	-,29977	,49891	-,15505	-,88693	,82444	,46496[a]	-,83898	,46026
	Anteil tierischer Fette	,83545	,55726	-,14367	,22193	,86260	-,97126	-,83898	,45227[a]	-,71028
	Vitamingehalt	-,75416	-,26705	-,14926	,01675	-,44878	,66229	,46026	-,71028	,27807[a]
	Natürlichkeit	,84396	-,02622	-,03962	-,41600	,37170	-,63532	-,52367	,57640	-,65216

Stichprobeneignung

Das variablenspezifische Kaiser-Meyer-Olkin-Kriterium (MSA-Kriterium) ist auf der Hauptdiagonalen der Anti-Image-Korrelations-Matrix abgetragen und weist nur die Variablen „Natürlichkeit“ und „Haltbarkeit“ als „kläglich“ bzw. „ziemlich gut“ für faktoranalytische Zwecke aus. Es ist deshalb angezeigt, Variablen aus der Analyse sukzessive auszuschließen (beginnend mit der Variablen „Vitamingehalt“), bis alle variablenspezifischen MSA-Kriterien größer als $0,5$ sind. In unserem Fall würde dieser Prozess dazu führen, dass insgesamt acht Variable aus der Analyse herausgenommen werden müssten. Aus *didaktischen Gründen* wird hier jedoch auf den Ausschluss von Variablen verzichtet.

MSA-Kriterium

Um im Rahmen der SPSS-Anwendung die beiden Eignungskriterien der Korrelationskoeffizienten und die Anti-Image-Matrix zu erhalten, sind im Dialogfeld Deskriptive Statistik die beiden Felder „Koeffizienten“ und „Anti-Image“ auszuwählen. Neben den in dieser Fallstudie vorgestellten Anwendungsmöglichkeiten der deskriptiven Statistik können hier weitere Korrelationsauswertungen selektiert werden (vgl. Abbildung 7.49).

SPSS-Anwendung

Abbildung 7.49: Dialogfeld „Deskriptive Statistik“

2. Bestimmung der Kommunalitäten

Bei diesem Schritt erfolgt der erste Eingriff des Anwenders in den Ablauf der Faktorenanalyse, da eine Schätzung der Kommunalitäten, also des Anteils der durch die gemeinsamen Faktoren zu erklärenden Varianz der Variablen, vorgenommen werden muss. Wir wollen hier nochmals die beiden theoretischen Konzepte der Hauptkomponentenanalyse und der Hauptachsenanalyse – als wichtiges Schätzverfahren der Faktorenanalyse – vergleichen: Obwohl beide Methoden die gleichen Rechenschritte durchlaufen, kommen sie aufgrund des Setzens unterschiedlicher Startwerte für die Kommunalitäten zu unterschiedlichen Ergebnisse, die in Abbildung 7.50 dargestellt sind. Während bei der Hauptkomponentenanalyse die Startwerte der Kommunalitätenschätzung immer auf Eins festgelegt werden, wird bei der Hauptachsenanalyse als Startwert das multiple Bestimmtheitsmaß der Variablen gewählt. Die Analysen führen bei Extraktion von drei Faktoren zu den ebenfalls in Abbildung 7.50 aufgeführten Ergebnissen (Endwerten). Es wird deutlich, dass die „Endwerte“ zum Teil erheblich von den Startwerten abweichen und die Ergebnisse einer Hauptkomponentenanalyse von denen der Hauptachsenanalyse abweichen.

Kommunalitätenschätzung

	Hauptkomponentenanalyse		Hauptachsenanalyse	
Variable	Kommunalität (Startwerte)	Kommunalität (Endwerte)	Kommunalität (Startwerte)	Kommunalität (Endwerte)
STREICHF	1.00000	.88619	.97414	.85325
PREIS	1.00000	.76855	.89018	.55717
HALTBARK	1.00000	.89167	.79497	.85754
UNGEFETT	1.00000	.85324	.85847	.91075
BACKEIGN	1.00000	.76043	.96501	.55819
GESCHMAC	1.00000	.84012	.98810	.82330
KALORIEN	1.00000	.80223	.97132	.73903
TIERFETT	1.00000	.92668	.99166	.94796
VITAMIN	1.00000	.63297	.78019	.40402
NATUR	1.00000	.88786	.96445	.87851

Abbildung 7.50: Vergleich der geschätzten Kommunalitäten

Bei der Hauptkomponentenanalyse liegt das darin begründet, dass hier weniger Faktoren als Variable extrahiert wurden. Würde man bei diesen beiden Verfahren ebenfalls 10 Faktoren extrahieren, so würden die Start- und Endwerte der Kommunalitätenschätzung übereinstimmen. Bei der mit iterativer Kommunalitätenschätzung durchgeführten Hauptachsenanalyse sind die „wahren Endwerte“ der Kommunalitätenschätzung unbekannt und werden aufgrund der Konvergenz des Iterationsprozesses (Konvergenzkriterium) bestimmt. Für die weiteren Betrachtungen sind die „Endwerte“ der Kommunalitätenschätzung jedoch von entscheidender Bedeutung, da der Erklärungswert der gefundenen Faktoren immer auch im Hinblick auf die zugrundeliegende Kommunalität zu beurteilen ist.

Konvergenzkriterium

3. Wahl der Extraktionsmethode

Es wurde gezeigt, dass die beiden grundlegenden Konzepte *Hauptkomponentenanalyse* und *Faktorenanalyse* auf unterschiedlichen theoretischen Modellen basieren. Für die Margarinestudie sei unterstellt, dass spezifische Varianzen und Messfehler relevant sind. Im Folgenden wird deshalb eine *Hauptachsenanalyse* angewendet.

SPSS-Dialog

Hierfür ist eine Änderung der SPSS-Voreinstellungen erforderlich. In der Dialogbox „Extraktion“ ist im Auswahlfenster „Methode“ aus der sich öffnenden Liste „Hauptachsen-Faktorenanalyse“ auszuwählen (vgl. Abbildung 7.51). Neben der Hauptachsenanalyse bietet SPSS aber noch weitere Verfahrensvarianten zur Faktorenextraktion, die in Abbildung 7.27 erläutert sind.

4. Zahl der Faktoren

Die Zahl der maximal möglichen Faktoren entspricht der Zahl der Variablen: Dann entspricht *jeder* Faktor *einer* Variablen. Da aber gerade die Zahl der Faktoren kleiner als die der Variablen sein soll, ist zu entscheiden, wie viele Faktoren (Zahl der Faktoren $<$ Zahl der Variablen) extrahiert werden sollen.

Wie bereits gezeigt, existieren zur Lösung dieses Problems verschiedene Vorschläge, ohne dass auf eine theoretisch befriedigende Alternative zurückgegriffen werden kann.

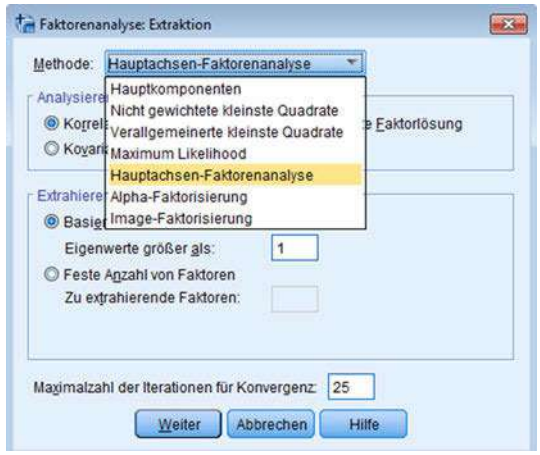


Abbildung 7.51: Dialogfeld „Extraktion“

Beispielhafte Alternativen, die von SPSS unterstützt werden, sind in Abbildung 7.52 aufgelistet.

SPSS-Lösungsalternativen

Unabhängig davon, welches Kriterium man zur Extraktion der Faktoren verwendet, ist es zunächst sinnvoll, so viele Faktoren zu extrahieren, wie Variablen vorhanden sind. Dies erfolgt bei Anwendung des Programmes SPSS automatisch. Hierbei werden allerdings unabhängig von der gewählten Extraktionsmethode die anfängliche Lösung und die Zahl der Faktoren nach der Haupt*komponenten*methode bestimmt. Die Anzahl der vorhandenen Variablen, die in einem korrelierten Verhältnis zueinander stehen, wird in diesem ersten Auswertungsschritt in eine gleich große Anzahl unkorrelierter Variablen umgewandelt. Die eigentliche Faktorenanalyse auf Basis der Hauptachsenanalyse hat zu diesem Zeitpunkt folglich noch nicht stattgefunden. Abbildung 7.53 zeigt den entsprechenden SPSS-Ausdruck der automatisierten Haupt-

	In der Literatur vorgeschlagene Kriterien zur Bestimmung der Faktoranzahl	Bei SPSS realisierte Alternativen
1.	Extrahiere solange, bis X % (z. B. 95 %) der Varianz erklärt sind.	Kann ex post manuell bestimmt werden.
2.	Extrahiere nur Faktoren mit Eigenwerten größer 1 (Kaiser-Kriterium).	Von SPSS automatisch verwandt, wenn keine andere Spezifikation.
3.	Extrahiere n (z. B. 3) Faktoren.	Anzahl kann im Dialogfeld ,,Extraktion “ ex ante manuell eingegeben werden.
4.	Scree-Test: Die Faktoren werden nach Eigenwerten in abfallender Reihenfolge geordnet. An die Faktoren mit den niedrigsten Eigenwerten wird eine Gerade angepasst. Der letzte Punkt auf der Geraden bestimmt die Faktorenzahl.	Erforderlicher Screeplot kann im Dialogfeld ,,Extraktion “ ebenfalls angefordert werden.
5.	Zahl der Faktoren soll kleiner als die Hälfte der Zahl der Variablen sein.	Kann ex post manuell bestimmt werden, sofern im Dialogfeld ,,Extraktion “ die eingetragene Zahl zu extrahierender Faktoren nicht kleiner als die Hälfte der Zahl der Variablen ist.
6.	Extrahiere alle Faktoren, die nach der Rotation interpretierbar sind.	Kann nach Einstellung des erwüschten Rotationsprinzips ex post manuell bestimmt werden.

Abbildung 7.52: Ausgewählte Faktorextraktionskriterien

Erklärte Gesamtvarianz

Komponente	Summen von quadrierten Faktorladungen für Extraktion: Gesamt	% der Varianz	Kumulierte %
1	5,05188	50,51883	50,51883
2	1,77106	17,71061	68,22944
3	1,42700	14,27002	82,49946
4	,81935	8,19349	90,69295
5	,42961	4,29611	94,98905
6	,24709	2,47085	97,45991
7	,15928	1,59275	99,05266
8	,06190	,61902	99,67168
9	,02943	,29434	99,96602
10	,00340	,03398	100,00000

Extraktionsmethode: Hauptkomponentenanalyse.

Abbildung 7.53: Extrahierte Faktoren mit Eigenwerten und Varianzerklärungsanteil

komponentenanalyse. Nach der Faustregel (95 % Varianzerklärung) würden sich fünf Faktoren ergeben.

Extraktionskriterium

Abbildung 7.54 zeigt die entsprechende Zahl der Faktoren für das Kaiser-Kriterium und den Scree-Test. Während nach dem Kaiser-Kriterium drei Faktoren zu extrahieren wären, legt der Scree-Test eine Ein-Faktor-Lösung nahe. Wegen der unterschiedlichen Ergebnisse der drei Extraktionskriterien muss sich der Anwender *subjektiv* für eine der Lösungen entscheiden.

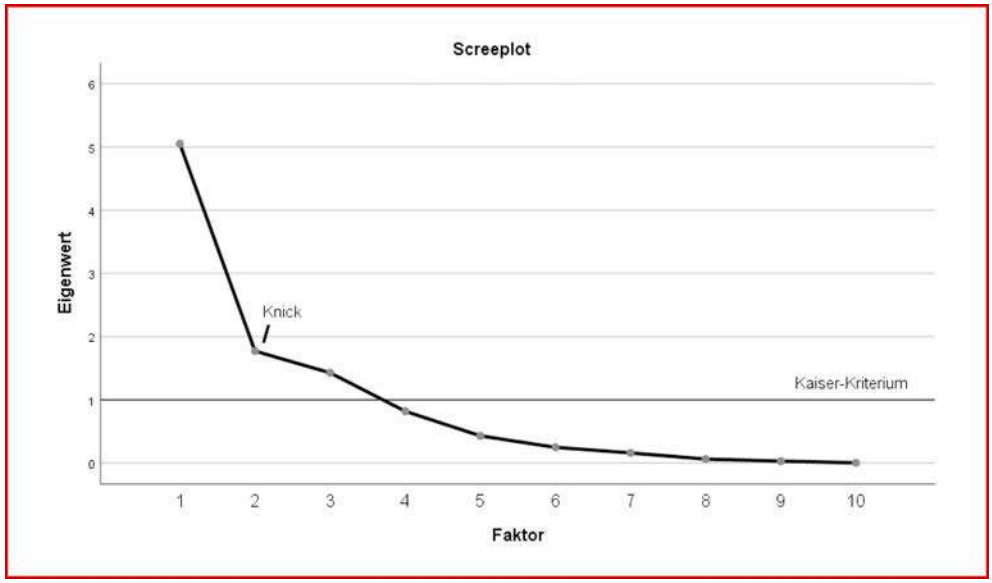


Abbildung 7.54: Scree-Test und Kaiser-Kriterium

Im vorliegenden Fallbeispiel wird das Kaiser-Kriterium als Extraktionskriterium verwendet. Dies entspricht der Voreinstellung von SPSS.

Beurteilung der Lösung mit SPSS

Um die Güte der 3-Faktorenlösung zu bestimmen, sind weitere SPSS-Outputs näher zu betrachten (vgl. Abbildung 7.55 bis Abbildung 7.57). Da es sich hierbei um keine anfängliche Lösung (automatisierte Hauptkomponentenanalyse), sondern um eine Lösung nach Durchführung zahlreicher Iterationsschritte handelt, ist die zuvor ausgewählte Extraktionsmethode (Hauptachsenanalyse) zur Anwendung gekommen.

Bei zehn Variablen beträgt die Gesamtvarianz wegen der Normierung jeder Einzelvarianz auf den Wert von 1 gleich 10. Das bedeutet z. B. für den ersten Faktor in Abbildung 7.55 mit einem Eigenwert von 4,864 im Verhältnis zu 10 einen Erklärungsanteil von ca. 48,6 % der Gesamtvarianz. Insgesamt beträgt die Summe der drei Eigenwerte 7,523. Setzt man diese Summe ins Verhältnis zur Gesamtvarianz von 10, so ergibt sich ein durch die Faktoren erklärter Varianzanteil von 75,3 % (vgl. Spalte „Kumulierte %“ in Abbildung 7.55).

Erklärte Gesamtvarianz

Faktor	Summen von quadrierten Faktorladungen für Extraktion		
	Gesamt	% der Varianz	Kumulierte %
1	4,86417	48,64170	48,64170
2	1,46805	14,68054	63,32225
3	1,19749	11,97491	75,29716

Extraktionsmethode: Hauptachsen-Faktorenanalyse.

Abbildung 7.55: Eigenwerte und Anteile erklärter Varianz

Die in der Übersicht ausgewiesenen Varianzerklärungsanteile (% der Varianz) geben also an, wieviel der jeweilige Faktor an Erklärunganteil in Bezug auf *alle* Ausgangsvariablen besitzt. Diese drei Faktoren erklären zusammen 75,3 % der Ausgangsvarianz, wobei der 1. Faktor 48,6 %, der 2. Faktor 14,7 % und der 3. Faktor 12,0 % der Ausgangsvarianz erklären. Der Eigenwert der drei Faktoren (erklärter Teil der Gesamtvarianz eines Faktors) kann in der Spalte „Gesamt“ (Abb. 7.55) abgelesen werden. Wie der Leser leicht nachvollziehen kann, ergibt sich nur 1 Faktor (vgl. Abb. 7.54).

Unrotierte Faktorladungsmatrix

Abbildung 7.56 enthält die unrotierte Faktorladungsmatrix, wobei die Faktorladungen der extrahierten Faktoren nach ihrer Ladungsgröße sortiert wurden. Dabei wird deutlich, dass die Variablen „Anteil tierischer Fette“, „Natürlichkeit“, „Streichfähigkeit“, „Kaloriengehalt“, „Geschmack“ und „Backeignung“ offenbar „viel mit Faktor 1 zu tun haben“, während Faktor 2 offenbar mit den Variablen „Ungesättigte Fettsäuren“, „Preis“ und „Vitamingehalt“ und Faktor 3 vor allem mit „Haltbarkeit“ korreliert.

Ergebnisverzerrungen

Die iterativ geschätzten Kommunalitäten auf Basis der Hauptachsenanalyse werden in Abbildung 7.57 widergespiegelt (vgl. auch Abbildung 7.50). Auffällig ist dabei vor allem, dass offenbar die Varianzanteile der Variablen „Preis“, „Backeignung“ und „Vitamingehalt“ nur zu einem sehr geringen Teil durch die gefundenen Faktoren erklärbar sind. Daraus ergibt sich die Konsequenz, dass diese Variablen tendenziell zu *Ergebnisverzerrungen* führen und von daher aus der Analyse ausgeschlossen werden sollten. Dabei handelt es sich aber gerade um diejenigen Variablen, die bereits nach dem Kaiser-Meyer-Olkin-Kriterium (vgl. Abbildung 7.48) aus der Analyse auszuschließen

Faktorenmatrix[a]

	Faktor 1	2	3
Streichfähigkeit	-,86273	,10548	,31276
Preis	,40050	-,62446	,08261
Haltbarkeit	-,56373	,23939	,69458
Anteil ungesättigter Fettsäuren	-,40207	,80846	-,30900
Brat- und Backeignung	,54555	,04984	,50802
Geschmack	,77638	,09144	,46062
Kaloriengehalt	,83090	,18542	-,11935
Anteil tierischer Fette	,94758	,22325	-,01469
Vitamingehalt	,38346	,48337	,15273
Natürlichkeit	,91885	,16537	-,08292

Extraktionsmethode: Hauptachsenfaktorenanalyse.

a. Es wurde versucht, 3 Faktoren zu extrahieren. Es werden mehr als 25 Iterationen benötigt. (Konvergenz=,002). Die Extraktion wurde abgebrochen.

Abbildung 7.56: Unrotierte Faktorenmatrix

waren. Aus didaktischen Gründen wird hier allerdings wiederum auf den Ausschluss von Variablen verzichtet.

Ergebnis-Interpretation

Um aus den vielen Möglichkeiten der Positionierung eines Koordinatenkreuzes die beste, d. h. interpretationsfähigste, bestimmen zu können, wird das oben ermittelte Faktorenmuster rotiert.

Kommunalitäten

	Extraktion
Streichfähigkeit	,85325
Preis	,55717
Haltbarkeit	,85754
Anteil ungesättigter Fettsäuren	,91075
Brat- und Backeignung	,55819
Geschmack	,82330
Kaloriengehalt	,73903
Anteil tierischer Fette	,94796
Vitamingehalt	,40402
Natürlichkeit	,87851

Extraktionsmethode: Hauptachsen-Faktorenanalyse.

Abbildung 7.57: Kommunalitäten

Abbildung 7.58: Dialogfeld „Rotation“

Die rechtwinklige Rotation kann im zweidimensionalen (wie im dreidimensionalen) Fall grundsätzlich auch graphisch erfolgen, indem der Untersuchende versucht, das Koordinatenkreuz so zu drehen, dass möglichst viele Punkte im Koordinatenkreuz (Faktorladungen) auf einer der beiden Achsen liegen. Im Mehr-als-drei-Faktoren-Fall ist es allerdings notwendig, die Rotation analytisch vorzunehmen. SPSS stellt hierfür unterschiedliche Möglichkeiten der Rotation zur Verfügung, die bei Auswahl der Dialogbox „Rotation“ erscheinen (vgl. Abbildung 7.58).

Rotation in SPSS

Bei der hier angewendeten Varimax-Rotationsmethode handelt es sich um eine orthogonale Rotation. Die Faktorachsen verbleiben folglich bei der Rotation in einem rechten Winkel zueinander, was unterstellt, dass die Achsen bzw. Faktoren nicht untereinander korrelieren. Da die Rotation der Faktoren zwar die Faktorladungen, nicht aber die Kommunalitäten des Modells verändert, ist die unrotierte Lösung primär für die Auswahl der Anzahl an Faktoren und für die Gütebeurteilung der Faktorlösung geeignet. Eine Interpretation der ermittelten Faktoren ist auf Basis eines unrotierten Modells allerdings nicht empfehlenswert, da sich durch Anwendung einer Rotationsmethode die Verteilung des erklärten Varianzanteils einer Variable auf die Faktoren verändert.

Varimax-Rotation

Die analytische Lösung von SPSS auf der Basis des Varimax-Kriteriums beim vorliegenden Beispiel zeigt Abbildung 7.59.

Vergleicht man die Lösung der rotierten Faktorladungen mit den unrotierten (Abbildung 7.56), dann zeigt sich eine erhebliche Veränderung. Nach Rotation laden z. T. andere Variablen auf bestimmte Faktoren im Vergleich zur nicht rotierten Faktorladungsmatrix.

Rotierte Faktorenmatrix[a]

	Faktor 1	2	3
Streichfähigkeit	-,36302	-,81410	,24230
Preis	,06852	,23296	-,70584
Haltbarkeit	,12616	-,90249	,16474
Anteil ungesättigter Fettsäuren	-,12951	-,07381	,94262
Brat- und Backeignung	,69978	-,01333	-,26139
Geschmack	,84729	,17468	-,27365
Kaloriengehalt	,57633	,63728	-,02717
Anteil tierischer Fette	,73476	,63626	-,05711
Vitamingehalt	,55369	,12349	,28669
Natürlichkeit	,65040	,67002	-,08109

Extraktionsmethode: Hauptachsenfaktorenanalyse.
Rotationsmethode: Varimax mit Kaiser-Normalisierung.

a. Die Rotation ist in 7 Iterationen konvergiert.

Abbildung 7.59: Varimax-rotierte Faktormatrix

5. Faktorinterpretation

Faktorinterpretation

Welche Interpretation lässt diese Rotation zu? Zur Veranschaulichung ist es häufig sinnvoll, die hochladenden Variablen –wie in Abbildung 7.60 dargestellt– mit einem + oder − (positive oder negative Korrelation) in Bezug auf den jeweiligen Faktor zu kennzeichnen.

	Faktor 1 Gesundheit	Faktor 2 Naturbelassenheit	Faktor 3 Preis-/Leistungs-verhältnis
Geschmack	+		
Tierfette	+	+	
Backeignung	+		
Vitamine	+		
Haltbarkeit		−	
Streichfähigkeit		−	
Natürlichkeit	+	+	
Kalorien	+	+	
Unges. Fettsäuren			+
Preis			−

Abbildung 7.60: Schematische Darstellung der rotierten Faktorladungen

Dabei wird deutlich, dass Faktor 2 durch hohe Ladungen der Variablen „Haltbarkeit“, „Streichfähigkeit“, „Natürlichkeit“, „Kaloriengehalt“ und „Anteil tierischer Fette“ gekennzeichnet ist, wobei die beiden erst genannten Variablen negativ auf den Faktor laden. Die negativen und positiven Ladungen sind damit erklärbar, dass bei „natürlichen Produkten“ meist der „Anteil tierischer Fette“, der „Kaloriengehalt“ und damit die „Natürlichkeit“ in einem gegensetzlichen Verhältnis zu „Haltbarkeit“ und „Streichfähigkeit“ stehen. Das bedeutet, dass z. B. eine hohe „Haltbarkeit“ und „Streichfähigkeit“ meist mit geringem „Anteil tierischer Fette“, „Kaloriengehalt“ und damit „Natürlichkeit“ einhergeht. Die Korrelationen zwischen diesen Variablen lässt sich deshalb auf die Beurteilungsdimension „*Naturbelassenheit*“ zurückführen.

Faktor 2 „Naturbelassenheit“

Der Leser möge selber versuchen, unseren *Interpretationsvorschlag* für die beiden übrigen Faktoren nachzuvollziehen. Dabei wird schnell deutlich werden, welche Schwierigkeiten eine gewissenhafte und sorgfältige Interpretation (entsprechend dem theoretischen Modell des angewandten Verfahrens) bereiten kann.

Schwierigkeiten bei der Faktorinterpretation

Häufig ist es allerdings notwendig, die Daten detaillierter zu analysieren, um die Ergebnisse einer Rotation richtig zu deuten. Gerade beim Rotationsproblem eröffnen sich erhebliche Manipulationsspielräume. Damit eröffnet die Faktorenanalyse auch Spielräume für Missbrauch.

6. Bestimmung der Faktorwerte (Factorscores)

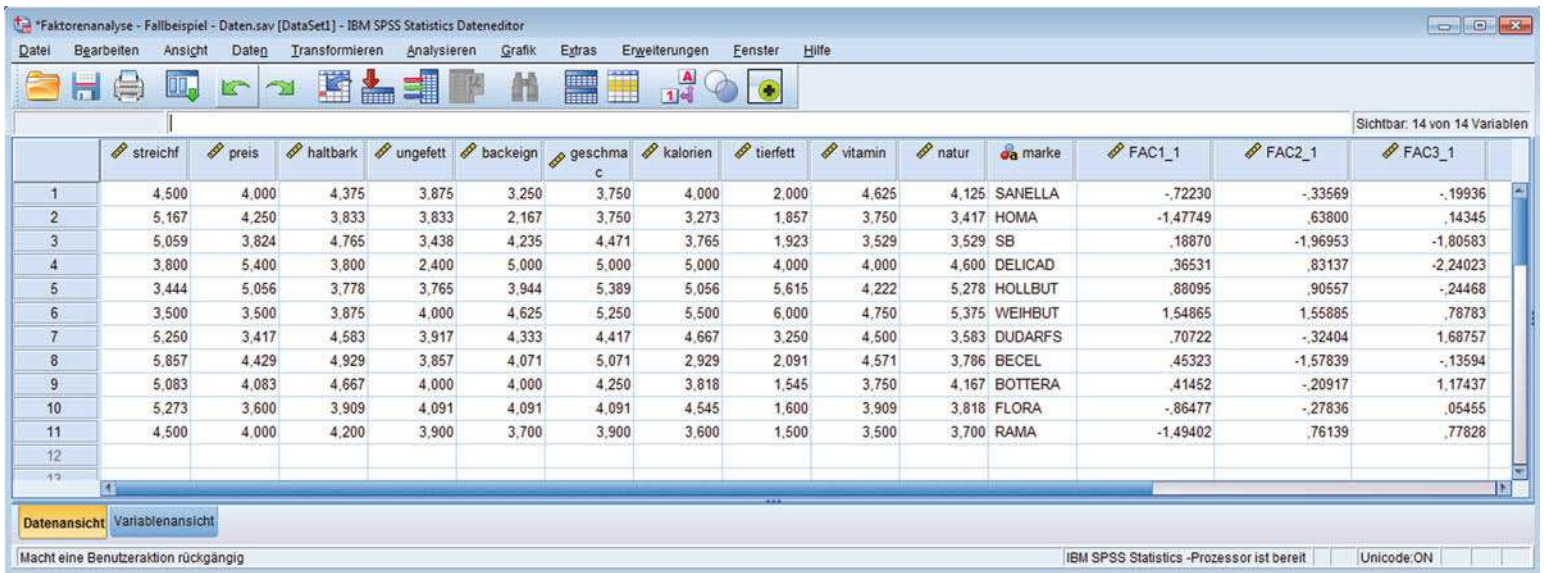
Nach Extraktion der drei Faktoren interessiert häufig auch, wie die verschiedenen Marken anhand dieser drei Faktoren beurteilt wurden. Auf dieser Basis lassen sich beispielsweise Produktpositionierungen vornehmen. Auch dazu sind Schätzungen notwendig. Empirische Untersuchungen haben gezeigt, dass je nach verwendeter Schätzmethode die Ergebnisse erheblich variieren können.

Faktorwerte ≠ Faktorladungen

Abbildung 7.61: Dialogfeld „Faktorscores“

In der Regel erfolgt die Schätzung der *Faktorwerte* (bei SPSS *Factorscores*), die streng von den Faktor*ladungen* zu trennen sind, –wie auch in SPSS– durch eine multiple Regressionsrechnung. SPSS bietet drei Verfahren zur Schätzung von Faktorwerten an, die zu unterschiedlichen Werten führen. Zur Einstellung der gewünschten Schätzmethode ist das Dialogfeld „Faktorscores“ auszuwählen (vgl. Abbildung 7.61). Alle drei zur Verfügung stehenden Schätzverfahren führen zu standardisierten Faktorwerten, mit einem Mittelwert von 0 und einer Standardabweichung von 1. Durch die Auswahl der hier verwendeten Methode „Regression“ können die zu ermittelnden

Faktorwerteschätzung in SPSS



	streichf	preis	haltbark	ungefett	backeign	geschmac	kalorien	tierfett	vitamin	natur	marke	FAC1_1	FAC2_1	FAC3_1
1	4,500	4,000	4,375	3,875	3,250	3,750	4,000	2,000	4,625	4,125	SANELLA	-,72230	-,33569	-,19936
2	5,167	4,250	3,833	3,833	2,167	3,750	3,273	1,857	3,750	3,417	HOMA	-1,47749	,63800	,14345
3	5,059	3,824	4,765	3,438	4,235	4,471	3,765	1,923	3,529	3,529	SB	,18870	-1,96953	-1,80583
4	3,800	5,400	3,800	2,400	5,000	5,000	5,000	4,000	4,000	4,600	DELICAD	,36531	,83137	-2,24023
5	3,444	5,056	3,778	3,765	3,944	5,389	5,056	5,615	4,222	5,278	HOLLBUT	,88095	,90557	-,24468
6	3,500	3,500	3,875	4,000	4,625	5,250	5,500	6,000	4,750	5,375	WEIHBUT	1,54865	1,55885	,78783
7	5,250	3,417	4,583	3,917	4,333	4,417	4,667	3,250	4,500	3,583	DUDARFS	,70722	-,32404	1,68757
8	5,857	4,429	4,929	3,857	4,071	5,071	2,929	2,091	4,571	3,786	BECEL	,45323	-1,57839	-,13594
9	5,083	4,083	4,667	4,000	4,000	4,250	3,818	1,545	3,750	4,167	BOTTERA	,41452	-,20917	1,17437
10	5,273	3,600	3,909	4,091	4,091	4,091	4,545	1,600	3,909	3,818	FLORA	-,86477	-,27836	,05455
11	4,500	4,000	4,200	3,900	3,700	3,900	3,600	1,500	3,500	3,700	RAMA	-1,49402	,76139	,77828
12														

Abbildung 7.62: Die Faktorwerte in der Datenmatrix

Faktorwerte korrelieren, obwohl – wie im Fall der Hauptachsenanalyse – die Faktoren orthogonal geschätzt wurden. Die zur Ermittlung der Faktorwerte erforderlichen *Regressionskoeffizienten* werden bei SPSS unter der Überschrift „Koeffizientenmatrix der Faktorwerte“ angezeigt. Hierbei handelt es sich nicht um die Faktorwerte, sondern um die Gewichtungsfaktoren, die mit den standardisierten Ausgangsdaten multipliziert werden müssen, um die endgültigen Faktorwerte zu errechnen. Der Datenmatrix werden die Faktorwerte der einzelnen Fälle bei SPSS als neue Variablen (fac1_1, fac2_1 und fac3_1) angehängt (vgl. Abbildung 7.62).

Für den Fall, dass für bestimmte Variable einzelne Probanden keine Aussagen gemacht haben (Problem der missing values), gilt:

Fehlende Werte

1. Die Fallzahl verringert sich für die entsprechende Variable.
2. Für diesen Fall können keine Faktorwerte berechnet werden.

Da in unsere Analyse nicht die Aussagen der einzelnen Probanden eingingen (vgl. dazu Kapitel 7.1), sondern für die elf Marken die Mittelwerte über alle Probanden, waren diese Effekte nicht relevant.

Faktorwerte-Plot

Stellt man die Faktorwerte der beiden ersten Faktoren graphisch dar (auf die Darstellung des 3. Faktors wird aus Anschauungsgründen verzichtet, da dies eine dreidimensionale Abbildung erfordern würde), so ergeben sich folgende *Produktpositionen* für die elf Aufstrichfette (Abbildung 7.63).

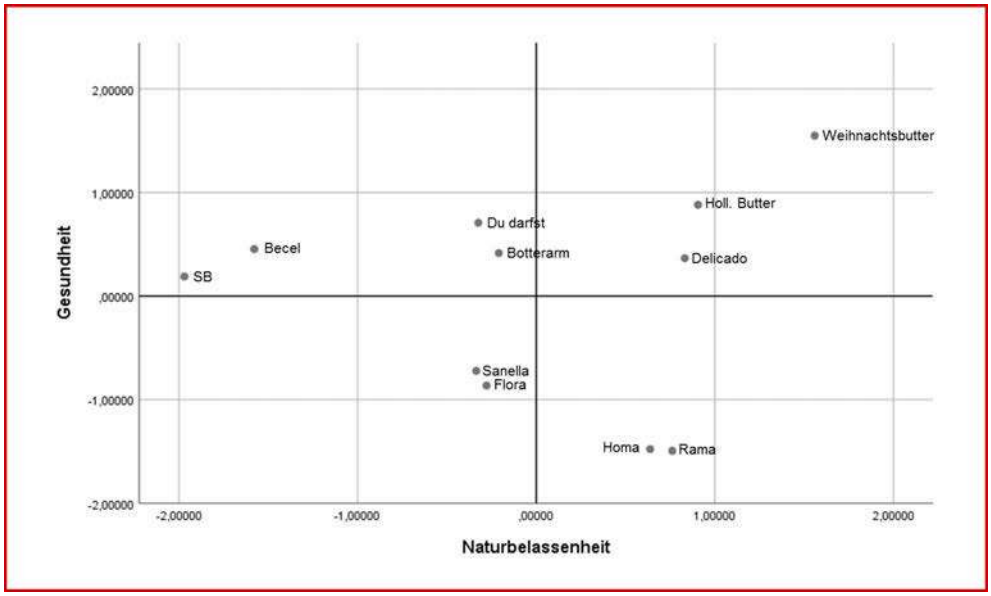


Abbildung 7.63: Graphische Darstellung der Faktorwerte

Die Achsen stellen in Abbildung 7.63 die beiden ersten extrahierten Faktoren dar und die Punkte im Koordinatenkreuz geben die jeweiligen Positionen der Marken in Bezug auf die beiden Faktoren an (Faktorwerte). Produkt 3 (SB) hat beispielsweise die Interpretation Koordinaten $0,189/-1,970$ (vgl. die Werte in Abbildung 7.62). Bei einer 2-faktoriellen Lösung gibt diese Position an, dass offenbar die Befragten, welche die ursprünglichen zehn Variablen bewertet hatten, bei einer „Bündelung“ der zehn Variablen zu zwei unabhängigen Faktoren Produkt 3 in Bezug auf Faktor 1 (Gesundheit) positiv und Faktor 2 (Naturbelassenheit) relativ negativ bewerten. Entsprechendes gilt für die Bewertung (Positionierung) der übrigen zehn Marken.

Als Ergebnis zeigt sich, dass z. B. die Marken „HOMA“ und „RAMA“ ebenso wie die Buttersorten (Holl. Butter, Weihnachtsbutter und Delicado Sahnebutter) im Vergleich zu den übrigen Produkten eine Extremposition einnehmen.

Bei der inhaltlichen Interpretation der Faktorwerte ist darauf zu achten, dass sie aufgrund der Standardisierung der Ausgangsdatenmatrix ebenfalls standardisierte Größen darstellen. Für die Interpretation der Faktorwerte bedeutet das folgendes:

- Ein negativer Faktorwert besagt, dass ein Produkt (Objekt) in Bezug auf diesen Faktor *im Vergleich zu allen anderen* betrachteten Objekten unterdurchschnittlich ausgeprägt ist.
- Ein Faktorwert von 0 besagt, dass ein Produkt (Objekt) in Bezug auf diesen Faktor eine *dem Durchschnitt entsprechende* Ausprägung besitzt.
- Ein positiver Faktorwert besagt, dass ein Produkt (Objekt) in Bezug auf diesen Faktor *im Vergleich zu allen anderen* betrachteten Objekten überdurchschnittlich ausgeprägt ist.

Damit sind z. B. die Koordinatenwerte der Marke SB mit $0,189/-1,970$ wie folgt zu interpretieren: Bei SB wird die Gesundheit (Faktor 1) im Vergleich zu den übrigen Marken als (leicht) überdurchschnittlich stark ausgeprägt angesehen, während die Naturbelassenheit (Faktor 2) als nur unterdurchschnittlich stark ausgeprägt eingeschätzt wird. Dabei ist zu beachten, dass die Faktorwerte unter Verwendung *aller* Faktorladungen aus der rotierten Faktorladungsmatrix (Abbildung 7.59) berechnet werden. Somit haben auch kleine Faktorladungen einen Einfluss auf die Größe der Faktorwerte. Das bedeutet in unserem Beispiel, dass insbesondere die Faktorwerte Generalfaktor bei Faktor 1, der einen *Generalfaktor* darstellt (d. h. durchgängig vergleichbar hohe Ladungen aufweist), *nicht nur* durch die in Abbildung 7.59 gegebenen Werte bestimmt werden, sondern auch *alle* anderen Variablen einen Einfluss –wenn z. T. auch nur einen geringen– auf die Bestimmung der Faktorwerte ausüben.

Solche Informationen lassen sich z. B. für Marktsegmentierungsstudien verwenden, Marktnischen indem durch die Faktorenanalyse Marktnischen aufgedeckt werden können. So findet sich z. B. im Bereich links unten (geringe Gesundheit und geringe Naturbelassenheit) kein Produkt. Stellt sich heraus, dass diese Kombination von Merkmalen für Emulsionsfette von ausreichend vielen Nachfragern gewünscht wird, so kann diese Marktnische durch ein neues Produkt mit eben diesen Eigenschaften geschlossen werden.

7.3.3 SPSS-Kommandos

Neben der Möglichkeit, die oben aufgezeigte explorative Faktorenanalyse menugestützt durchzuführen, kann die Auswertung ebenfalls mit der nachfolgenden Syntaxdatei gerechnet werden.

```
*MVA: Fallbeispiel Faktorenanalyse
* DATENDEFINITION.
Data List Fixed / Streichfähigkeit Preis Haltbarkeit ungesFettsaueren Backeignung Geschmack
Kaloriengehalt
            Tierfett Vitamingehalt Natuerlichkeit 1-50(3) Marke 51-60 (A).

BEGIN DATA
4500 4000 4375 3875 3250 3750 4000 2000 4625 4125 SANELLA
5167 4250 3833 3833 2167 3750 3273 1857 3750 3417 HOMA
5059 3824 4765 3438 4235 4471 3765 1923 3529 3529 SB
3800 5400 3800 2400 5000 5000 5000 4000 4000 4600 DELICAD
3444 5056 3778 3765 3944 5389 5056 5615 4222 5278 HOLLBUT
3500 3500 3875 4000 4625 5250 5500 6000 4750 5375 WEIHBUT
5250 3417 4583 3917 4333 4417 4667 3250 4500 3583 DUDARFS
5857 4429 4929 3857 4071 5071 2929 2091 4571 3786 BECEL
5083 4083 4667 4000 4000 4250 3818 1545 3750 4167 BOTTERA
5273 3600 3909 4091 4091 4091 4545 1600 3909 3818 FLORA
4500 4000 4200 3900 3700 3900 3600 1500 3500 3700 RAMA
END DATA.


* PROZEDUR.
* Hauptachsenanalyse (PAF) für den Margarinemarkt.
FACTOR
 /VARIABLES Streichfähigkeit TO Natuerlichkeit
 /MISSING LISTWISE
 /ANALYSIS ALL
 /PRINT INITIAL CORRELATION SIG DET KMO INV REPR AIC EXTRACTION ROTATION FSCORE
 /PLOT EIGEN
 /CRITERIA MINEIGEN(1) ITERATE(25)
 /EXTRACTION PAF
 /CRITERIA ITERATE(25)
 /ROTATION VARIMAX
 /SAVE REG(ALL,FAC1_)
 /METHOD=CORRELATION.


* Ausgabe der Faktorwerte für alle Margarinemarken.
FORMATS FAC1_1 TO FAC1_3 (f8.5).
LIST VARIABLES = FAC1_1 TO FAC1_3 Marke.
```

Abbildung 7.64: SPSS-Job zur Faktorenanalyse

7.4 Anwendungsempfehlungen

7.4.1 Probleme bei der Anwendung der Faktorenanalyse

7.4.1.1 Unvollständig beantwortete Fragebögen: Das Missing Value-Problem

Beim praktischen Einsatz der Faktorenanalyse steht der Anwender häufig vor dem Problem, dass die Fragebögen nicht alle vollständig ausgefüllt sind.[17] Um die fehlenden Werte (missing values) im Programm handhaben zu können, bietet SPSS drei Optionen an. Zur Auswahl einer der Alternativen ist die Dialogbox „Optionen“ zu öffnen (vgl. Abbildung 7.65).

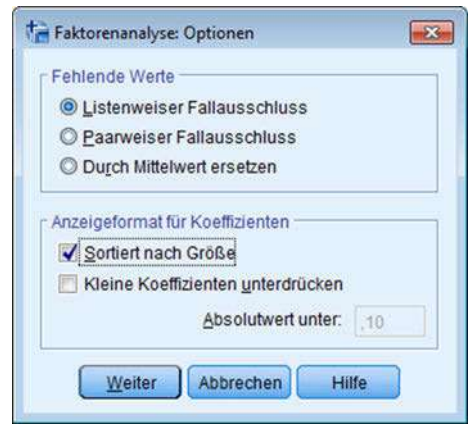


Abbildung 7.65: Das Dialogfeld „Optionen“

Folgende Optionen stehen dem Anwender konkret zur Auswahl:

Listwise Deletion

1. Die Werte werden „fallweise“ ausgeschlossen („*Listenweiser Fallausschluss*“), d. h. sobald ein fehlender Wert bei einer Variablen auftritt, wird der gesamte Fragebogen aus der weiteren Analyse ausgeschlossen. Dadurch wird die Fallzahl häufig erheblich reduziert!

Pairwise Deletion

2. Die Werte werden *variablenweise* ausgeschlossen („*Paarweiser Fallausschluss*“), d. h. bei Fehlen eines Wertes wird nicht der gesamte Fragebogen eliminiert, sondern lediglich die betroffene Variable. Dadurch wird zwar nicht die Fallzahl insgesamt reduziert, aber bei der Durchschnittsbildung liegen pro Variable unterschiedliche Fallzahlen vor. Dadurch kann es zu einer Ungleichgewichtung der Variablen kommen.

Mittelwert-Imputation

3. Es erfolgt überhaupt kein Ausschluss. Für die fehlenden Werte pro Variable werden *Durchschnittswerte* („*Durch Mittelwert ersetzen*“) eingefügt.

Je nachdem, welches Verfahren der Anwender zugrunde legt, können unterschiedliche Ergebnisse resultieren, sodass hier ein weiterer Manipulationsspielraum vorliegt.

[17] Vgl. grundsätzlich Backhaus/Blechschmidt (2009).

7.4.1.2 Starke Streuung der Antworten: Das Problem der Durchschnittsbildung

In unserem Fallbeispiel hatte die Befragung eine *dreidimensionale Matrix* ergeben (Abbildung 7.66).

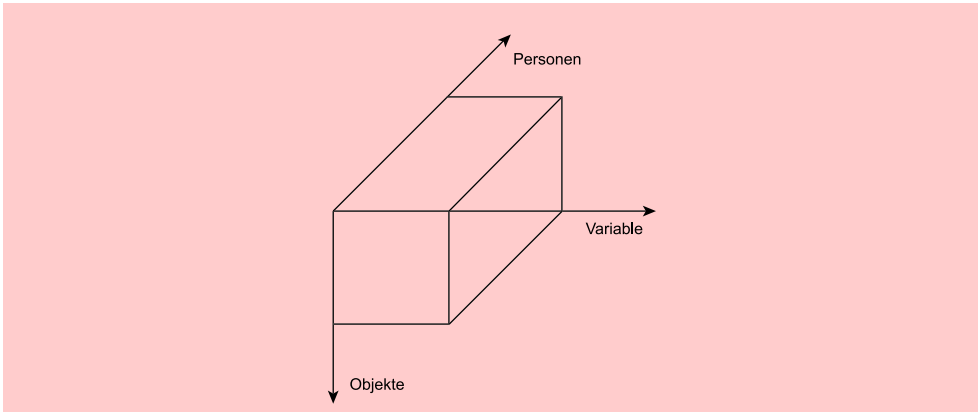


Abbildung 7.66: Der „Datenquader“

3-modale Analyse

18 Personen hatten 11 Objekte (Marken) anhand von 10 Eigenschaften beurteilt. Diese dreidimensionale Datenmatrix hatten wir durch Bildung der Durchschnitte über die 18 Personen auf eine zweidimensionale Objekte/Variablen-Matrix verdichtet. Diese Durchschnittsbildung verschenkt aber die Informationen über die personenbezogene Streuung der Daten. Ist diese Streuung groß, wird also auch viel Informationspotential verschenkt.

Eine Möglichkeit, die personenbezogene Streuung in den Daten mit in die Analyse einfließen zu lassen, besteht darin, die Beurteilung der jeweiligen Marke für jede Person aufrecht zu erhalten, indem jede einzelne Markenbeurteilung durch jede Person als *ein* Objekt betrachtet wird. Die dreidimensionale Matrix in Abbildung 7.66 wird dann zu einer vergrößerten zweidimensionalen Matrix (Abbildung 7.67).

In diesem Falle werden aus den ursprünglich (durchschnittlich) bewerteten 11 Objekten (Marken) $11 \cdot 18 = 198$ Objekte (Da in unserem Fallbeispiel jedoch nicht alle Personen alle Marken beurteilt hatten, ergaben sich nur 127 Objekte).

Ergebnisunterschiede

Vergleicht man die Ergebnisse des „Durchschnittsverfahrens“ mit dem „personenbezogenen Objektverfahren“, dann können *erhebliche Unterschiede* in den Ergebnissen der Faktorenanalyse auftreten. Abbildung 7.68 stellt die Ergebnisse bei einer Zweifaktoren-Lösung gegenüber.

Dabei wird deutlich, dass sich die Faktorladungen erheblich verschoben haben. Unterschiede ergeben sich auch in den Positionierungen der Marken anhand der Faktorwerte. Abbildung 7.69 zeigt die Durchschnittspositionen der 11 Marken. Vergleicht man die Positionen von Rama und SB aus der Durchschnittspositionierung mit der personenbezogenen Positionierung in Abbildung 7.70, dann werden die Ergebnisunterschiede besonders deutlich.[18]

Heterogene Faktorwerte

Die Vielzahl unterdrückter Informationen bei der Mittelwertbildung führt über verschiedene Faktormuster letztlich auch zu recht heterogenen Faktorwertstrukturen und

[18] Zu einer grundsätzlichen Diskussion der Behandlung von dreidimensionalen Ausgangsdatenmatrizen vgl. Krolak-Schwerdt (1991), passim.

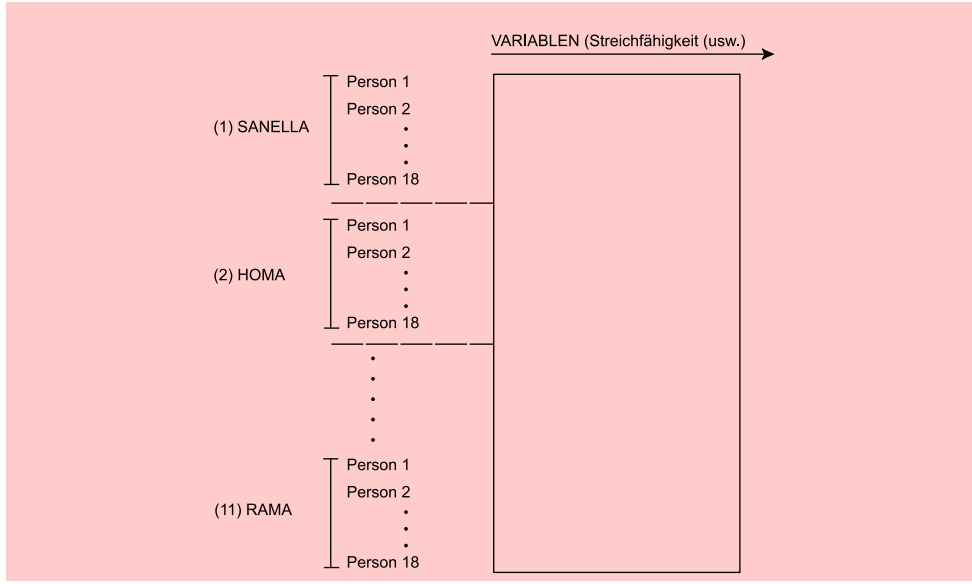


Abbildung 7.67: Die personenbezogene Objektmatrix

	Durchschnittsverfahren N = 11		Objektverfahren N = 127	
	FAKTOR 1	FAKTOR 2	FAKTOR 1	FAKTOR 2
STREICHF	−.77533	.35540	.29203	−.88490
PREIS	.15782	−.82054	.24729	−.00013
HALTBARK	−.43414	.29059	.50559	−.38055
UNGEFETT	−.13003	.80776	.1511	−.05290
BACKEIGN	.49826	−.15191	.58184	.11287
GESCHMAC	.71213	−.21740	.79836	.20807
KALORIEN	.86186	−.07113	.31326	.30129
TIERFETT	.98085	−.10660	.22904	.61220
VITAMIN	.51186	.30277	.62307	.03999
NATUR	.92891	−.15978	.53825	.36232

Abbildung 7.68: Die Faktorladungen im Vergleich

damit Positionen. Dadurch, dass sich bei den Analysen unterschiedliche Faktorenmuster ergeben, sind die Positionierungen in letzter Konsequenz nicht mehr vergleichbar.

7.4.1.3 Entdeckungs- oder Begründungszusammenhang: Explorative versus konfirmatorische Faktorenanalyse

Entdeckungszusammenhang

Bei einer Vielzahl wissenschaftlicher und praktischer Fragestellungen ist es von Interesse, Strukturen in einem empirischen Datensatz zu erkennen. Der Anwender hat keine konkreten Vorstellungen über den Zusammenhang zwischen Variablen, und es werden lediglich hypothetische Faktoren als verursachend für empirisch beobachtete

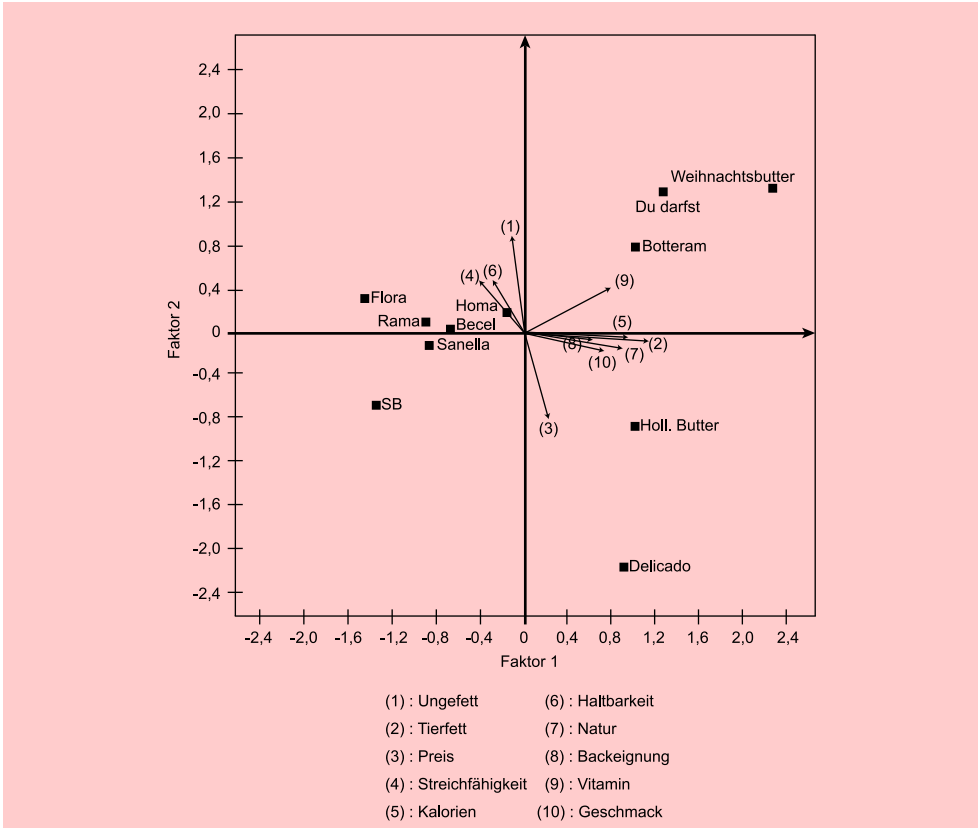


Abbildung 7.69: Die zweidimensionale Positionierung beim Durchschnittsverfahren

Korrelationen zwischen den Variablen angesehen, ohne dass der Anwender genaue Kenntnisse über diese Faktoren besitzt. In einer solchen Situation bietet die in diesem Kapitel beschriebene Faktorenanalyse ein geeignetes Analyseinstrumentarium zur Aufdeckung unbekannter Strukturen. Die Faktorenanalyse ist damit im Hinblick auf den methodologischen Standort in den *Entdeckungszusammenhang* einzuordnen. Sie kann deshalb auch als *Hypothesengenerierungsinstrument* bezeichnet werden, und wir sprechen in diesem Fall von einer *explorativen Faktorenanalyse (EFA)*.

Demgegenüber existieren bei vielen Anwendungsfällen aber bereits a priori konkrete Vorstellungen über mögliche hypothetische Faktoren, die hinter empirisch beobachteten Korrelationen zwischen Variablen zu vermuten sind. Aufgrund *theoretischer* Vorüberlegungen werden Hypothesen über die Beziehung zwischen direkt beobachtbaren Variablen und dahinter stehenden, nicht beobachtbaren Faktoren aufgestellt, und es ist von Interesse, diese Hypothesen an einem empirischen Datensatz zu prüfen. Hier kann die Faktorenanalyse zur *Hypothesenprüfung* herangezogen werden. Wir befinden uns damit im *Begründungszusammenhang*. In solchen Anwendungsfällen spricht man von einer *konfirmatorischen Faktorenanalyse (KFA)*. Die konfirmatorische Faktorenanalyse basiert ebenfalls auf dem Fundamentaltheorem der Faktorenanalyse. Die Anwendung einer solchen Faktorenanalyse setzt allerdings voraus, dass der Anwender die Beziehungen zwischen beobachteten Variablen und Faktoren aufgrund intensiver theoretischer Überlegungen *vor* Anwendung der Faktorenanalyse festlegt. Welche

Begründungszusammenhang

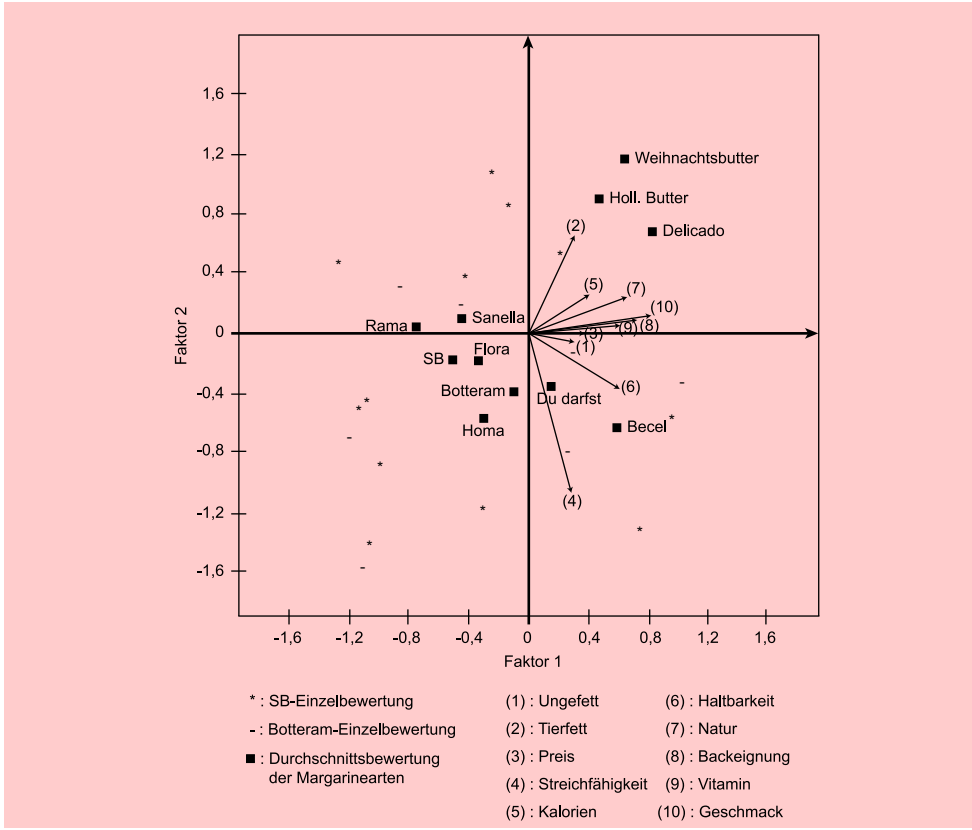


Abbildung 7.70: Die zweidimensionale Positionierung beim Objektverfahren

Auswirkungen daraus resultieren, wird in dem Buch Fortgeschrittene Multivariate Analysemethoden behandelt (Backhaus, Erichson, Weiber (2015), Fortgeschrittene Multivariate Analysemethoden, 3. Auflage, Springer 2015).

7.4.2 Empfehlungen zur Durchführung einer Faktorenanalyse

Die obigen Ausführungen haben gezeigt, dass eine Faktorenanalyse bei gleichen Ausgangsdaten zu unterschiedlichen Ergebnissen führen kann, je nachdem, wie die subjektiv festzulegenden Einflussgrößen „eingestellt“ werden. Gerade für denjenigen, der neu in diesem Gebiet tätig werden will, mögen einige Empfehlungen (Abbildung 7.71) für die vom Anwender subjektiv festzulegenden Größen eine erste Hilfestellung bedeuten. Die Vorschläge sind dabei daran orientiert, inwieweit sie sich bei der Fülle bereits durchgeführter Faktorenanalysen bewährt haben.

Empfehlungen

Abschließend sei nochmals betont, dass diese Empfehlungen lediglich an denjenigen gerichtet sind, der sich neu mit der Faktorenanalyse befasst. Die Leser, die tiefer in die Materie eindringen möchten, seien vor allem auf das Buch von Hair et al. verwiesen, das eine der eingängisten Quellen zur Faktorenanalyse ist. Hier finden sich weitere ins Detail gehende Erläuterungen und Empfehlungen.[19]

[19] Vgl. Hair et al. (2014).

Notwendige Schritte der Faktorenanalyse	Empfehlungen bzw. Voraussetzungen
1. Ausgangserhebung	– Daten müssen metrisch skaliert sein (mindestens Intervallskala). – Fallzahl sollte mindestens der dreifachen Variablenzahl entsprechen, mindestens größer 50 sein.
2. Erstellen der Ausgangsdatenmatrix	– Standardisierung der Variablen
3. Berechnung der Korrelationsmatrix	
4. Kommunalitätenschätzung und Faktorextraktion	– Entscheidung, ob eine Hauptkomponentenanalyse oder eine Faktorenanalyse durchgeführt werden soll. Wird ein Extraktionsverfahren der Faktorenanalyse gewählt, so sind die Schätzverfahren ,,Hauptachsen-Faktorenanalyse “ und ,,Maximum Likelihood “ die am häufigsten gewählten Schätzmethoden.
5. Bestimmung der Faktorenzahl	– Scree-Test
6. Rotation	– Varimax-Kriterium
7. Interpretation	– Höchstens Faktorladungen $> 0,5$ verwenden (Konvention)
8. Bestimmung der Faktorwerte	– Regressionsschätzung

Abbildung 7.71: Empfehlungen zur Faktoranalyse

7.5 Anhang: Mathematische Darstellung der Faktorextraktion

Für den mathematisch interessierten Leser soll nachfolgend die rechnerische Durchführung der Faktorenextraktion dargestellt werden. Wir beschränken uns dabei auf die Hauptkomponentenmethode.

Die Grundgleichung der Faktorenanalyse lässt sich wie folgt schreiben, wenn die Matrix der Ausgangsdaten $\mathbf{Z}$ transponiert wird, so dass die Variablen in den Zeilen und die Fälle in den Spalten aufgeführt werden:

$$\mathbf{Z} = \mathbf{A} \cdot \mathbf{P}$$

d. h. die Matrix der standardisierten Ausgangsdaten lässt sich zerlegen in das Muster der Faktorladungen $\mathbf{A}$ und der Faktorwerte $\mathbf{P}$. Wegen der Orthogonalität der Faktoren lässt sich diese Gleichung umformen in

$$\begin{aligned} \mathbf{P} &= \mathbf{A}^{-1} \cdot \mathbf{Z} \\ &= \mathbf{A}' \cdot \mathbf{Z} \\ \text{bzw.} \quad \mathbf{p_q} &= \sum_j a_{j_q} \cdot z_j = \mathbf{a'_q}\, \mathbf{z_j} \end{aligned}$$

Da jeder Faktor q eine Linearkombination der Variablen $z_j (j = 1, 2, ..., J)$ bildet, gilt für die Varianz eines Faktors q:

$$s_q^2 = \sum_{j=1}^{J} \sum_{l=1}^{J} a_{j_q} \cdot a_{l_q} \cdot r_{j_l} = \mathbf{a'_q}\, \mathbf{R}\, \mathbf{a}_q$$

Da die Korrelationsmatrix $\mathbf{R}$ vorgegeben ist, folgt: Die Varianz eines Faktors ist abhängig von seinen Ladungen. Da ein Faktor möglichst viel Varianz der Variablen aufnehmen soll, ergibt sich folgendes *Optimierungsproblem* zur Ermittlung der Faktoren:

$$\max_{\mathbf{a_q}}\{s_q^2\} \quad \text{unter der Nebenbedingung } \mathbf{a_q'a_q} = 1$$

(Reparametrisierungsbedingung)

Zur Lösung dieses Optimierungsproblems für Faktor 1 lässt sich der folgende Lagrange-Ansatz verwenden:

$$\begin{aligned} & \frac{\partial[s_1^2 + \lambda_1(1-\mathbf{a_1'}\,\mathbf{a_1})]}{\partial\,\mathbf{a_1}} \\ = \quad & \frac{\partial[\mathbf{a_1'}\,\mathbf{R}\,\mathbf{a_1}+\lambda_1(1-\mathbf{a_1'}\,\mathbf{a_1})}{\partial\,\mathbf{a_1}} \\ = \quad & 2\,(\mathbf{R}-\lambda_1\mathbf{E})\,\mathbf{a_1} = \mathbf{0} \end{aligned}$$

Die gesuchten Ladungen von Faktor 1 sind damit bestimmt durch das homogene Gleichungssystem

$$(\mathbf{R}-\lambda_1\mathbf{E})\cdot\mathbf{a_1} = \mathbf{0} \qquad \text{(Eigenwertproblem)}$$

mit: λ_1 = Eigenwert

$\mathbf{a_1}$ = Eigenvektor

Wegen $\mathbf{a_1'}\cdot\mathbf{a_1} = 1$

gilt $\lambda_1 = \mathbf{a_1'}\,\mathbf{R}\,\mathbf{a_1} = s_1^2$ Varianz von Faktor 1

Da s_1^2 maximiert werden soll, muss λ_1 der **größte Eigenwert von R** sein. Der zugehörige Eigenvektor liefert die Ladungen von Faktor 1. Analog liefert der zweitgrößte Eigenwert die Ladungen von Faktor 2, usw.

Allgemein gilt: Die Faktoren q $(q = 1, ..., Q)$ sind bestimmt durch die Eigenwerte der Korrelationsmatrix $\mathbf{R}$

$$\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_Q$$

Eigenwertprobleme

Die Lösung von Gleichungssystemen der Art

$$(\mathbf{R}-\lambda\mathbf{E})\,\mathbf{a} = \mathbf{0} \quad \text{mit:} \quad \lambda = \textit{Eigenwert}$$

$\mathbf{a}$ = Eigenvektor

$$\text{und} \qquad (\mathbf{R}-\lambda\mathbf{E}) = \begin{bmatrix} r_{11}^{-\lambda} & r_{12} & \cdots & r_{1J} \\ r_{21} & r_{22}^{-\lambda} & \cdots & r_{2J} \\ \vdots & \vdots & \ddots & \vdots \\ r_{J1} & r_{J2} & \cdots & r_{JJ}^{-\lambda} \end{bmatrix}$$

nennt man ein Eigenwertproblem. Es gilt:

- Zu jedem Eigenwert $\lambda_q \neq 0$ existiert ein Eigenvektor $\mathbf{a}_q$
- Die Anzahl der Q der von 0 verschiedenen Eigenwerte von $\mathbf{R}$ ist begrenzt durch den Rang von $\mathbf{R}$ (Ist $\mathbf{R}$ symmetrisch, so gilt $R_g\,(\mathbf{R}) = Q$).

Berechnung der Eigenwerte

Da $(\mathbf{R}-\lambda\mathbf{E})\mathbf{a} = \mathbf{0}$ ein **homogenes Gleichungssystem** bildet, kann es nur für solche Werte von λ eine nichttriviale Lösung $\mathbf{a}$ besitzen, für die die Koeffizientendeterminante

$$|\mathbf{R} - \lambda\mathbf{E}|_{det} = 0$$

ist. Damit ist ein Weg zur Berechnung der Eigenwerte der Matrix $\mathbf{R}$ aufgezeigt.

Die Berechnung der Determinante von $(\mathbf{R} - \lambda\mathbf{E})$ führt zur **charakteristischen Gleichung**

$$|\mathbf{R} - \lambda\mathbf{E}|_{det} = c_Q\lambda^Q + c_{Q-1}\lambda^{Q-1} + ... + c_1\lambda + c_0 = 0$$

Als **Nullstellen** dieses Polynoms erhält man die Eigenwerte $\lambda_1, ..., \lambda_Q$ der Korrelationsmatrix $\mathbf{R}$.

Beispiel: $\mathbf{R} = \begin{bmatrix} 4 & 6 \\ 2 & 3 \end{bmatrix}$

$$\begin{aligned} \begin{vmatrix} 4-\lambda & 6 \\ 2 & 3-\lambda \end{vmatrix}_{det} &= (4-\lambda)(3-\lambda) - 6 \cdot 2 \\ &= \lambda^2 - 7 \cdot \lambda \\ &= \lambda(\lambda - 7) = 0 \end{aligned}$$

$$\rightarrow \lambda_1 = 7, \lambda_2 = 0$$

Normierung der Faktorladungen

Nach der Extraktion sind die Faktorladungen geeignet zu normieren:

$$a_{jq} := a_{jq} \cdot \frac{\sqrt{\lambda_q}}{\sqrt{a_{1_q}^2 + a_{2_q}^2 + ... + a_{J_q}^2}}$$

Damit gilt:

$$\sum_{j=1}^{J} a_{jq}^2 = \lambda_q \qquad \text{Durch Faktor } q \text{ erklärte Varianz}$$

$$\sum_{q=1}^{Q} a_{jq}^2 = h_j^2 \qquad \text{Durch die Faktoren erklärte Varianz von Variable } j$$

Die Summation der Ladungsquadrate

- eines Faktors q über alle Variablen ergibt somit die durch den Faktor erklärte Varianz (Eigenwert),
- aller Faktoren für eine Variable ergibt deren erklärte Varianz (Kommunalität).

Literaturhinweise

A. Basisliteratur zur Faktorenanalyse

Child, D. (2006), The Essentials of Factor Analysis, 3. Auflage, London u. a.

Hair, J./Black, W./Babin, B./Anderson, R. (2014), Multivariate Data Analysis, 7. Auflage, Englewood Cliffs (N.J.), Kapitel 3.

Hüttner, M./Schwarting U. (2007), Exploratorische Faktorenanalyse, in: Herrmann, A./Homburg C./Klarmann, A.: Handbuch Marktforschung – Methoden, Anwendungen und Praxisbeispiele, 3. Auflage, Wiesbaden, S. 381–412.

Kim, J./Mueller, C. (1978), Introduction to Factor Analysis, Series Number 07-013, Beverly Hills u. a.

Revensdorf, D. (1976), Lehrbuch der Faktorenanalyse, Stuttgart.

Schlittgen, R. (2009), Multivariate Statistik, München.

Überla, K. (1977), Faktorenanalyse, 2. Auflage, Berlin u. a.

B. Zitierte Literatur

Backhaus, K./Blechschmidt, B. (2009), Fehlende Werte und Datenqualitat, in: *Die Betriebswirtschaft*, Vol. 69, Nr. 2, S. 265.

Carroll, J. (2004), Human Cognitive Abilities – A survey of factor-analytic studies, Cambridge.

Cureton, E./D'Agostino, R. (1993), Factor Analysis – An Applied Approach, Hillsdale (N.J.).

Dziuban, C./Shirkey, E. (1974), When is a Correlation Matrix Appropriate for Factor Analysis? in: *Psychological Bulletin*, Vol. 81, Nr. 6, S. 358–361.

Gellert, W./Küstner, H./Hellwich, M./Kästner, H. (1977), Mathematik, Leipzig.

Guttmann, L. (1953), Image Theory for the Structure of Quantitative Vanates, in: *Psychometrika*, Vol. 18, Nr. 4, S. 277–296.

Hair, J./Black, W./Babin, B./Anderson, R. (2014), Multivariate Data Analysis, 7. Auflage, Englewood Cliffs (N.J.).

Harmann, H. (1976), Modern Factor Analysis, 3. Auflage, Chicago.

Kaiser, H. (1970), A Second Generation Little Jiffy, in: *Psychometrika*, Vol. 35, Nr. 4, S. 401–415.

Kaiser, H./Rice, J. (1974), Little Jiffy, Mark IV, in: *Educational and Psychological Measurement*, Vol. 34, Nr. 1, S. 111–117.

Krolak-Schwerdt, S. (1991), Modelle der dreimodalen Faktorenanalyse, Frankfurt am Main.

Litfin, T./Teichmann, M.-H./Clement, M. (2000), Beurteilung der Güte von Explorativen Faktoranalysen im Marketing, in: *Wirtschaftswissenschaftliches Studium*, Vol. 29, Nr. 5, S. 283–286.

Loehlin, J. (2004), Latent variable models: factor, path, and structural equation analysis, 4. Auflage, New Jersey.

Plinke, W. (1985), Erlösplanung im industriellen Anlagengeschäft, Wiesbaden.

Stewart, D. (1981), The Application and Misapplication of Factor Analysis in Marketing Research, in: *Journal of Marketing Research*, Vol. 18, Nr. 1, S. 51–62.

Überla, K. (1977), Faktorenanalyse, 2. Auflage, Berlin u. a.

Weiber, R./Mühlhaus, D. (2014), Strukturgleichungsmodellierung, 2. Auflage, Berlin Heidelberg.

8 Clusteranalyse

K. Backhaus, B. Erichson, W. Plinke, R. Weiber, *Multivariate Analysemethoden*,
https://doi.org/10.1007/978-3-662-56655-8_9

8.1 Problemstellung

Bei vielen Anwendungen z. B. in der Medizin, der Archäologie, der Soziologie, der Linguistik, der Biologie oder den Wirtschaftswissenschaften ist die Frage von Bedeutung, ob zwischen den betrachteten Untersuchungsobjekten (z. B. Personen, Unternehmen, Produkte, Pflanzen, Biokulturen) Ähnlichkeiten bestehen. Ziel ist es dabei häufig, solche Untersuchungsobjekte zu Gruppen (Cluster) zusammenzufassen, die im Hinblick auf die betrachteten Eigenschaften oder Merkmale als möglichst homogen zu bezeichnen sind. Gleichzeitig sollten die Gruppen untereinander eine möglichst große Heterogenität aufweisen, d. h. möglichst unähnlich sein. Als Beispiel seien hier die 20.000 eingeschriebenen Studierenden einer Universität als Fälle (Untersuchungsobjekte) genannt, von denen das Geschlecht, das Studienfach, die Semesterzahl, der Studienwohnort, die Nationalität und der Familienstand als Eigenschaften (Merkmalsvariable) erhoben wurden. Ausgehend von diesen Daten möchte der Forscher nun wissen, ob sich die Erhebungsgesamtheit nach Studierendengruppen aufteilen lässt, wobei die Studierenden innerhalb einer Gruppe eine möglichst hohe Ähnlichkeit aufweisen sollten, während zwischen den Studierendengruppen (so gut wie) keine Ähnlichkeiten bestehen sollten.

Unter dem Begriff Clusteranalyse werden unterschiedliche Verfahren zur Gruppenbildung zusammengefasst, die sich vor allem im Hinblick auf folgende zwei Aspekte unterscheiden:

Proximitätsmaße

1. Wahl des Proximitätsmaßes, d. h. das statistische Maß, mit dem die Ähnlichkeit bzw. Unähnlichkeit (Distanzmaße) zwischen Objekten gemessen wird.

Gruppierungsverfahren

2. Wahl des Gruppierungsverfahrens, d. h der Vorgehensweise, nach der eine Zusammenfassung von ähnlichen Objekten zu Gruppen (Fusionierungsalgorithmen) oder aber die Zerlegung einer Erhebungsgesamtheit in Gruppen (Partitionierungsalgorithmem) erfolgen soll.

Es ist ein wesentliches Charakteristikum von Clusterverfahren, dass zur Gruppenbildung *alle* vorliegenden Eigenschaften der Untersuchungsobjekte *gleichzeitig* zur Gruppenbildung herangezogen werden. Weiterhin zählen Clusterungsmethoden zu den explorativen Verfahren der multivariaten Datenanalyse, da dem Forscher die Gruppen im Ausgangspunkt *unbekannt* sind und er mit Hilfe eines Clusterverfahrens erst eine solche Gruppierung herbeiführt, d. h. Objektgruppen in der Erhebungsgesamtheit identifiziert. In Abbildung 8.1 sind einige *Anwendungsbeispiele* der Clusteranalyse aus dem Bereich der Wirtschaftswissenschaften zusammengestellt. Sie vermitteln einen Einblick in die Problemstellung, die Zahl und Art der Merkmale, die Zahl und Art der Untersuchungseinheiten sowie die ermittelte Gruppenzahl.

Problemstellung	Zahl und Art der Merkmale	Zahl der Untersuchungs-einheiten	Ermittelte Gruppenzahl
Segmentierung von Internetusern[1]	4 Merkmale zur Interneterfahrung, z. B.: Selbsteinschätzung der User	86 Studierende	4
Segmentierung von Bedürfnistypen[2]	5 bzgl. Wichtigkeit beurteilte Unterstützungsleistungen im Kundenprozess	1.055 Privatpersonen	3
Bildung von Innovationstypen[3]	5 Merkmale zur Innovationsleistung von Unternehmen; 3 Merkmale zur Innovationskompetenz	115 Unternehmen	2
Segmentierung von Kunden im Einzelhandel[4]	4 grundlegende Einkaufsmotive: Leistungsumfangs-, Qualitäts-, Preis- und Zeit-Orientierung	2.000 Konsumenten	4
Gruppierung von spezialisierten Marktfruchtunternehmen[5]	3 Merkmale: Betriebsaufwand, Reinertrag, Naturalertrag	24 landwirtschaftliche Betriebe	8

Abbildung 8.1: Anwendungsbeispiele der Clusteranalyse

8.2 Vorgehensweise

Ablaufschritte der Clusteranalyse

Die Ablaufschritte einer Clusteranalyse werden teilweise durch das gewählte Clusterverfahren bestimmt, wobei hier zwischen partitionierenden und hierarchischen Verfahren unterschieden werden kann (vgl. 8.2.2). Da letztere Verfahrensgruppe bei praktischen Anwendungen die größte Verbreitung gefunden hat, steht diese auch im Vordergrund der nachfolgenden Betrachtungen. Es lassen sich folgende grundlegende Ablaufschritte unterscheiden:

Abbildung 8.2: Ablaufschritte der hierarchischen Clusteranalyse

1. Schritt: *Bestimmung der Ähnlichkeiten bzw. Distanzen*
Für jeweils zwei Personen bzw. Objekte werden die Ausprägungen der Beschreibungsmerkmale geprüft und die Unterschiede bzw. Übereinstimmungen durch einen Zahlenwert (Proximitätsmaß) gemessen.

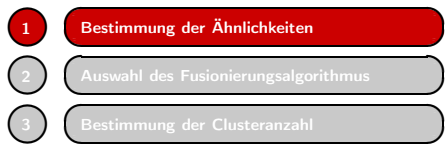
[1] Vgl. Meyer (2004), S. 188 ff.
[2] Weiber/Fälsch (2007), S. 97 ff.
[3] Vgl. Wiesener (2014), S. 285 ff.
[4] Vgl. Swoboda/Hälsig/Morschett (2007), S.19 ff.
[5] Vgl. Herink/Petersen (2004), S. 290 ff.

2. Schritt: *Auswahl des Fusionierungsalgorithmus*
Aufgrund der Ähnlichkeits- bzw. Distanzwerte werden die Fälle so zu Gruppen zusammengefasst, dass sich diejenigen Objekte oder Personen mit weitgehend übereinstimmend ausgeprägten Beschreibungsmerkmalen in einer Gruppe wiederfinden. Entsprechend der Vorschriften des Fusionierungsalgorithmus fasst die (agglomerative) Clusteranalyse die betrachteten Fälle solange zusammen, bis am Ende alle Fälle in einer einzigen Gruppe enthalten sind.

3. Schritt: *Bestimmung der optimalen Clusterzahl*
Anschließend ist zu entscheiden, welche Anzahl an Clustern die „beste“ Lösung darstellt und im Ergebnis verwendet werden soll. Hier gilt es vor allem den Zielkonflikt zwischen Handhabbarkeit (geringe Clusterzahl) und Homogenitätsanforderung (große Clusterzahl) zu lösen.

Diesen Schritten entsprechend sind die nachfolgenden Überlegungen aufgebaut. Da unter dem Begriff Clusteranalyse sehr unterschiedliche Verfahren zur Gruppenbildung zusammengefasst sind, wird im Rahmen der obigen Schritte nicht nur ein Verfahren der Clusteranalyse behandelt, sondern es werden jeweils unterschiedliche Verfahrensvarianten zur Bestimmung der Ähnlichkeiten sowie zur Fusionierung von Fällen besprochen.

8.2.1 Bestimmung der Ähnlichkeiten



Den Ausgangspunkt der Clusteranalyse bildet eine *Rohdatenmatrix* mit K Objekten (z. B. Personen, Unternehmen, Produkte), die durch J Variable beschrieben werden und deren allgemeiner Aufbau Abbildung 8.3 verdeutlicht.

Rohdatenmatrix

	Variable 1	Variable 2	...	Variable J
Objekt 1				
Objekt 2				
-				
-				
-				
Objekt K				

Abbildung 8.3: Aufbau der Rohdatenmatrix

Im Inneren dieser Matrix stehen die objektbezogenen metrischen und/oder nicht metrischen Variablenwerte. Im ersten Schritt geht es zunächst um die *Quantifizierung der Ähnlichkeit* zwischen den Objekten durch eine statistische Maßzahl. Zu diesem Zweck wird die Rohdatenmatrix in eine *Distanz- oder Ähnlichkeitsmatrix* (Abbildung 8.4) überführt, die immer eine quadratische (KxK-)Matrix darstellt.

Distanz- oder Ähnlichkeitsmatrix

Diese Matrix enthält die Ähnlichkeits- oder Unähnlichkeitswerte (Distanzwerte) zwischen den betrachteten Objekten, die unter Verwendung der objektbezogenen Variablenwerte aus der Rohdatenmatrix berechnet werden. Maße, die eine Quantifizie-

	Objekt 1	Objekt 2	...	Objekt K
Objekt 1				
Objekt 2				
-				
-				
-				
Objekt K				

Abbildung 8.4: Aufbau einer Distanz- oder Ähnlichkeitsmatrix

Proximitätsmaße

rung der Ähnlichkeit oder Distanz zwischen den Objekten ermöglichen, werden allgemein als *Proximitätsmaße* bezeichnet. Es lassen sich zwei Arten von Proximitätsmaßen unterscheiden:

Ähnlichkeitsmaße

- *Ähnlichkeitsmaße* spiegeln die Ähnlichkeit zwischen zwei Objekten wider: Je größer der Wert eines Ähnlichkeitsmaßes wird, desto ähnlicher sind sich zwei Objekte.

Distanzmaße

- *Distanzmaße* messen die Unähnlichkeit zwischen zwei Objekten: Je größer die Distanz wird, desto unähnlicher sind sich zwei Objekte. Sind zwei Objekte als vollkommen identisch anzusehen, so ergibt sich eine Distanz von Null.

In Abhängigkeit des Skalenniveaus der betrachteten Merkmale existiert eine Vielzahl an Proximitätsmaßen. Diese werden in SPSS nach Proximitätsmaßen für Variable mit metrischem Skalenniveau (Intervall), mit nominalem Skalenniveau (Häufigkeitsdaten) und mit binärer Ausprägung (0/1-Variable) unterschieden. Abbildung 8.5 zeigt eine Auswahl der in SPSS verfügbaren Proximitätsmaße, von denen im Folgenden je Skalenniveau-Kategorie drei gebräuchliche Maße zur Bestimmung der Ähnlichkeit bzw. Distanz im Detail diskutiert werden.

8.2.1.1 Ähnlichkeitsermittlung bei binärer Variablenstruktur

Eine binäre Variablenstruktur liegt vor, wenn alle Merkmalsvariable nur die Ausprägungen Null und Eins annehmen, wobei der Wert 1 i. d. R. „Eigenschaft vorhanden“ bedeutet und der Wert 0 für „Eigenschaft nicht vorhanden“ verwendet wird. Bei der Ermittlung der Ähnlichkeit zwischen zwei Objekten geht die Clusteranalyse immer von einem Paarvergleich aus, d. h. für jeweils zwei Objekte werden alle Eigenschaftsausprägungen miteinander verglichen. Wie Abbildung 8.6 zu entnehmen ist, lassen sich im Fall binärer Merkmale beim Vergleich zweier Objekte bezüglich einer Eigenschaft vier Fälle unterscheiden:

Kombinationsmöglichkeiten binärer Variable

- bei beiden Objekten ist die Eigenschaft vorhanden (Feld a)
- nur Objekt 2 weist die Eigenschaft auf (Feld b)
- nur Objekt 1 weist die Eigenschaft auf (Feld c)
- bei beiden Objekten ist die Eigenschaft nicht vorhanden (Feld d)

Für die Ermittlung von Ähnlichkeiten zwischen Objekten mit binärer Variablenstruktur ist in der Literatur eine Vielzahl von Maßzahlen entwickelt worden, die sich größtenteils auf folgende allgemeine Ähnlichkeitsfunktion zurückführen lassen:[6]

$$S_{ij} = \frac{a + \delta \cdot d}{a + \delta \cdot d + \lambda(b + c)} \tag{8.1}$$

[6] Vgl. Kaufman/Rousseeuw (2008), S. 24 ff.

	Skalenniveau der Merkmalsvariablen		
	metrisch (Intervall)	nominal (Häufigkeiten)	binär (0/1)
Ähnlichkeitsmaße	• Kosinus • Pearson-Korrelation	in SPSS nicht verfügbar, zu berücksichtigen durch • Tranformation in binäre Variable oder • Analyse von Häufigkeitsdaten	• Würfelmaß (Dice- oder Czekanowski-Koeffiz.) • Jaccard-Koeffizient • M-Koeffizient (Einfache Übereinstimmung) • Kulczynski-Koeffizient • Rogers und Tanimoto • Russel & Rao (RR) Koeffizient
Distanzmaße	• (Quadrierte) Euklidische Distanz • Minkowski Metrik • Block-Metrik • Tschebyscheff-Metrik	• Chi-Quadrat-Maß • Phi-Quadrat-Maß	• Binäre Euklidische Distanz • Lance-Williams-Maß • Binäre Form-Differenz • Größendifferenz • Varianz

Abbildung 8.5: Ausgewählte Proximitätsmaße für die hierarchische Clusteranalyse in SPSS

	Objekt 2		Zeilensumme
Objekt 1	Eigenschaft vorhanden (1)	Eigenschaft nicht vorhanden (0)	
Eigenschaft vorhanden (1)	a	c	**a+c**
Eigenschaft nicht vorhanden (0)	b	d	**b+d**
Spaltensumme	**a+b**	**c+d**	**M**

Abbildung 8.6: Kombinationsmöglichkeiten von binären Variablen

mit

$$\begin{aligned} S_{ij} &= \text{Ähnlichkeit zwischen den Objekten i und j} \\ \delta, \lambda &= \text{mögliche (konstante) Gewichtungsfaktoren} \end{aligned}$$

Dabei entsprechen die Variablen a, b, c und d den Kennungen in Abbildung 8.6, wobei z. B. die Variable a der Anzahl der Eigenschaften entspricht, die bei beiden Objekten (1 und 2) vorhanden ist. Je nach Wahl der Gewichtungsfaktoren δ und λ erhält man unterschiedliche Ähnlichkeitsmaße für Objekte mit binären Variablen. Abbildung 8.7 gibt darüber einen Überblick.[7]

[7] Eine Darstellung weiterer Ähnlichkeitskoeffizienten liefern u. a. Bacher/Pöge/Wenzig (2010), S. 195 ff., Handl (2010), S. 83 ff.

Name des Ähnlichkeitskoeffizienten	Gewichtungsfaktoren		Definition
	δ	λ	
Jaccard	0	1	$\frac{a}{a+b+c}$
M-Koeffizient	1	1	$\frac{a+d}{M}$
Russel & Rao (RR)	-	-	$\frac{a}{M}$
Dice	0	$\frac{1}{2}$	$\frac{2a}{2a+(b+c)}$
Kulczynski	-	-	$\frac{a}{b+c}$

Abbildung 8.7: Definition ausgewählter Ähnlichkeitsmaße bei binären Variablen

Ähnlichkeitskoeffizienten

Im Folgenden werden die *Ähnlichkeitskoeffizienten nach Jaccard, Russel & Rao* sowie der *M-Koeffizient* näher betrachtet, die bei praktischen Anwendungen im Fall binärer Merkmale häufig zur Anwendung kommen.

Zu diesem Zweck betrachten wir das in Abbildung 8.8 dargestellte Beispiel, das elf Butter- und Margarinemarken mit jeweils zehn binären Eigenschaften enthält. Bezüglich der Merkmale wird angegeben, ob ein Produkt die jeweilige Eigenschaft aufweist (1) oder nicht (0).

Eigenschaften / Emulsionsfette	Lagerzeit mehr als 1 Monat	Diätprodukt	Nationale Werbung	Becherverpackung	Pfundgröße	Verkaufshilfen	Eignung für Sonderangebote	Direktbezug vom Hersteller	Handelsspanne mehr als 20 %	Beanstandungen im letzten Jahr
Becel	1	1	1	1	0	0	1	0	0	0
Du darfst	1	1	0	1	0	1	0	1	0	1
Rama	1	0	1	1	1	1	1	1	1	0
Delicado Sahnebutter	0	0	1	1	0	0	1	0	1	0
Holländische Butter	0	0	0	0	0	1	0	0	0	0
Weihnachtsbutter	0	0	0	0	1	0	1	0	0	1
Homa	1	0	0	1	1	1	0	1	0	1
Flora	1	1	1	1	1	0	1	0	1	0
SB	1	1	0	1	1	1	0	0	1	0
Sanella	1	0	1	1	1	0	1	1	1	0
Botteram	0	0	1	1	1	1	0	0	0	1

Abbildung 8.8: Ausgangsdatenmatrix zur Darstellung von Ähnlichkeitskoeffizienten bei binärer Variablenstruktur

8.2.1.1.1 Jaccard-Koeffizient

Jaccard-Koeffizient

Der *Jaccard-Koeffizient* misst den relativen Anteil gemeinsamer Eigenschaften bezogen auf die Variablen, die mindestens eine 1 aufweisen. Zunächst wird festgestellt, wie viele Eigenschaften beide Produkte übereinstimmend aufweisen. In unserem Beispiel sind dies bei den Margarinemarken „Becel“ und „Du darfst“ drei Merkmale („Lagerzeit mehr als 1 Monat“, „Diätprodukt“ und „Becherverpackung“). Anschließend werden die Eigenschaften gezählt, die lediglich bei einem Produkt vorhanden sind. In unserem Beispiel lassen sich fünf Attribute finden („Nationale Werbung“, „Verkaufshilfen“, „Eignung für Sonderangebote“, „Direktbezug vom Hersteller“ und „Beanstandungen im letzten Jahr“). Setzt man die Anzahl der Eigenschaften, die bei beiden Produkten vorhanden sind, in den Zähler (a=3) und addiert hierzu für den Nenner die Anzahl der Eigenschaften, die nur bei einem Produkt vorhanden sind (b+c=5), so beträgt der Jaccard-Koeffizient für die Produkte „Becel“ und „Du darfst“ $3/8 = 0{,}375$.

Auf dem gleichen Weg werden für alle anderen Objektpaare die entsprechenden Ähnlichkeiten berechnet. Abbildung 8.9 gibt die Ergebnisse wieder. Bezüglich der dargestellten Matrix ist auf zwei Dinge hinzuweisen:

- Die Ähnlichkeit zweier Objekte wird nicht durch ihre Reihenfolge beim Vergleich beeinflusst, d. h. es ist unerheblich, ob die Ähnlichkeit zwischen „Becel“ und „Du darfst“ oder zwischen „Du darfst“ und „Becel“ gemessen wird (Symmetrie-Eigenschaft). Damit ist auch zu erklären, dass die Ähnlichkeit der Produkte in Abbildung 8.9 nur durch die untere Dreiecksmatrix wiedergegeben wird.

Symmetrie-Eigenschaft

	Becel	Du darfst	Rama	Delicado Sahnebutter	Holländische Butter	Weihnachts-butter	Homa	Flora	SB	Sanella	Botteram
Becel	1										
Du darfst	0,375	1									
Rama	0,444	0,4	1								
Delicado Sahnebutter	0,5	0,111	0,5	1							
Holländische Butter	0	0,167	0,125	0	1						
Weihnachtsbutter	0,143	0,125	0,222	0,167	0	1					
Homa	0,222	0,714	0,556	0,111	0,167	0,286	1				
Flora	0,714	0,3	0,667	0,571	0	0,25	0,3	1			
SB	0,375	0,5	0,556	0,25	0,167	0,125	0,5	0,625	1		
Sanella	0,5	0,3	0,875	0,571	0	0,25	0,444	0,75	0,444	1	
Botteram	0,25	0,375	0,444	0,286	0,2	0,333	0,571	0,333	0,375	0,333	1

Abbildung 8.9: Ähnlichkeiten nach dem Jaccard-Koeffizient

- Die Werte der Ähnlichkeitsmessung liegen zwischen 0 („totale Unähnlichkeit“, a=0) und 1 („totale Ähnlichkeit“, b=c=0). Wird die Übereinstimmung der Merkmale bei einem einzigen Produkt geprüft, so gelangt man zum Ergebnis der vollständigen Übereinstimmung. Somit ist auch verständlich, dass man in der Diagonalen der Matrix lediglich die Zahl 1 vorfindet.

Vollständige Übereinstimmung

Die Erläuterungen versetzen uns nunmehr in die Lage, das ähnlichste und das unähnlichste Paar zu ermitteln. Die größte Übereinstimmung weisen die Margarine-

sorten „Rama“ und „Sanella“ auf (Jaccard-Koeffizient=0,875). Als völlig unähnlich werden fünf Paare bezeichnet: „Holländische Butter“ – „Becel“, „Holländische Butter“ – „Delicado Sahnebutter“, „Weihnachtsbutter“ – „Holländische Butter“, „Flora“ – „Holländische Butter“ und „Sanella“ – „Holländische Butter“ (Jaccard-Koeffizient=0, da a=0).

8.2.1.1.2 Russel und Rao-Koeffizient

Russel&Rao-Koeffizient

Auf eine etwas andere Art und Weise wird die Ähnlichkeit der Objektpaare beim Koeffizienten von *Russel und Rao* (RR-Koeffizienten) gemessen. Der Unterschied zum Jaccard-Koeffizienten besteht darin, dass nunmehr im Nenner auch die Fälle, bei denen beide Objekte das Merkmal nicht aufweisen (d), mit aufgenommen werden. Somit finden sich alle in der jeweiligen Untersuchung berücksichtigten Eigenschaften im Nenner des Ähnlichkeitsmaßes wieder. Abgesehen von den Extremwerten (0 und 1) ergeben sich in unserem Beispiel nur „Zehntel-Brüche“ als RR-Koeffizient, da das Beispiel insgesamt 10 Eigenschaften je Objekt betrachtet und deshalb der Nenner des RR-Koeffizienten gleich 10 ist. Existiert beim Paarvergleich der Fall, dass wenigstens eine Eigenschaft bei beiden Objekten nicht vorhanden ist, so weist der RR-Koeffizient einen kleineren Ähnlichkeitswert auf als der Jaccard-Koeffizient. Dieser Fall ist beim Produktpaar „Becel“ – „Du darfst“ zu verzeichnen. Beide Margarinemarken weisen nicht die Eigenschaften „Pfundgröße“ und „Handelsspanne mehr als 20 %“ auf. Somit „sinkt“ ihr Ähnlichkeitswert im Vergleich zum Jaccard-Koeffizienten von 0,375 nun auf 0,3. Fehlt keine Eigenschaft (d=0), gelangen beide Ähnlichkeitsmaße zum gleichen Ergebnis. Die einzelnen Werte für den RR-Koeffizienten enthält Abbildung 8.10. Dabei ist zu beachten, dass auf der Hauptdiagonalen der Ähnlichkeitsmatrix nach Russel & Rao von SPSS nicht die „1“, sondern der Anteil der je Untersuchungsobjekt vorhandenen Merkmale (Kodierung mit 1) ausgewiesen wird. Die Werte der Hauptdiagonalen lassen sich mit Hilfe der Ausgangsdatenmatrix in Abbildung 8.8 leicht nachvollziehen.

	Becel	Du darfst	Rama	Delicado Sahnebutter	Holländische Butter	Weihnachts-butter	Homa	Flora	SB	Sanella	Botteram
Becel	0,500	0,300	0,400	0,300	0,000	0,100	0,200	0,500	0,300	0,400	0,200
Du darfst	0,300	0,600	0,400	0,100	0,100	0,100	0,500	0,300	0,400	0,300	0,300
Rama	0,400	0,400	0,800	0,400	0,100	0,200	0,500	0,600	0,500	0,700	0,400
Delicado Sahnebutter	0,300	0,100	0,400	0,400	0,000	0,100	0,100	0,400	0,200	0,400	0,200
Holländische Butter	0,000	0,100	0,100	0,000	0,100	0,000	0,100	0,000	0,100	0,000	0,100
Weihnachtsbutter	0,100	0,100	0,200	0,100	0,000	0,300	0,200	0,200	0,100	0,200	0,200
Homa	0,200	0,500	0,500	0,100	0,100	0,200	0,600	0,300	0,400	0,400	0,400
Flora	0,500	0,300	0,600	0,400	0,000	0,200	0,300	0,700	0,500	0,600	0,300
SB	0,300	0,400	0,500	0,200	0,100	0,100	0,400	0,500	0,600	0,400	0,300
Sanella	0,400	0,300	0,700	0,400	0,000	0,200	0,400	0,600	0,400	0,700	0,300
Botteram	0,200	0,300	0,400	0,200	0,100	0,200	0,400	0,300	0,300	0,300	0,500

Dies ist eine Ähnlichkeitsmatrix

Abbildung 8.10: Ähnlichkeitskoeffizient nach Russel & Rao (RR-Koeffizient)

8.2.1.1.3 M-Koeffizient

Simple-Matching-Koeffizient (Einfache Übereinstimmung)

Beim *M-Koeffizienten* (auch „Einfache Übereinstimmung“ oder „Simple-Matching-Koeffizient“ genannt) werden im Zähler *alle* übereinstimmenden Komponenten erfasst. Zu den bereits oben genannten Merkmalen kommen daher beim Vergleich von „Becel“ und „Du darfst“ noch die beiden Eigenschaften „Pfundgröße“ und „Handelsspanne mehr als 20 %“ hinzu. Die Ähnlichkeit, die sich entsprechend des Bruchs ($\frac{a+d}{M}$) berechnet, hat für das genannte Produktpaar folglich einen Wert von 0,5. Die Werte für die anderen Objektpaare kann man Abbildung 8.11 entnehmen.

	Becel	Du darfst	Rama	Delicado Sahnebutter	Holländische Butter	Weihnachts-butter	Homa	Flora	SB	Sanella	Botteram
Becel	1										
Du darfst	0,5	1									
Rama	0,5	0,4	1								
Delicado Sahnebutter	0,7	0,2	0,6	1							
Holländische Butter	0,4	0,5	0,3	0,5	1						
Weihnachtsbutter	0,4	0,3	0,3	0,5	0.6	1					
Homa	0,3	0,8	0,6	0,2	0,5	0,5	1				
Flora	0,8	0,3	0,7	0,7	0,2	0,4	0,3	1			
SB	0,5	0,6	0,6	0,4	0,5	0,3	0,6	0,7	1		
Sanella	0,6	0,3	0,9	0,7	0,2	0,4	0,5	0,8	0,5	1	
Botteram	0,4	0,5	0,5	0,5	0,6	0,6	0,7	0,4	0,5	0,4	1

Abbildung 8.11: Simple-Matching- oder M-Koeffizient (Einfache Übereinstimmung)

8.2.1.1.4 Vergleich der drei Ähnlichkeitskoeffizienten für binäre Merkmalsvariable

Vergleich von Ähnlichkeitskoeffizienten

Alle drei dargestellten Ähnlichkeitsmaße gelangen zum gleichen Ergebnis, wenn *keine* Eigenschaft beim Paarvergleich gleichzeitig fehlt, d. h. wenn d=0 ist. Ist dies jedoch nicht gegeben, so weist grundsätzlich der RR-Koeffizient den geringsten und der M-Koeffizient den höchsten Ähnlichkeitswert auf. Eine Mittelposition nimmt das Jaccard-Ähnlichkeitsmaß ein. Jaccard- und M-Koeffizient kommen jedoch dann zum gleichen Ergebnis, wenn lediglich die Fälle (a) und (d) existieren, d. h. nur ein gleichzeitiges Vorhandensein bzw. Fehlen von Eigenschaften beim Paarvergleich zu verzeichnen ist.

An dieser Stelle kann nicht ausführlich auf alle *Unterschiede der Ähnlichkeitsrangfolge* in unserem Beispiel eingegangen werden, die sich aufgrund der drei vorgestellten Koeffizienten ergeben. Es sei jedoch kurz auf einige Differenzen hingewiesen:

- Die Objektpaare „SB“ und „Rama“ bzw. „Homa“ und „Rama“ belegen z. B. beim RR-Koeffizienten den dritten Rang in der Ähnlichkeitsreihenfolge. Bei den beiden anderen Ähnlichkeitsmaßen sind die Produkte nicht unter den ersten neun ähnlichsten Paaren zu finden.
- Während „Weihnachtsbutter“ und „Holländische Butter“ nach dem Jaccard- und RR-Koeffizienten keinerlei Ähnlichkeit aufweisen, beläuft sich ihr Ähnlichkeitswert nach dem M-Koeffizienten auf 0,6.

Welches Ähnlichkeitsmaß im Rahmen einer empirischen Analyse vorzuziehen ist, lässt sich nicht allgemeingültig sagen. Eine große Bedeutung bei dieser nur im Einzelfall zu treffenden Entscheidung hat die Frage, ob das Nichtvorhandensein eines Merkmals für die Problemstellung die gleiche Bedeutung bzw. Aussagekraft besitzt wie das Vorhandensein der Eigenschaft. Machen wir uns diesen Sachverhalt am Beispiel der eingangs erwähnten Studenten-Untersuchung klar. Beim Merkmal „Geschlecht“ kommt z. B. dem Vorhandensein der Eigenschaftsausprägung „männlich“ die gleiche Aussagekraft zu wie dem Nichtvorhandensein. Dies gilt nicht für das Merkmal „Nationalität“ mit den Ausprägungen „Deutscher“ und „Nicht-Deutscher“; denn durch die Aussage „Nicht-Deutscher“ lässt sich die genaue Nationalität, die möglicherweise von Interesse ist, nicht bestimmen. Wenn also das Vorhandensein einer Eigenschaft (eines Merkmals) dieselbe Aussagekraft für die Gruppierung besitzt wie das Nichtvorhandensein, so ist Ähnlichkeitsmaßen, die im Zähler alle Übereinstimmungen berücksichtigen (z. B. M-Koeffizient) der Vorzug zu gewähren. Umgekehrt ist es ratsam, den Jaccard-Koeffizienten oder mit ihm verwandte Proximitätsmaße heranzuziehen. Insbesondere auch bei ungleich verteilten Merkmalen (z. B. Leiden an einer sehr seltenen Krankheit), führt die unreflektierte Anwendung von Proximitätsmaßen zu inhaltlich-sachlichen Verzerrungen, wenn bspw. der höchstwahrscheinliche Fall, dass zwei Personen nicht an ein und demselben seltenen Leiden erkranken, als Ähnlichkeit interpretiert wird.

Nichtvorhandensein einer Eigenschaft

8.2.1.2 Ähnlichkeitsermittlung bei nominaler Variablenstruktur

Bei nominal skalierten Variablen bzw. mehrkategorialen Merkmalen bestehen in SPSS grundsätzlich zwei Möglichkeiten, diese im Rahmen einer Clusteranalyse zu berücksichtigen:

- Transformation in binäre Variable
- Analyse von Häufigkeitsdaten

8.2.1.2.1 Transformation in binäre Variable

Nominale Merkmale, die mehr als zwei mögliche Merkmalsausprägungen aufweisen, können in binäre (Hilfs-)Variable zerlegt werden, und jeder Merkmalsausprägung (Kategorie) wird entweder der Wert 1 (Eigenschaft vorhanden) oder der Wert 0 (Eigenschaft nicht vorhanden) zugewiesen (sog. Binärzerlegung). Damit lassen sich mehrkategoriale Merkmale in Binärvariable (0/1-Variable) zerlegen, weshalb die im vorangegangenen Abschnitt behandelten Proximitätsmaße bei binärer Variablenstruktur auf einen solchen Datensatz angewandt werden können. Dabei ist allerdings zu berücksichtigen, dass bei großer und insbesondere bei stark unterschiedlicher Anzahl von Kategorien solche Proximitätsmaße für binäre Daten zu deutlichen Verzerrungen führen können, die den gemeinsamen Nichtbesitz einer Eigenschaft als Übereinstimmung von Objekten betrachten (z. B. M-Koeffizient). Die Vorgehensweise sei an folgendem Beispiel verdeutlicht:

Binärzerlegung

Betrachtet wird die nominale Variable „Beanstandungen einer Lieferung“, die vier Beanstandungskategorien (Merkmalsausprägungen) beinhalten soll. Abbildung 8.12 zeigt neben den vier Beanstandungskategorien auch die Transformation in vier binäre Variable, wobei durch die Merkmalsausprägungen (A bis D) keine Rangordnung zum Ausdruck gebracht werden kann.

Beispiel

Art der Beanstandung	Beanstandungskategorien	Transformation in mehrere binäre Variable
Fehlerhafte Ware	A	1000
Unvollständige Lieferung	B	0100
Verpackungsschäden	C	0010
Verspätete Lieferung	D	0001

Abbildung 8.12: Beispiel einer Datentransformation (Binärzerlegung)

Die Zahl der Kategorien (Ausprägungen) einer nominalen Variablen bestimmt also die Länge des aus Nullen und Einsen bestehenden Feldes. In unserem Fall umfasst das Feld somit vier Stellen bzw. vier binäre Variable. Für jede Beanstandungsart ist jeweils eine Spalte vorgesehen, die bei Gültigkeit mit einer Eins versehen wird (= Merkmalsausprägung vorhanden). Tritt beispielsweise ein Verpackungsschaden auf, so wird die für diese Klasse vorgesehene dritte Spalte mit einer Eins versehen und die restlichen Spalten erhalten jeweils eine Null. Bezüglich der Verwendung der Ähnlichkeitskoeffizienten bei nominalen (mehrstufigen) Variablen ist darauf hinzuweisen, dass bei großer und/oder unterschiedlicher Stufenzahl der Merkmale die Maße, die den gemeinsamen Nicht-Besitz als Übereinstimmung interpretieren (d. h. der Wert wird mit in den Zähler genommen), wegen der Verzerrungsgefahr möglichst keine Berücksichtigung finden sollten. Würden wir beispielsweise die Ähnlichkeit zweier Objekte bezüglich der Zahl der Beanstandungen überprüfen, so ergäbe sich im obigen Beispiel dem M-Koeffizienten entsprechend – unabhängig von der Wahl der beiden differierenden Beanstandungsstufen – immer mindestens ein Ähnlichkeitswert von 0,5. Dass dieses Ergebnis wenig sinnvoll ist, bedarf keiner besonderen Erläuterung.

8.2.1.2.2 Analyse von Häufigkeitsdaten

Analyse von Häufigkeitsdaten

Auch bei der Verwendung von *Häufigkeitsdaten* werden die Kategorien einer nominalen Variable als eigenständige Variable betrachtet, und es werden die Häufigkeiten der Nennung einer Merkmalskategorie in einem Datensatz analysiert. Zur Verdeutlichung sei auf folgendes Beispiel zurückgegriffen:

In einer Befragung seien 100 Personen nach ihrer Einschätzung von fünf Emulsionsfetten (Rama, Homa, Flora, SB und Holländische Butter) im Hinblick auf die gewünschte Verpackung befragt worden. Als Antwortkategorien seien „Plastikbecher“, „Kartonbox“ und „Papier“ als mögliche Verpackungsarten vorgegeben worden. Abbildung 8.13 enthält die Häufigkeiten der Nennung je Verpackungsart für die entsprechenden Emulsionsfette, wobei Mehrfachnennungen möglich waren.

	Verpackungsart		
Marke	Plastikbecher	Kartonbox	Papier
Rama	24	65	12
Homa	83	30	21
Flora	75	28	22
SB	35	55	21
Holl. Butter	20	40	75

Abbildung 8.13: Häufigkeitsdaten zur Verpackungsart

Chi-Quadrat-Homogenitätstest

Die Daten in Abbildung 8.13 können als Kreuztabelle der beiden nominal skalierten Variablen „Marke“ und „Verpackungsart“ interpretiert werden, auf die sich nun ein *Chi-Quadrat-Homogenitätstest* anwenden lässt.[8] Diesem Test liegt die *Nullhypothese* zugrunde, dass ein Merkmal in den zwei oder mehr Stichproben jeweils zugrunde liegenden Grundgesamtheiten jeweils die gleiche Verteilung besitzt. Diese Homogenitätshypothese impliziert, dass die empirischen Häufigkeiten mit den theoretischen zu erwartenden Häufigkeiten übereinstimmen. Überschreitet der empirische Chi-Quadrat-Wert für jeweils zwei Stichproben (hier: Marken) einem dem Signifikanzniveau entsprechenden Wert der Chi-Quadrat-Tabelle der theoretischen Chi-Quadrat-Verteilung (vgl. Anhang A.4), so ist die Nullhypothese mit der vorher festgelegten Irrtumswahrscheinlichkeit zu verwerfen. Je größer demzufolge der Chi-Quadrat-Wert ausfällt, desto größer ist die Wahrscheinlichkeit, dass die Häufigkeiten der betrachteten beiden Stichproben (hier: Marken bzw. Fälle) nicht der gleichen Grundgesamtheit entstammen und somit als sehr unterschiedlich einzustufen sind. Das bei Häufigkeitsdaten zur Berechnung der Distanzmatrix heranzuziehende *Chi-Quadrat-Maß* (= Quadratwurzel aus dem empirischen Chi-Quadrat-Wert) basiert auf dem Chi-Quadrat-Test, mit dem die Gleichheit zweier Häufigkeiten-Sets gemessen wird, und es stellt somit ein Distanz- oder Unähnlichkeitsmaß dar. Die Distanzmatrix zu obigen Häufigkeitsdaten nach dem Chi-Quadrat-Maß sind in Abbildung 8.14 wiedergegeben.

Chi-Quadrat-Maß

	Näherungsmatrix Chi-Quadrat zwischen Häufigkeiten-Sets				
Fall	1: Rama	2: Homa	3: Flora	4: SB	5: Holl. Butter
1: Rama	0,000	6,642	6,470	2,209	6,931
2: Homa	6,642	0,000	0,430	4,994	8,387
3: Flora	6,470	0,430	0,000	4,754	7,941
4: SB	2,209	4,994	4,754	0,000	5,901
5: Holl. Butter	6,931	8,387	7,941	5,901	0,000

Dies ist eine Unähnlichkeitsmatrix

Abbildung 8.14: Distanzmatrix der Häufigkeitsdaten nach dem Chi-Quadrat-Maß

Die Ergebnisse zeigen, dass die Häufigkeitsdaten von „Homa“ und „Flora“ mit einem Wert des Chi-Quadrat-Maßes von 0,430 die geringste Distanz (größte Ähnlichkeit) aufweisen und somit auf der ersten Stufe zu fusionieren wären. Entsprechend sind im nächsten Schritt „Rama“ und „SB“ bei einem Wert des Chi-Quadrat-Maßes von 2,209 zu fusionieren. Neben dem Chi-Quadrat-Maß kann in SPSS für Häufigkeitsdaten auch noch das *Phi-Quadrat-Maß* herangezogen werden, dessen Berechnung auf dem Chi-Quadrat-Maß basiert und zusätzlich eine Normalisierung der Daten vornimmt. Eine solche Normalisierung ist immer dann zweckmäßig, wenn die absoluten Häufigkeiten stark zwischen den einzelnen Paarvergleichen variieren.

Phi-Quadrat-Maß

8.2.1.3 Ähnlichkeitsermittlung bei metrischer Variablenstruktur

Zur Erläuterung von Proximitätsmaßen bei metrischem Skalenniveau der Beschreibungsmerkmale der Objekte wird folgendes Beispiel verwendet: In einer Befragung seien Personen nach ihrer Einschätzung von Emulsionsfetten (Butter, Margarine) be-

[8] Vgl. zum Chi-Quadrat-Test ausführlich das Kapitel „Kreuztabellierung und Kontingenzanalyse“ in diesem Buch.

fragt worden. Dabei seien die Marken Rama, Homa, Flora, SB und Holländische Butter anhand der Variablen Kaloriengehalt, Preis und Vitamingehalt auf einer siebenstufigen Skala von hoch bis niedrig beurteilt worden. Die Abbildung 8.15 enthält die durchschnittlichen subjektiven Beurteilungswerte der 30 befragten Personen für die entsprechenden Emulsionsfette.

	Eigenschaften		
Marken	Kaloriengehalt	Preis	Vitamingehalt
Rama	1	2	1
Homa	2	3	3
Flora	3	2	1
SB	5	4	7
Holl. Butter	6	7	6

Abbildung 8.15: Ausgangsdatenmatrix mit metrischer Variablenstruktur

8.2.1.3.1 Minkowski-Metriken oder L-Normen

Ein weit verbreitetes Distanzmaß bei metrischer Variablenstruktur bilden die sog. *Minkowski-Metriken* oder *L-Normen*, die wie folgt berechnet werden:

Minkowski-Metrik

Minkowski-Metrik

$$d_{k,l} = [\sum_{j=1}^{J} |x_{kj} - x_{lj}|^r]^{\frac{1}{r}} \tag{8.2}$$

mit

$d_{k,l}$ = Distanz der Objekte k und l
x_{kj}, x_{lj} = Wert der Variablen j bei Objekt k,l (j=1,2,...,J)
$r \geq 1$ = Minkowski-Konstante

Dabei stellt r eine positive Konstante dar. Für r=1 erhält man die *City-Block-Metrik* (L1-Norm) und für r=2 die *Euklidische Distanz* (L2-Norm). Die *City-Block-Metrik* (auch Manhattan- oder Taxifahrer-Metrik genannt) spielt bei praktischen Anwendungen z. B. bei der Clusterung von Standorten eine bedeutende Rolle. Sie wird berechnet, indem man die Differenz bei jeder Eigenschaft für ein Objektpaar bildet und die sich ergebenden absoluten Differenzwerte addiert. Die Berechnung dieser Distanz (d) sei beispielhaft für das Objektpaar „Rama“ und „Homa“ (vgl. Abbildung 8.15) durchgeführt, wobei die erste Zahl bei der Differenzbildung jeweils den Eigenschaftswert von „Rama“ darstellt.

City-Block-Metrik (L1-Norm)

$$\begin{aligned} d_{\text{Rama,Homa}} &= |1-2| + |2-3| + |1-3| \\ &= 1 + 1 + 2 \\ &= 4 \end{aligned}$$

Zwischen den Produkten „Rama“ und „Homa“ ergibt sich somit aufgrund der L1-Norm eine Distanz von 4. In der gleichen Weise werden für alle anderen Objektpaare die Abstände ermittelt. Das Ergebnis der Berechnungen zeigt Abbildung 8.16.

	Rama	Homa	Flora	SB	Holl. Butter
Rama	0				
Homa	4	0			
Flora	2	4	0		
SB	12	8	10	0	
Holl. Butter	15	11	13	5	0

Abbildung 8.16: Distanzmatrix entsprechend der L1-Norm (Block-Metrik)

Da ein Objekt zu sich selbst immer eine Distanz von Null besitzt, besteht die Hauptdiagonale einer Distanzmatrix immer aus Nullen. Aus diesem Grund wollen wir im Folgenden bei der Aufstellung einer Distanzmatrix die Hauptdiagonalwerte jeweils vernachlässigen, d. h. die erste Zeile und die letzte Spalte der Distanzmatrix in Abbildung 8.16 können eliminiert werden. Diese Abbildung macht deutlich, dass mit einem Abstandswert von 2 das Produktpaar „Flora“ und „Rama“ die größte Ähnlichkeit aufweist. Die geringste Ähnlichkeit besteht demgegenüber zwischen „Holländischer Butter“ und der Margarinemarke „Rama“. Hier beträgt die Differenz 15.

Anwendungsvoraussetzungen

Bei der Anwendung der *Minkowski-Metriken* ist darauf zu achten, dass *vergleichbare Maßeinheiten* zugrunde liegen. Das ist in unserem Beispiel erfüllt, da alle Eigenschaftsmerkmale der Margarinemarken auf einer von 1 bis 7 gehenden Ratingskala erhoben wurden. Ist diese Voraussetzung *nicht* erfüllt, so müssen die Ausgangsdaten zuerst z. B. mit Hilfe einer *Standardisierung* vergleichbar gemacht werden.

8.2.1.3.2 Einfache und quadrierte Euklidische Distanz

Die Euklidische Distanz zählt zu den weit verbreiteten Distanzmaßen bei empirischen Anwendungen. Dabei werden für jedes Objektpaar die Differenzwerte jeder Eigenschaft quadriert und anschließend addiert. Die Euklidische Distanz ergibt sich, indem anschließend aus der Summe die Quadratwurzel gezogen wird.

Für unsere Beispieldaten (vgl. Abbildung 8.15) lässt sich z. B. für das Produktpaar „Rama“ und „Flora“ zunächst die *quadrierte Euklidische Distanz* wie folgt berechnen:

$$\begin{aligned} d^2_{\text{Rama,Flora}} &= (1-3)^2 + (2-2)^2 + (1-1)^2 \\ &= 4 \qquad\qquad\qquad\qquad\qquad\qquad = 4+0+0 \end{aligned}$$

Quadrierte Euklidische Distanz

Durch die Quadrierung werden große Differenzwerte bei der Berechnung der Distanz stärker berücksichtigt, während geringen Differenzwerten ein kleineres Gewicht zukommt. Die *Euklidische Distanz* ergibt sich dann durch Radizieren der quadrierten Euklidischen Distanz; in obigem Beispiel erhält man den Wert 2. Sowohl die quadrierte Euklidische Distanz als auch die Euklidische Distanz können als Maß für die *Unähnlichkeit* zwischen Objekten (Distanz) herangezogen werden. Da eine Reihe von Algorithmen auf der quadrierten Euklidischen Distanz aufbaut, stützen sich die folgenden Betrachtungen ebenfalls auf die quadrierte Euklidische Distanz. Die Abbildung 8.17 fasst die quadrierten Euklidischen Distanzen für unser 5-Produkte-Beispiel zusammen.

L1-Norm und L2-Norm im Vergleich

Bezüglich des ähnlichsten und des unähnlichsten Paares gelangt man bei der quadrierten Euklidischen Distanz zur gleichen Aussage wie bei der City-Block-Metrik.

	Rama	Homa	Flora	SB
Homa	6			
Flora	4	6		
SB	56	26	44	
Holländische Butter	75	41	59	11

Abbildung 8.17: Distanzmatrix nach der quadrierten Euklidischen Distanz (Ausgangsdistanzmatrix)

Fasst man die Reihenfolge der Ähnlichkeiten nach beiden Metriken in einer Tabelle zusammen (Abbildung 8.18), so wird deutlich, dass sich bei den Produktpaaren „SB“ und „Flora“ sowie „Holländische Butter“ und „Homa“ eine Verschiebung der Reihenfolge der Ähnlichkeiten ergeben hat. Die Wahl des Distanzmaßes beeinflusst somit die Ähnlichkeitsreihenfolge der Untersuchungsobjekte.

	Rama	Homa	Flora	SB
Homa	2(2)			
Flora	1(1)	2(2)		
SB	7(7)	4(4)	5(6)	
Holländische Butter	9(9)	6(5)	8(8)	3(3)

Abbildung 8.18: Reihenfolge der Ähnlichkeiten entsprechend der quadrierten Euklidischen Distanz (Klammerwerte der Tabelle) sowie der L1-Norm

Die unterschiedlichen Ergebnisse sind auf die abweichende Behandlung der Differenzen zurückzuführen, da bei der L1-Norm alle Differenzwerte gleichgewichtig in die Berechnung eingehen.

8.2.1.3.3 Pearson-Korrelation oder Q-Korrelationskoeffizient

Soll die Bestimmung der Ähnlichkeit zwischen Objekten mit metrischer Variablenstruktur nicht über ein Distanzmaß, sondern direkt durch ein Ähnlichkeitsmaß erfolgen, so ist der *Q-Korrelationskoeffizient* hierfür ein gebräuchliches Maß, das sich wie folgt berechnen lässt:

Q-Korrelationskoeffizient

$$r_{k,l} = \frac{\sum_{j=1}^{J}(x_{jk} - \bar{x}_k) \cdot (x_{jl} - \bar{x}_l)}{\{\sum_{j=1}^{J}(x_{jk} - \bar{x}_k)^2 \cdot \sum_{j=1}^{J}(x_{jl} - \bar{x}_l)^2\}^{\frac{1}{2}}} \tag{8.3}$$

mit

$x_{j,k}$ = Ausprägung der Eigenschaft j bei Objekt (Cluster) k (bzw. 1), wobei: j= 1, 2, ..., J

$\bar{x}_k$ = Durchschnittswert aller Eigenschaften bei Objekt (Cluster) k (bzw. 1)

Pearson Korrelationskoeffizient

Der Q-Korrelationskoeffizient (in SPSS als *Pearson Korrelationskoeffizient* bezeichnet) berechnet die Ähnlichkeit zwischen zwei Objekten k und l unter Berücksichtigung aller Variablen eines Objektes. So ergibt sich z. B. für „Rama“ ein Variablendurchschnitt von (1+2+1)/3 = 4/3 (=$\bar{x}_k$) und für „Homa“ ein Variablendurchschnitt von (2+3+3)/3 = 8/3 (=$\bar{x}_l$).

Mit Hilfe dieser Variablendurchschnitte lässt sich die Ähnlichkeit zwischen „Rama“ und „Homa“ unter Verwendung der Ausgangsdaten aus Abbildung 8.15 wie folgt bestimmen:

$x_{jk} - \bar{x}_k$	$x_{jl} - \bar{x}_l$	$(x_{jk} - \bar{x}_k)(x_{jl} - \bar{x}_l)$	$(x_{jk} - \bar{x}_k)^2$	$(x_{jl} - \bar{x}_l)^2$
$-1/3$	$-2/3$	$2/9$	$1/9$	$4/9$
$2/3$	$1/3$	$2/9$	$4/9$	$1/9$
$-1/3$	$1/3$	$-1/9$	$1/9$	$1/9$
		$3/9$	$6/9$	$6/9$

Abbildung 8.19: Berechnungstabelle zur Bestimmung des Q-Korrelationskoeffizienten

$$r_{k,l} = \frac{3/9}{\sqrt{6/9 \cdot 6/9}} = 0,5$$

mit

k = Rama
l = Homa

Werden diese Berechnung für alle Produktpaare durchgeführt, so ergibt sich für unser Beispiel die in Abbildung 8.20 dargestellte Ähnlichkeitsmatrix auf Basis des Q- bzw. Pearson-Korrelationskoeffizienten.

	Rama	Homa	Flora	SB	Holl. Butter
Rama	1,000				
Homa	0,500	1,000			
Flora	0,000	-0,866	1,000		
SB	-0,756	0,189	-0,655	1,000	
Holl. Butter	1,000	0,500	0,000	-0,756	1,000

Abbildung 8.20: Ähnlichkeitsmatrix entsprechend dem Q-Korrelationskoeffizienten

Ähnlichkeiten und Distanzen im Vergleich

Werden diese Ähnlichkeitswerte mit den Distanzwerten aus Abbildung 8.17 verglichen, so wird deutlich, dass sich die Beziehungen zwischen den Objekten stark verschoben haben. Nach der quadrierten Euklidischen Distanz sind sich „Holländische Butter“ und „Rama“ am unähnlichsten, während sie nach dem Q-Korrelationskoeffizienten als das ähnlichste Markenpaar erkannt werden. Ebenso sind nach Euklid „Flora“ und „Rama“ mit einer Distanz von 4 sehr ähnlich, während sie mit einer Korrelation von 0 in Abbildung 8.20 als vollkommen unähnlich gelten. Diese Vergleiche machen deutlich, dass bei der Wahl des Proximitätsmaßes vor allem inhaltliche Überlegungen eine Rolle spielen. Betrachten wir zu diesem Zweck einmal die in Abbildung 8.21 dargestellten Profilverläufe von „Rama“ und „Holländischer Butter“ entsprechend den Ausgangsdaten in unserem Beispiel.

Die Profilverläufe zeigen, dass „Rama“ und „Holländische Butter“ zwar sehr weit voneinander entfernt liegen, der Verlauf ihrer Profile aber vollkommen gleich ist. Von daher lässt sich erklären, warum sie bei Verwendung eines Distanzmaßes als

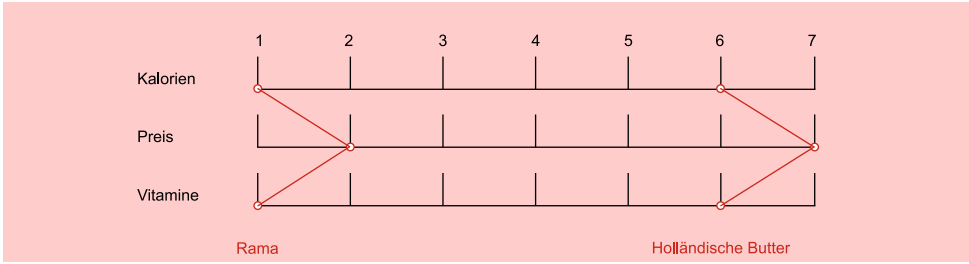


Abbildung 8.21: Profilverläufe von „Rama“ und „Holländischer Butter“

vollkommen unähnlich und bei Verwendung des Q-Korrelationskoeffizienten als vollkommen ähnlich erkannt werden. Allgemein lässt sich somit festhalten:

Zur Messung der Ähnlichkeit zwischen Objekten sind

- *Distanzmaße* immer dann geeignet, wenn der absolute Abstand zwischen Objekten von Interesse ist und die Unähnlichkeit dann als um so größer anzusehen ist, wenn zwei Objekte weit entfernt voneinander liegen; Distanzmaße
- *Ähnlichkeitsmaße* basierend auf Korrelationswerten immer dann geeignet, wenn der primäre Ähnlichkeitsaspekt im Gleichlauf zweier Profile zu sehen ist, unabhängig davon, auf welchem Niveau die Objekte liegen. Ähnlichkeitsmaße

Betrachten wir hierzu ein Beispiel: Eine Reihe von Unternehmen wird durch die Umsätze eines bestimmten Produktes im Ablauf von fünf Jahren (= Variable) beschrieben. Mit Hilfe der Clusteranalyse sollen solche Unternehmen zusammengefasst werden, die

1. im Zeitablauf ähnliche *Umsatzgrößen* mit diesem Produkt erzielt haben.
2. im Zeitablauf eine ähnliche *Umsatzentwicklungen* im Sinne der Veränderungsraten aufweisen.

Im ersten Fall ist für die Clusterung die *Umsatzhöhe* von Bedeutung. Folglich muss die Proximität zwischen den Unternehmen mit Hilfe eines *Distanzmaßes* ermittelt werden. Im zweiten Fall hingegen spielt die Umsatzhöhe keine Rolle, sondern die *Umsatzentwicklung*, und ein Ähnlichkeitsmaß (z. B. der Pearson Korrelationskoeffizient) ist das geeignete Proximitätsmaß.

8.2.1.4 Ähnlichkeitsermittlung bei gemischt skalierter Variablenstruktur

Durch die bisherige Darstellung wurde deutlich, dass die clusteranalytischen Verfahren kein spezielles Skalenniveau der Merkmale verlangen. Dieser Vorteil der allgemeinen Verwendbarkeit ist allerdings mit dem Problem der Behandlung *gemischter Variablen* verbunden, denn man verzeichnet in empirischen Studien sehr häufig sowohl metrische als auch nicht-metrische Eigenschaften der zu klassifizierenden Objekte. Ist dies der Fall, so ist eine Antwort auf die Frage zu finden, wie Merkmale mit unterschiedlichem Skalenniveau gemeinsam Berücksichtigung finden können. Zwei grundsätzliche *Verfahrensweisen* stehen zur Verfügung:[9] Gemischt skalierte Variable

[9] Vgl. Kaufmann/Pape (1996), S. 452 f.

Getrennte Behandlung unterschiedlich skalierter Variablen

Eine erste Möglichkeit besteht in der für metrische und nicht-metrische Variablen *getrennten Berechnung von Ähnlichkeitskoeffizienten bzw. Distanzen*. Die Gesamtähnlichkeit ermittelt sich dann als ungewichteter oder gewichteter Mittelwert der im vorherigen Schritt berechneten Größen. Verdeutlichen wir uns die Vorgehensweise am Beispiel der Produkte „Rama“ und „Flora“. Die Ähnlichkeit der Produkte soll anhand der nominalen und der metrischen Eigenschaften bestimmt werden. Als M-Koeffizient für diese beiden Produkte hatten wir einen Wert von 0,7 ermittelt (vgl. Abbildung 8.11). Die sich daraus ergebende Distanz der beiden Margarinesorten beläuft sich auf 0,3. Man erhält sie, indem man den Wert für die Ähnlichkeit von der Zahl 1 subtrahiert. Bei den metrischen Eigenschaften hatten wir für die beiden Produkte eine quadrierte euklidische Distanz von 4 (Abbildung 8.17) berechnet. Verwendet man nun das *ungewichtete arithmetische Mittel* als gemeinsames Distanzmaß, so erhalten wir in unserem Beispiel einen Wert von 2,15. Zu einer anderen Distanz kann man bei Anwendung des *gewichteten arithmetischen Mittels* gelangen. Hier besteht einmal die Möglichkeit, mehr oder weniger willkürlich extern Gewichte für den metrischen und den nicht-metrischen Abstand vorzugeben. Zum anderen kann man auch den jeweiligen Anteil der Variablen an der Gesamt-Variablenzahl als Gewichtungsfaktor heranziehen. Würde man den letztgenannten Weg beschreiten, so ergäben sich in unserem Beispiel keine Veränderungen gegenüber der Verwendung des ungewichteten arithmetischen Mittels, wenn wir sowohl zehn nominale als auch zehn metrische Merkmale zur Klassifikation benutzt hätten.

Transformation des Skalenniveaus von Variablen

Eine zweite Möglichkeit zur Behandlung gemischt skalierter Variablen besteht in der *Transformation von einem höheren auf ein niedrigeres Skalenniveau*. Welche Varianten sich in dieser Hinsicht ergeben, sei am Beispiel des Merkmals „Preis“ verdeutlicht: Für die betrachteten 5 Emulsionsfette im „metrischen Fall“ habe man die nachstehenden durchschnittlichen Verkaufspreise ermittelt (bezogen auf eine 250-Gramm-Packung).

Holländische Butter	2,05 €
Rama	1,75 €
Flora	1,65 €
SB	1,59 €
Homa	1,35 €

Dichotomisierung

Eine Möglichkeit zur Umwandlung der vorliegenden Verhältnisskalen in binäre Skalen besteht in der *Dichotomisierung*. Hierbei ist eine Schwelle festzulegen, die zu einer Trennung der niedrig- und hochpreisigen Emulsionsfette führt. Wird diese Grenze z. B. bei 1,60 € angenommen, so erhalten die Preisausprägungen bis zu 1,59 € als Schlüssel eine Null und die darüber hinausgehenden Preise eine Eins. Vorteilhaft an dieser Vorgehensweise sind ihre Einfachheit sowie schnelle Anwendungsmöglichkeit. Als problematisch ist demgegenüber der hohe Informationsverlust anzusehen, denn „Flora“ stünde in preislicher Hinsicht mit „Holländischer Butter“ auf einer Stufe, obwohl die letztgenannte Marke wesentlich teurer ist. Ein weiterer Problemaspekt

Schwellenwert

besteht in der Festlegung der Schwelle. Ihre willkürliche Bestimmung kann leicht zu Verzerrungen der realen Gegebenheiten führen, was wiederum einen Einfluss auf das Gruppierungsergebnis hat.

Der Informationsverlust lässt sich verringern, wenn *Preisintervalle* gebildet werden und jedes Intervall binär derart kodiert wird, dass – wenn der Preis für ein Produkt in das Intervall fällt – eine Eins und ansonsten eine Null vergeben wird. Diese Vorgehensweise wurde bereits in Abschnitt 8.2.1.2.1 ausführlich dargestellt (vgl. Abbildung 8.12).

Abschließend sei eine dritte Möglichkeit genannt, die ebenfalls auf einer Einteilung in Preisklassen beruht. In unserem Beispiel gehen wir von vier Intervallen (vgl. Abbildung 8.22) aus. Zur Verschlüsselung benötigen wir dann drei binäre Merkmale. Die Kodierung einer Null bzw. einer Eins erfolgt entsprechend der Antwort auf die nachfolgenden Fragen. Intervalle

Merkmal 1:	Preis gleich oder größer als 1,40 €? nein=0 ja=1
Merkmal 2:	Preis gleich oder größer als 1,70 €? nein=0 ja=1
Merkmal 3:	Preis gleich oder größer als 2,00 €? nein=0 ja=1

Das erste Preisintervall verschlüsselt man somit durch drei Nullen, da jede Frage mit nein beantwortet wird. Geht man auch bei den anderen Klassen in der beschriebenen Weise vor, so ergibt sich die in Abbildung 8.22 enthaltene Kodierung. Wird nun die erhaltene Binärkombination z. B. zur Verschlüsselung von „Rama“ verwendet, so erhalten wir für dieses Produkt die Zahlenfolge „1 1 0“. Abbildung 8.23 enthält die weiteren Verschlüsselungen der Emulsionsfette. Verschlüsselung

	Binäres Merkmal		
PREIS	1	2	3
bis 1,40 €	0	0	0
1,41-1,69 €	1	0	0
1,70-1,99 €	1	1	0
2,00-2,30 €	1	1	1

Abbildung 8.22: Kodierung von Preisklassen

Produkte	Binär-Schlüssel		
Höllandische Butter	1	1	1
Rama	1	1	0
Flora	1	0	0
SB	1	0	0
Homa	0	0	0

Abbildung 8.23: Verschlüsselung der Emulsionsfette

Der besondere Vorteil dieses Verfahrens liegt in seinem geringen Informationsverlust, der umso geringer ausfällt, je kleiner die jeweilige Klassenspanne ist. Bei sieben Preisklassen könnte man beispielsweise zu einer Halbierung der Spannweite und damit zu einer besseren Wiedergabe der tatsächlichen Preisunterschiede gelangen. Ein Nachteil einer derartigen Verschlüsselung ist in der Zunahme des Gewichts der betreffenden Vorteil

Eigenschaft zu sehen. Gehen wir nämlich davon aus, dass in unserer Studie neben dem Merkmal „Preis“ nur noch Eigenschaften mit zwei Ausprägungen existieren, so lässt sich erkennen, dass dem Preis bei fünf Preisklassen ein vierfaches Gewicht zukommt. Eine Halbierung der Spannweiten führt dann zu einem achtfachen Gewicht. Inwieweit eine stärkere Berücksichtigung eines einzelnen Merkmals erwünscht ist, muss man im Einzelfall klären. Zur Behandlung von Daten mit gemischten Merkmalen schlägt *Gower* folgenden Koeffizienten vor:[10]

$$d_{ij} = \frac{\sum_{k=1}^{p} \delta_{ij}^{(k)} d_{ij}^{(k)}}{\sum_{f=1}^{p} \delta_{ij}^{(k)}} \tag{8.4}$$

Gower-Koeffizient

Der *Koeffizient von Gower* bietet den Vorteil, dass über die Integration des δ-(Delta)-Terms sowohl fehlende Werte (δ nimmt dann den Wert 0 an) als auch asymmetrische Binärmerkmale, bei denen das Vorhandensein bzw. das Nicht-Vorhandensein eines Merkmals anders behandelt wird, berücksichtigt werden können. Dabei nimmt δ dann den Wert 0 an, wenn beide Objekte das betreffende Merkmal nicht aufweisen und dies auch nicht als Ähnlichkeit gewertet wird. Abhängig vom Messniveau der Variablen wird die Distanz zwischen den Objekten i und j und ihren Merkmalsausprägungen x_{ij} bzw. x_{jk} wie folgt bestimmt:

Binäre oder nominalskalierte Merkmale:

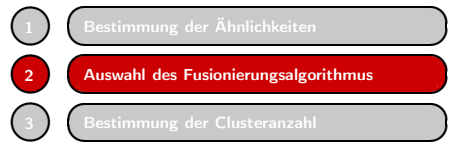
$$d_{ij}^{(k)} = \begin{cases} 1 & \text{wenn} \quad x_{ik} \neq x_{jk} \\ 0 & \text{wenn} \quad x_{ik} = x_{jk} \end{cases}$$

Metrische Merkmale:

$$d_{ij}^{(k)} = \frac{|x_{ik} - x_{jk}|}{R(k)} \quad \text{mit} \quad R(k) = \max_i x_{ik} - \min_i x_{ik}$$

8.2.2 Auswahl des Fusionierungsalgorithmus

Fusionierungsalgorithmen der Clusteranalyse



Die bisherigen Ausführungen haben gezeigt, wie sich mit Hilfe von Proximitätsmaßen eine Distanz- oder Ähnlichkeitsmatrix aus den Ausgangsdaten ermitteln lässt. Die gewonnene Distanz- oder Ähnlichkeitsmatrix bildet nun den Ausgangspunkt der Clusteralgorithmen, die eine Zusammenfassung der Objekte zum Ziel haben. Die Clusteranalyse bietet dem Anwender ein breites Methodenspektrum an Algorithmen (Cluster-Algorithmen) zur Gruppierung einer gegebenen Objektmenge. Nach der Zahl der Variablen, die beim Fusionierungsprozess Berücksichtigung finden, lassen sich *monothetische und polythetische Verfahren* unterscheiden. *Monothetische Verfahren* sind dadurch gekennzeichnet, dass sie zur Gruppierung jeweils nur eine Variable heranziehen. Der große Vorteil der Clusteranalyse liegt aber gerade darin, simultan alle relevanten Beschreibungsmerkmale (Variable) zur Gruppierung der Objekte heranzuziehen. Da dieser Zielsetzung aber nur *polythetische Verfahren* entsprechen, werden

Monothetische Verfahren

Polythetische Verfahren

[10] Vgl. Gower (1971), passim.

auch nur diese im Folgenden betrachtet. Eine weitere Einteilung der Clusteralgorithmen lässt sich entsprechend der Vorgehensweise im Fusionierungsprozess vornehmen. Die Abbildung 8.24 gibt hierzu einen entsprechenden Überblick.

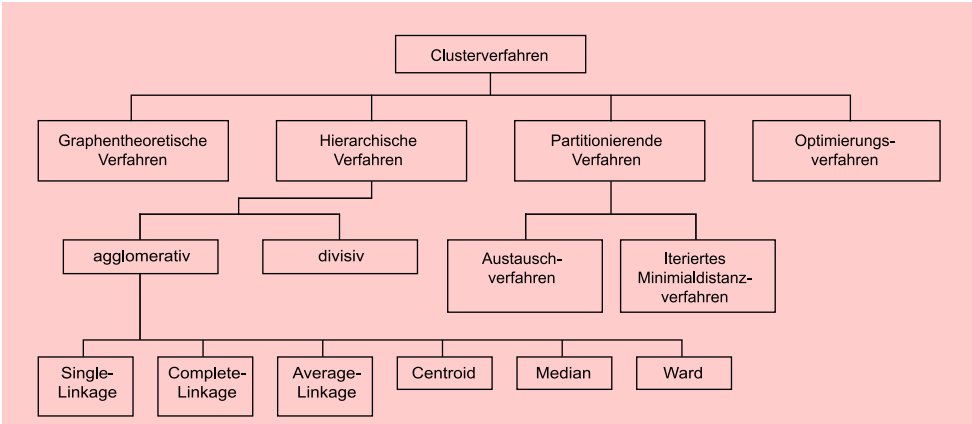


Abbildung 8.24: Überblick über ausgewählte Cluster-Algorithmen

Aus der Vielzahl existierender Verfahren werden im Folgenden, entsprechend ihrer Bedeutung, der Ablauf bei partitionierenden und bei hierarchischen Verfahren beispielhaft dargestellt:

Partitionierende Verfahren

- Die *partitionierenden Verfahren* gehen von einer gegebenen Gruppierung der Objekte (Startpartition) und, damit verbunden, einer festgelegten Zahl an Clustern aus und ordnen die einzelnen Elemente mit Hilfe eines Austauschalgorithmus zwischen den Gruppen so lange um, bis eine gegebene Zielfunktion ein Optimum erreicht. Während bei den hierarchischen Verfahren eine einmal gebildete Gruppe im Analyseprozess nicht mehr aufgelöst werden kann, haben die partitionierenden Verfahren den Vorteil, dass während des Fusionierungsprozesses Elemente zwischen den Gruppen getauscht werden können.

Hierarchische Verfahren

- Bei den *hierarchischen Verfahren* wird zwischen agglomerativen und divisiven Algorithmen unterschieden. Während bei den agglomerativen Verfahren von der feinsten Partition (sie entspricht der Anzahl der Untersuchungsobjekte) ausgegangen wird, bildet die gröbste Partition (alle Untersuchungsobjekte befinden sich in einer Gruppe) den Ausgangspunkt der divisiven Algorithmen. Somit lässt sich der Ablauf der ersten Verfahrensart durch die Zusammenfassung von Gruppen und der der zweiten Verfahrensart durch die Aufteilung einer Gesamtheit in Gruppen charakterisieren.

Im Folgenden wird die grundsätzliche Vorgehensweise dieser beiden Gruppen von Cluster-Algorithmen dargestellt. Dabei liegt der Schwerpunkt der Betrachtungen auf den agglomerativen Verfahren, da ihnen in der Praxis die größte Bedeutung zukommt. Demgegenüber werden die divisiven Verfahren aufgrund ihrer nach wie vor eher geringeren Bedeutsamkeit nicht weiter betrachtet. Allerdings stehen in SPSS mit dem *Klassifizierungsbaum* (*Menüfolge: Analysieren* → *Klassifizieren* → *Baum*) divisive Clusteralgorithmen zur Verfügung.

Klassifizierungsbaum

8.2.2.1 Partitionierende Verfahren

Die Gemeinsamkeit partitionierender Verfahren besteht darin, dass – ausgehend von einer vorgegebenen Gruppeneinteilung – durch Verlagerung der Objekte in andere Gruppen versucht wird, zu einer besseren Lösung zu gelangen.[11] Die in diesem Bereich existierenden Verfahren unterscheiden sich in zweierlei Hinsicht. Erstens ist in diesem Zusammenhang auf die Art und Weise, wie die Verbesserung der Clusterbildung gemessen wird, hinzuweisen. Ein zweiter Unterschied besteht in der Regelung des Austausches der Objekte zwischen den Gruppen.

Austauschverfahren

Im Folgenden sei beispielhaft das *Austauschverfahren* kurz erläutert, wobei die Verbesserung einer Gruppenbildung durch das Varianzkriterium gemessen werden soll (vgl. Abschnitt 8.2.2.2.2). Das Austauschverfahren beinhaltet folgende Ablaufschritte:

1. Schritt: Es wird eine Anfangspartition vorgegeben.

2. Schritt: Pro Gruppe wird je Eigenschaft das arithmetische Mittel berechnet.

3. Schritt: Für die jeweils gültige Gruppenzuordnung wird die Fehlerquadratsumme (Varianzkriterium) über alle Gruppen ermittelt.

4. Schritt: Die Objekte werden daraufhin untersucht, ob durch eine Verlagerung das Varianzkriterium vermindert werden kann.

5. Schritt: Das Objekt, das zu einer maximalen Verringerung führt, wird in die entsprechende Gruppe verlagert.

6. Schritt: Für die empfangende und die abgebende Gruppe müssen die neuen Mittelwerte berechnet werden.

Das Verfahren setzt den nächsten Durchlauf mit dem 3. Schritt fort. Beendet wird die Clusterung, wenn alle Objekte bezüglich ihrer Verlagerung untersucht wurden und sich keine Verbesserung des Varianzkriteriums mehr erreichen lässt. Der Abbruch an dieser Stelle muss erfolgen, da nicht alle grundsätzlich möglichen Gruppenbildungen auf ihren Zielfunktionswert hin untersucht werden können. Diese Aussage lässt sich leicht dadurch erklären, dass für m Objekte und g Gruppen g^m Einteilungsmöglichkeiten existieren. Gehen wir beispielsweise von 10 Objekten und drei Gruppen aus, so existieren bereits $3^{10} = 59.049$ Möglichkeiten zur Clusterbildung. Aufgrund dieser hohen Zahl möglicher Clusterbildungen ist i. d. R. eine vollständige Enumeration nicht sinnvoll, und man gelangt häufig nur zu lokalen und nicht zu globalen Optima. Daher ist es bei den partitionierenden Verfahren erforderlich, zu einer Verbesserung der Lösung durch eine *Veränderung der Startpartition* zu gelangen. Inwieweit hierdurch eine homogenere Gruppenbildung erzielt wird, lässt sich anhand des Varianzkriteriums ablesen. Ist der Zielfunktionswert gesunken, so ist man dem Vorhaben der Zusammenfassung gleichartiger Objekte näher gekommen. Hinter der einfachen Feststellung

Startpartition

„*Veränderung der Startpartition*“ verbergen sich zwei Entscheidungsprobleme:

[11] Vgl. zu den partitionierenden Verfahren auch: Fett (2008), S. 26 ff.; Kaufman/Rousseeuw (2008), S. 38 ff.

1. Es ist festzulegen, auf wie viele Gruppen die Objekte zu verteilen sind.
2. Es ist festzulegen, nach welchem Modus die Untersuchungsobjekte auf die Startgruppen zu verteilen sind. Hierzu kann beispielsweise eine Zufallszahlentabelle herangezogen werden. Eine andere Möglichkeit besteht darin, dass die Objekte entsprechend der Reihenfolge ihrer Nummerierung den Gruppen 1,2,..., g_1; 1,2,..., g_2; usw. zugeordnet werden. Weiterhin lassen sich auch die Ergebnisse hierarchischer Verfahren für die Festlegung der Startpartition heranziehen.

Werden die agglomerativen hierarchischen und die partitionierenden Verfahren verglichen, so ergibt sich ein zentraler Unterscheidungspunkt: Während bei den erstgenannten Verfahren sich ein einmal konstruiertes Cluster in der Analyse nicht mehr auflösen lässt, kann bei den partitionierenden Verfahren jedes Element von Cluster zu Cluster beliebig verschoben werden. Die partitionierenden Verfahren zeichnen sich somit durch eine größere Variabilität aus. Sie haben jedoch bei praktischen Anwendungen eine deutlich geringere Verbreitung gefunden. Dieser Umstand ist vor allem durch folgende Punkte begründet:

Größere Variabilität partitionierender Verfahren

- Die Ergebnisse der partitionierenden Verfahren werden verstärkt durch die der „Umordnung“ der Objekte zugrunde liegenden Zielfunktion beeinflusst.
- Die Wahl der Startpartition ist häufig subjektiv begründet und kann ebenfalls die Ergebnisse des Clusterprozesses beeinflussen. Sofern die Startpartition zufällig initialisiert wird, was bei einigen Prozeduren wie z. B. K-Means üblich ist, führt dies u. U. dazu, dass die entsprechend erzielten Cluster-Lösungen variieren und somit die Ergebnisse nicht vergleichbar sind.
- Bei partitionierenden Verfahren ergeben sich häufig nur lokale und keine globalen Optima, wenn keine vollständige Enumeration durchgeführt wird.

In SPSS ist mit der K-Means-Clusteranalyse (*Menüfolge: Analysieren → Klassifizieren → K-Means-Cluster*) ein partitionierender Clusteralgorithmus implementiert, der in Vorgängerversionen unter der Bezeichnung „Clusterzentrenanalyse“ aufgeführt war. Zur Durchführung von K-Means (Prozedur: QUICK CLUSTER) muss der Anwender vorab eine Anzahl von k Clustern vorgeben, zwischen denen dann auf Basis der betrachteten Variablen ein Austausch zwischen den Objekten zu den Gruppen solange vorgenommen wird, bis ein vorgegebenes Zielkriterium erfüllt ist. Als Zielkriterium wird dabei das Varianzkriterium verwendet und zur Distanzberechnung die einfache euklidische Distanz. Entsprechend müssen die zur Objektbeschreibung verwendeten Variablen metrisch skaliert sein.

K-Means-Clusteranalyse

8.2.2.2 Hierarchische Verfahren

8.2.2.2.1 Ablauf der agglomerativen Verfahren

Die in der Praxis häufig zur Anwendung kommenden *agglomerativen Algorithmen* umfassen die in Abbildung 8.24 dargestellten sechs Verfahren. Die Vorgehensweise dieser Verfahren kann durch die folgenden *allgemeinen Ablaufschritte* beschrieben werden:

Agglomerative Verfahren

1. Schritt: Gestartet wird mit der feinsten Partition, d. h. jedes Objekt stellt ein Cluster dar. In unserem Beispiel aus Abbildung 8.15 gehen wir somit von fünf Gruppen aus.

2. Schritt: Für alle in die Untersuchung eingeschlossenen Objekte wird die paarweise Distanz bzw. Ähnlichkeit der Objekte berechnet. In unserem Fall erhalten wir somit $\binom{5}{2} = 10$ Distanzen. Für den weiteren Verlauf gehen wir von den in Abbildung 8.17 enthaltenen quadrierten Euklidischen Distanzen aus.

3. Schritt: Es werden die beiden Cluster mit der größten Ähnlichkeit bzw. geringsten Distanz zueinander gesucht. Im ersten Durchlauf weisen die beiden Margarinemarken „Rama“ und „Flora“ den geringsten Abstand auf ($d^2 = 4$).

4. Schritt: Die beiden Gruppen mit der größten Ähnlichkeit werden zu einem neuen Cluster zusammengefasst. Die Zahl der Gruppen nimmt somit um 1 ab. Zum Ende des ersten Durchgangs existieren in unserem Beispiel noch vier Gruppen.

5. Schritt: Es werden die Abstände zwischen den neuen und den übrigen Gruppen berechnet, wodurch man zu einer *reduzierten* Distanzmatrix gelangt.

6. Schritt: Die Schritte 3 bis 5 werden solange wiederholt, bis alle Untersuchungsobjekte in einer Gruppe enthalten sind (sog. Ein-Cluster-Lösung).

Unterschiede in agglomerativen Verfahren

Die Unterschiede zwischen den agglomerativen Verfahren ergeben sich nur daraus, wie die Distanz zwischen einem Objekt (Cluster) R und dem neuen Cluster (P+Q) ermittelt wird. Sind zwei Objekte (Gruppen) P und Q zu vereinigen, so ergibt sich die Distanz D(R;P+Q) zwischen irgendeiner Gruppe R und der neuen Gruppe (P+Q) durch folgende Transformation:[12]

$$\begin{aligned} D(R;P+Q) &= A \cdot D(R,P) + B \cdot D(R,Q) + E \cdot D(P,Q) \\ &\quad + G \cdot |D(R,P) - D(R,Q)| \end{aligned} \tag{8.5}$$

mit

$$\begin{aligned} D(R,P) &= \text{Distanz zwischen den Gruppen R und P} \\ D(R,Q) &= \text{Distanz zwischen den Gruppen R und Q} \\ D(P,Q) &= \text{Distanz zwischen den Gruppen P und Q} \end{aligned}$$

Die Größen A, B, E und G sind Konstanten, die je nach verwendetem Algorithmus variieren. Die in Abbildung 8.25 dargestellten agglomerativen Verfahren erhält man durch Zuweisung entsprechender Werte für die Konstanten in Gleichung (8.5). Die Abbildung 8.25 zeigt die jeweiligen Wertzuweisungen und die sich damit ergebenden Distanzberechnungen bei ausgewählten agglomerativen Verfahren.[13]

Während bei den ersten vier Verfahren grundsätzlich alle möglichen Proximitätsmaße verwendet werden können, ist die Anwendung der Verfahren „Centroid“, „Median“ und „Ward“ nur sinnvoll bei Verwendung eines Distanzmaßes. Bezüglich des Skalenniveaus der Ausgangsdaten lässt sich festhalten, dass die Verfahren sowohl bei metrischen als auch bei nicht-metrischen Ausgangsdaten angewandt werden können. Entscheidend ist hier nur, dass die verwendeten Proximitätsmaße auf das Skalenniveau der Daten abgestimmt sind, denn nicht-metrische Proximitätsmaße stellen relative Häufigkeiten dar, die im Ergebnis metrisch interpretiert werden können.

[12] Vgl. Steinhausen/Langer (1977), S. 76.
[13] Vgl. Steinhausen/Langer (1977), S. 77.

Verfahren	Konstante A	B	E	G	Distanzberechnung (D(R;P+Q)) nach Gleichung (8.5):
Single Linkage	0,5	0,5	0	-0,5	$0,5 \cdot \{D(R,P)+D(R,Q)-\lvert D(R,P)-D(R,Q)\rvert\}$
Complete Linkage	0,5	0,5	0	0,5	$0,5 \cdot \{D(R,P)+D(R,Q)+\lvert D(R,P)-D(R,Q)\rvert\}$
Average Linkage (ungewichtet)	0,5	0,5	0	0	$0,5 \cdot \{D(R,P)+D(R,Q)\}$
Average Linkage (gewichtet)	$\frac{NP}{NP+NQ}$	$\frac{NQ}{NP+NQ}$	0	0	$\frac{1}{NP+NQ} \cdot \{NP \cdot D(R,P)+NQ \cdot D(R,Q)\}$
Centroid	$\frac{NP}{NP+NQ}$	$\frac{NQ}{NP+NQ}$	$-\frac{NP \cdot NQ}{(NP+NQ)^2}$	0	$\frac{1}{NP+NQ} \cdot \{NP \cdot D(R,P)+NQ \cdot D(R,Q)\} - \frac{NP \cdot NQ}{(NP+NQ)^2} \cdot D(P,Q)$
Median	0,5	0,5	-0,25	0	$0,5 \cdot \{D(R,P)+D(R,Q)\} - 0,25 \cdot D(P,Q)$
Ward	$\frac{NR+NP}{NR+NP+NQ}$	$\frac{NR+NQ}{NR+NP+NQ}$	$-\frac{NR}{NR+NP+NQ}$	0	$\frac{1}{NR+NP+NQ} \cdot \{(NR+NP) \cdot D(R,P) + (NR+NQ) \cdot D(R,Q) - NR \cdot D(P,Q)\}$

NR: Zahl der Objekte in Gruppe R
NP: Zahl der Objekte in Gruppe P
NQ: Zahl der Objekte in Gruppe Q

Abbildung 8.25: Distanzberechnung bei ausgewählten agglomerativen Verfahren[14]

8.2.2.2.2 Vorgehensweise der hierarchisch agglomerativen Clusterverfahren Single-Linkage, Complete-Linkage und Ward

Single-Linkage-Verfahren (Nächstgelegener Nachbar)

Single-Linkage-Verfahren (nächstgelegener Nachbar)

Das *Single-Linkage-Verfahren* vereinigt im ersten Schritt die Objekte, die gemäß der Distanzmatrix aus Abbildung 8.17 die *kleinste* Distanz aufweisen, d. h. die Objekte, die sich am ähnlichsten sind. Somit werden im ersten Durchlauf die Objekte „Rama“ und „Flora“ mit einer Distanz von 4 vereinigt. Da „Rama“ und „Flora“ nun eine eigenständige Gruppe bilden, muss im nächsten Schritt der Abstand dieser Gruppe zu allen übrigen Objekten bestimmt werden. Als Distanz zwischen der neuen Gruppe „Rama, Flora“ und einem Objekt (Gruppe) R wird nun der *kleinste* Wert der Einzeldistanzen zwischen „Rama“ und R bzw. „Flora“ und R herangezogen, sodass sich die

[14] In der deutschsprachigen Version von SPSS sind die Verfahren Single Linkage als „Nächstgelegener Nachbar“, Complete Linkage als „Entferntester Nachbar“ und Average Linkage als „Linkage zwischen den Gruppen“ bezeichnet. SPSS bietet bei dem Verfahren „Linkage zwischen den Gruppen“ jedoch nur die ungewichtete Variante.

neue Distanz gemäß Gleichung (8.5) wie folgt bestimmt (vgl. Abbildung 8.25):

$$D(R; P+Q) = 0,5\{D(R,P) + D(R,Q) - |D(R,P) - D(R,Q)|\} \tag{8.6}$$

Vereinfacht ergibt sich diese Distanz auch aus der Beziehung:

$$D(R; P+Q) = min\{D(R,P); D(R,Q)\}$$

Nearest Neighbour

Das Single-Linkage-Verfahren weist somit einer neu gebildeten Gruppe die kleinste Distanz zu, die sich aus den alten Distanzen der in der Gruppe vereinigten Objekte zu einem bestimmten anderen Objekt ergibt. Man bezeichnet diese Methode deshalb auch als „*Nearest-Neighbour-Verfahren*“ (*Nächstgelegener Nachbar*). Verdeutlichen wir uns dieses Vorgehen beispielhaft an der Distanzbestimmung zwischen der Gruppe „Rama, Flora“ und der Marke „SB“. Zur Berechnung der neuen Distanz sind die Abstände zwischen „Rama“ und „SB“ sowie zwischen „Flora“ und „SB“ heranzuziehen. Aus der Ausgangsdistanzmatrix (Abbildung 8.17) ist ersichtlich, dass die erstgenannte Distanz 56 und die zweitgenannte Distanz 44 beträgt. Somit wird für den zweiten Durchlauf als Distanz zwischen der Gruppe „Rama, Flora“ und der Marke „SB“ eine Distanz von 44 zugrunde gelegt. Abbildung 8.26 fasst die Vorgehensweise noch einmal graphisch zusammen. Der „Kreis“ um „Rama“ und „Flora“ soll verdeutlichen, dass sich die beiden Produkte bereits in einem Cluster befinden.

Formal lassen sich diese Distanzen auch mit Hilfe von Gleichung (8.6) bestimmen. Dabei ist P+Q die Gruppe „Flora (P) und Rama (Q)“, und R stellt jeweils ein verbleibendes Objekt dar. Die neuen Distanzen zwischen „Flora, Rama“ und den übrigen Objekten ergeben sich in unserem Beispiel dann wie folgt (vgl. die Werte in Abbildung 8.17):

$$\begin{aligned} D(\text{Homa; Flora+Rama}) &= 0,5 \cdot \{(6+6) - |6-6|\} &&= 6 \\ D(\text{SB; Flora+Rama}) &= 0,5 \cdot \{(44+56) - |44-56|\} &&= 44 \\ D(\text{Butter; Flora+Rama}) &= 0,5 \cdot \{(59+75) - |59-75|\} &&= 59 \end{aligned}$$

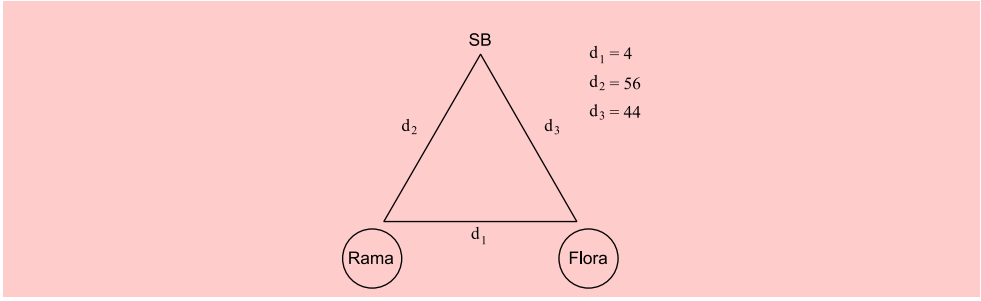


Abbildung 8.26: Berechnung der neuen Distanz beim Single-Linkage-Verfahren

Reduzierte Distanzmatrix

Die reduzierte Distanzmatrix ergibt sich, indem die Zeilen und Spalten der fusionierten Cluster aus der für den betrachteten Durchgang gültigen Distanzmatrix entfernt und dafür eine neue Spalte und Zeile für die gerade gebildete Gruppe eingefügt wird. Am Ende des ersten Durchgangs ergibt sich eine reduzierte Distanzmatrix (Abbildung 8.27), die im zweiten Schritt Verwendung findet.

	Flora, Rama	Homa	SB
Homa	6		
SB	44	26	
Holländische Butter	59	41	11

Abbildung 8.27: Distanzmatrix nach dem ersten Fusionsschritt beim Single-Linkage-Verfahren

Entsprechend der reduzierten Distanzmatrix werden im nächsten Schritt die Objekte (Cluster) vereinigt, die die geringste Distanz aufweisen. Im vorliegenden Fall wird „Homa“ in die Gruppe „Flora, Rama“ aufgenommen, da hier die Distanz (d=6) am kleinsten ist. Für die reduzierte Distanzmatrix im zweiten Durchlauf errechnen sich dann die Abstände der Gruppe „Flora, Rama, Homa“ zu „SB“ bzw. „Holländische Butter“ wie folgt:

$$D(\text{SB; Flora+Rama+Homa}) = 0,5 \cdot \{(44 + 26) - |44 - 26|\} = 26$$
$$D(\text{Butter; Flora+Rama+Homa}) = 0,5 \cdot \{(59 + 41) - |59 - 41|\} = 41$$

Damit ergibt sich die reduzierte Distanzmatrix im zweiten Schritt gemäß Abbildung 8.28.

	Flora, Rama, Homa	SB
SB	26	
Holländische Butter	41	11

Abbildung 8.28: Distanzmatrix nach dem zweiten Fusionsschritt beim Single-Linkage Verfahren

Den Werten in Abbildung 8.28 entsprechend werden im nächsten Schritt die Marken „SB“ und „Holländische Butter“ zu einer eigenständigen Gruppe zusammengefasst. Die Distanz zwischen den verbleibenden Gruppen „Flora, Rama, Homa“ und „SB, Holländische Butter“ ergibt sich dann auf Basis von Abbildung 8.28 wie folgt:

$$D(\text{Flora,Rama,Homa; SB,Butter}) = 0,5 \cdot \{(26 + 41) - |26 - 41|\} = 26$$

Das Ergebnis der Cluster-Analyse nach dem Single-Linkage-Verfahren lässt sich graphisch durch das in Abbildung 8.29 dargestellte Dendrogramm verdeutlichen. Dendrogramm

Dadurch, dass das Single-Linkage-Verfahren als neue Distanz zwischen zwei Gruppen immer den kleinsten Wert der Einzeldistanzen heranzieht, ist es geeignet, „Ausreißer“ in einer Objektmenge zu erkennen. Da das Single-Linkage-Verfahren dazu neigt, viele kleine und wenige große Gruppen zu bilden (kontrahierendes Verfahren), bilden die kleinen Gruppen einen Anhaltspunkt für die Identifikation von „Ausreißern“ in der Objektmenge. Das Verfahren hat dadurch aber den Nachteil, dass es aufgrund der großen Gruppen zur Kettenbildung neigt, wodurch „schlecht“ getrennte Gruppen nicht aufgedeckt werden.[15] Ausreißer-Entdeckung

[15] Vgl. hierzu die Ausführungen im Abschnitt 8.2.2.3.

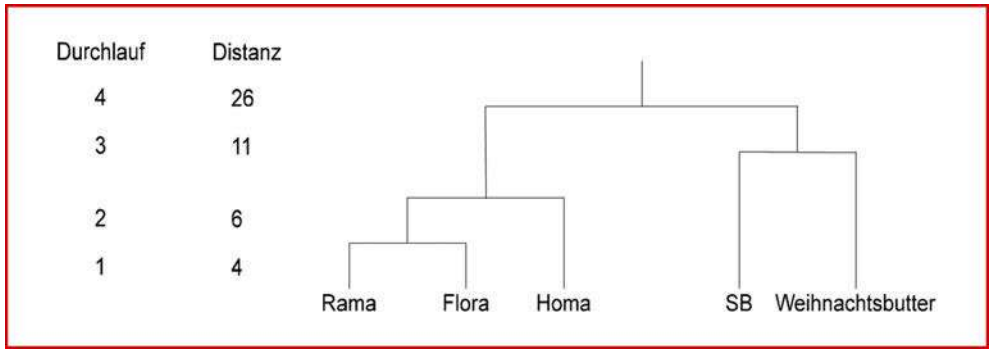


Abbildung 8.29: Dendrogramm für das Single-Linkage-Verfahren

Complete-Linkage-Verfahren (Entferntester Nachbar)

Complete-Linkage-Verfahren (entferntester Nachbar)

Der Unterschied zwischen dem Single-Linkage- und dem *Complete-Linkage-Verfahren* besteht in der Vorgehensweise bei der neuen Distanzbildung im vierten Schritt. Diese berechnet sich gemäß Gleichung (8.5) wie folgt (vgl. Abbildung 8.25):

$$D(R; P+Q) = 0,5 \cdot \{D(R,P) + D(R,Q) + |D(R,P) - D(R,Q)|\} \tag{8.7}$$

Es werden also nicht die geringsten Abstände als neue Distanz herangezogen – wie beim Single-Linkage-Verfahren –, sondern die größten Abstände, sodass sich für (8.6) auch schreiben lässt:

$$D(R; P+Q) = max\{D(R,P); D(R,Q)\}$$

Furthest Neighbour

Man bezeichnet dieses Verfahren deshalb auch als „*Furthest-Neighbour-Verfahren*“ (Entferntester Nachbar). Ausgehend von der Distanzmatrix in Abbildung 8.17 werden im ersten Schritt auch hier die Objekte „Rama“ und „Flora“ vereinigt. Der Abstand dieser Gruppe zu z. B. „SB“ entspricht aber jetzt in der reduzierten Distanzmatrix dem größten Einzelabstand, der entsprechend Abbildung 8.26 jetzt 56 beträgt. Formal ergeben sich die Einzelabstände gemäß (8.7) wie folgt:

$$\begin{aligned} D(\text{Homa; Flora+Rama}) &= 0,5 \cdot \{(6+6) + |6-6|\} &&= 6 \\ D(\text{SB; Flora+Rama}) &= 0,5 \cdot \{(44+56) + |44-56|\} &&= 56 \\ D(\text{Butter; Flora+Rama}) &= 0,5 \cdot \{(59+75) + |59-75|\} &&= 75 \end{aligned}$$

Damit erhalten wir die in Abbildung 8.30 dargestellte reduzierte Distanzmatrix.

	Flora, Rama	Homa	SB
Homa	6		
SB	56	26	
Holländische Butter	75	41	11

Abbildung 8.30: Reduzierte Distanzmatrix nach dem ersten Fusionsschritt beim Complete-Linkage-Verfahren

Im nächsten Durchlauf wird auch hier die Marke „Homa“ in die Gruppe „Rama, Flora“ aufgenommen, da entsprechend Abbildung 8.30 hier die kleinste Distanz mit

d=6 auftritt. Der Prozess setzt sich nun ebenso wie beim Single-Linkage-Verfahren fort, wobei die jeweiligen Distanzen immer nach Gleichung (8.7) bestimmt werden. Hier sei nur das Endergebnis anhand eines Dendrogramms aufgezeigt (Abbildung 8.31).

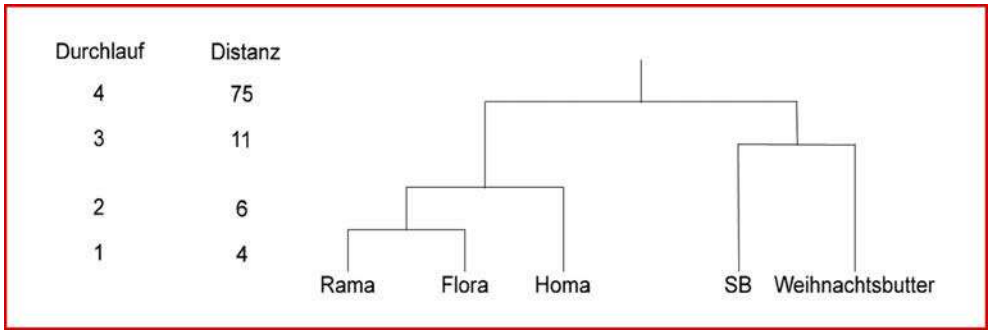


Abbildung 8.31: Dendrogramm für das Complete-Linkage-Verfahren

Obwohl in diesem Beispiel der Fusionierungsprozess beim Single- und Complete-Linkage-Verfahren nahezu identisch verläuft, tendiert das Complete-Linkage-Verfahren eher zur Bildung kleiner Gruppen. Das liegt darin begründet, dass als neue Distanz jeweils der größte Wert der Einzeldistanzen herangezogen wird. Von daher ist das Complete-Linkage-Verfahren im Gegensatz zum Single-Linkage-Verfahren nicht dazu geeignet, „Ausreißer“ in einer Objektgesamtheit zu entdecken. Diese führen beim Complete-Linkage-Verfahren eher zu einer Verzerrung des Gruppierungsprozesses und sollten daher vor Anwendung dieses Verfahrens (etwa mit Hilfe des Single-Linkage-Verfahrens) eliminiert werden.[16]

Ward-Verfahren

Das *Ward-Verfahren* hat in der Praxis eine weite Verbreitung gefunden. Es unterscheidet sich von den vorhergehenden nicht nur durch die Art der neuen Distanzbildung, sondern auch durch die Vorgehensweise bei der Fusion von Gruppen. Der Abstand zwischen dem zuletzt gebildeten Cluster und den anderen Gruppen wird wie folgt berechnet (vgl. Abbildung 8.25):

Ward-Verfahren

$$\begin{aligned} D(R; P+Q) &= \tfrac{1}{NR+NP+NQ}\{(NR+NP)\cdot D(R,P) \\ &\quad +(NR+NQ)\cdot D(R,Q) - NR\cdot D(P,Q)\} \end{aligned} \tag{8.8}$$

Das Ward-Verfahren unterscheidet sich von den bisher dargestellten Linkage-Verfahren insbesondere dadurch, dass nicht diejenigen Gruppen zusammengefasst werden, die die geringste Distanz aufweisen, sondern es werden die Objekte (Gruppen) vereinigt, die ein vorgegebenes *Heterogenitätsmaß* am wenigsten vergrößern. Das Ziel des Ward-Verfahrens besteht folglich darin, jeweils diejenigen Objekte (Gruppen) zu vereinigen, die die Streuung (Varianz) in einer Gruppe möglichst wenig erhöhen. Dadurch werden möglichst homogene Cluster gebildet. Als Heterogenitätsmaß wird das *Varianzkriterium* verwendet, das auch als Fehlerquadratsumme bezeichnet wird.

Heterogenitätsmaß

Varianzkriterium

[16] Vgl. hierzu auch die Ausführungen in Abschnitt 8.2.2.3.

Fehlerquadratsumme

Die *Fehlerquadratsumme* (Varianzkriterium) errechnet sich für eine Gruppe g wie folgt:

$$V_g = \sum_{k=1}^{K_g} \sum_{j=1}^{J} (x_{kjg} - \bar{x}_{jg})^2 \tag{8.9}$$

mit

x_{kjg} = Beobachtungswert der Variablen j (j = 1,...,J) bei Objekt k (für alle Objekte k = 1,...,K_g in Gruppe g)

$\bar{x}_{jg}$ = Mittelwert über die Beobachtungserte der Variablen j in Gruppe g $\left(= 1/K_g \sum_{k=1}^{K_g} x_{kjg}\right)$

Wird dem Ward-Verfahren als Proximitätsmaß die quadrierte Euklidische Distanz zugrunde gelegt, so werden auch hier im ersten Schritt die quadrierten Euklidischen Distanzen zwischen allen Objekten berechnet. Somit hat auch das Ward-Verfahren für unser 5-Produkte-Beispiel die in Abbildung 8.17 berechnete Distanzmatrix als Ausgangspunkt. Da in Abbildung 8.17 noch keine Objekte vereinigt wurden, besitzt die Fehlerquadratsumme im ersten Schritt einen Wert von Null; d. h. jedes Objekt ist eine „eigenständige Gruppe“, und folglich tritt auch bei den Variablenwerten dieser Objekte noch keine Streuung auf. Das Zielkriterium beim Ward-Verfahren für die Zusammenfassung von Objekten (Gruppen) lautet nun:

„Vereinige diejenigen Objekte (Gruppen), die die

Fehlerquadratsumme am wenigsten erhöhen.“

Es lässt sich zeigen, dass die Werte der Distanzmatrix in Abbildung 8.17 (quadrierte Euklidische Distanzen) bzw. die mit Hilfe von Gleichung (8.8) berechneten Distanzen genau der *doppelten Zunahme der Fehlerquadratsumme* gemäß Gleichung (8.9) bei Fusionierung zweier Objekte (Gruppen) entsprechen.

Zusammenhang zwischen quadrierter Euklidischer Distanz und Fehlerquadratsumme

Dieser Zusammenhang lässt sich für das vorliegende Beispiel wie folgt verdeutlichen: Entsprechend Abbildung 8.17 sind im ersten Schritt die Objekte mit der kleinsten quadrierten Euklidischen Distanz zu vereinigen. Das sind in unserem Beispiel die Produkte „Rama“ und „Flora“, die eine quadrierte Euklidischen Distanz von 4 besitzen. Entsprechend des oben formulierten Zusammenhangs muss dieser Wert der doppelten Zunahme der Fehlerquadratsumme entsprechen bzw. die *Zunahme* der Fehlerquadratsumme beträgt nach Vereinigung dieser Produkte $1/2 \cdot 4 = 2$. Da die Fehlerquadratsumme im Ausgang Null war (es wurden zwei Objekte vereinigt), beträgt sie nach Vereinigung der Produkte „Rama“ und „Flora“ für diese neue Gruppe ebenfalls 2. Abbildung 8.32 verdeutlicht diesen Zusammenhang für unser Beispiel.

Dabei ist zu beachten, dass die Ausgangswerte für die Variablen „Preis“ und „Vitamingehalt“ bei Rama und Flora identisch sind (vgl. Abbildung 8.15), sodass sich die quadrierte Euklidische Distanz zwischen diesen beiden Objekten allein aufgrund der unterschiedlichen Werte der Variablen „Kaloriengehalt“ bestimmt. Für die quadrierte Euklidische Distanz folgt damit:

$$D(\text{Rama,Flora}) = (1-3)^2 = 4$$

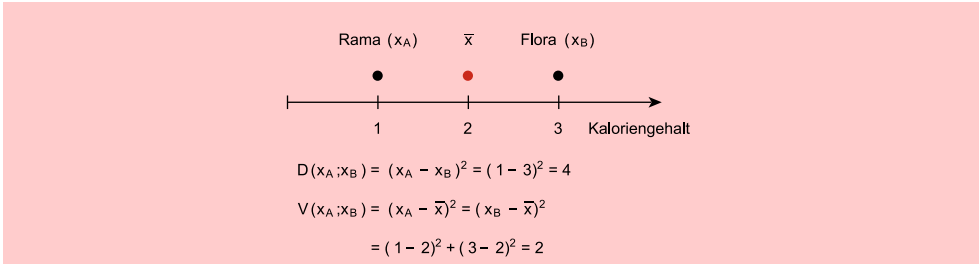


Abbildung 8.32: Zusammenhang zwischen quadrierter Euklidischer Distanz und Fehlerquadratsumme

Berücksichtigt man, dass der Mittelwert der Variablen „Kaloriengehalt“ $(1+3)/2 = 2$ beträgt, so ergibt sich für die Fehlerquadratsumme der Wert:

$$V(\text{Rama,Flora}) = (1-2)^2 + (3-2)^2 = 2$$

Im zweiten Schritt müssen nun die Distanzen zwischen der Gruppe „Rama, Flora“ und den verbleibenden Objekten gemäß Gleichung (8.8) bestimmt werden. Wir verwenden zu diesem Zweck die Distanzen aus Abbildung 8.17:

$$\begin{aligned} D(\text{Homa;Rama+Flora}) &= \frac{1}{3}\{(1+1)\cdot 6 + (1+1)\cdot 6 - 1\cdot 4\} &= 6,667 \\ D(\text{SB;Rama+Flora}) &= \frac{1}{3}\{(1+1)\cdot 56 + (1+1)\cdot 44 - 1\cdot 4\} &= 65,333 \\ D(\text{Butter;Rama+Flora}) &= \frac{1}{3}\{(1+1)\cdot 75 + (1+1)\cdot 59 - 1\cdot 4\} &= 88,000 \end{aligned}$$

Wir erhalten damit im zweiten Schritt die reduzierte Distanzmatrix im Ward-Verfahren, die ebenfalls die doppelte Zunahme der Fehlerquadratsumme bei Fusionierung zweier Objekte (Gruppen) enthält (Abbildung 8.33).

	Rama, Flora	Homa	SB
Homa	6,667		
SB	65,333	26	
Holländische Butter	88,000	41	11

Abbildung 8.33: Matrix der doppelten Heterogenitätszuwächse nach dem ersten Fusionsschritt beim Ward-Verfahren

Die doppelte Zunahme der Fehlerquadratsumme ist bei Hinzunahme von „Homa“ in die Gruppe „Rama, Flora“ am geringsten. In diesem Fall wird die Fehlerquadratsumme nur um $1/2 \cdot 6,667 = 3,333$ erhöht. Die gesamte Fehlerquadratsumme beträgt nach diesem Schritt:

$$V_g = 2 + 3,333 = 5,333,$$

wobei der Wert 2 die Zunahme der Fehlerquadratsumme aus dem ersten Schritt darstellt. Nach Abschluss dieser Fusionierung sind die Produkte „Rama“, „Flora“ und „Homa“ in einer Gruppe, und die Fehlerquadratsumme beträgt 5,333.

Dieser Wert lässt sich auch mit Hilfe von Gleichung (8.9) unter Verwendung der Ausgangsdaten in Abbildung 8.15 berechnen:

Wir müssen zu diesem Zweck zunächst die Mittelwerte für die Variablen „Kaloriengehalt“ (x_1), „Preis“ (x_2) und „Vitamingehalt“ (x_3) über die Objekte „Rama, Homa, Flora“ berechnen. Wir erhalten aus Abbildung 8.17:

$$\bar{x}_1 = 2; \qquad \bar{x}_2 = 2\frac{1}{3}; \qquad \bar{x}_3 = 1\frac{2}{3};$$

Nun bilden wir gemäß Gleichung (8.9) die quadrierten Differenzen zwischen den Beobachtungswerten (x_{kj}) einer jeden Variablen bei jedem Produkt und summieren diese Werte. Es folgt:

$$\begin{aligned}
V_g &= \underbrace{(1-2)^2 + (2-2\frac{1}{3})^2 + (1-1\frac{2}{3})^2}_{\text{Rama}} + \underbrace{(2-2)^2 + (3-2\frac{1}{3})^2 + (3-1\frac{2}{3})^2}_{\text{Homa}} \\
&\quad + \underbrace{(3-2)^2 + (2-2\frac{1}{3})^2 + (1-1\frac{2}{3})^2}_{\text{Flora}} \\
&= (-1)^2 + (-\frac{1}{3})^2 + (-\frac{2}{3})^2 + (0)^2 + (\frac{2}{3})^2 + (1\frac{1}{3})^2 + (1)^2 + (-\frac{1}{3})^2 + (-\frac{2}{3})^2 \\
&= 1 + \frac{1}{9} + \frac{4}{9} + 0 + \frac{4}{9} + \frac{16}{9} + 1 + \frac{1}{9} + \frac{4}{9} = 5\frac{3}{9} \\
&= 5,333
\end{aligned}$$

Im nächsten Schritt müssen nun die Distanzen zwischen der Gruppe „Rama, Flora, Homa“ und den verbleibenden Produkten bestimmt werden. Wir verwenden hierzu wiederum Gleichung (8.8) und die Ergebnisse aus Abbildung 8.33 des ersten Durchlaufs:

$$\begin{aligned}
D(\text{SB;Rama+Flora+Homa}) &= \frac{1}{4}\{(1+2)\cdot 65,333 + (1+1)\cdot 26 - 1\cdot 6,667\} \\
&= 60,333 \\
D(\text{Butter;Rama+Flora+Homa}) &= \frac{1}{4}\{(1+2)\cdot 88,000 + (1+1)\cdot 41 - 1\cdot 6,667\} \\
&= 84,833
\end{aligned}$$

Damit erhalten wir im zweiten Durchlauf beim Ward-Verfahren das in Abbildung 8.34 dargestellte Ergebnis. Es wird deutlich, dass die doppelte Zunahme in der Fehlerquadratsumme dann am kleinsten ist, wenn wir im nächsten Schritt die Objekte „SB“ und „Holländische Butter“ vereinigen. Die Fehlerquadratsumme erhöht sich dann nur um $1/2 \cdot 11 = 5,5$ und beträgt nach dieser Fusionierung:

$$V_g = 5,333 + 5,5 = 10,833$$

Der Wert 10,833 spiegelt dabei die Höhe der Fehlerquadratsumme nach Abschluss des dritten Fusionierungsschrittes wider. Entsprechend Gleichung (8.9) splittet sich der Gesamtwert korrekt in folgende zwei Einzelwerte auf: V(Rama, Flora, Homa) $= 5,333$ und V(SB, Holl. Butter) $= 5,5$.

	Rama, Flora, Homa	SB
SB	60,333	
Holländische Butter	84,833	11

Abbildung 8.34: Matrix der doppelten Heterogenitätszuwächse nach dem zweiten Fusionsschritt beim Ward-Verfahren

Werden im letzten Schritt die Gruppen „Rama, Flora, Homa“ und „SB, Holl. Butter“ vereinigt, so bedeutet das eine doppelte Zunahme der Fehlerquadratsumme um:

$$\begin{aligned} D(\text{Rama,Flora,Homa;Butter,SB}) &= \tfrac{1}{5}\{(3+1)\cdot 84,833 + (3+1)\cdot 60,333 \\ &\quad - 3\cdot 11\} \\ &= 109,533 \end{aligned}$$

Nach diesem Schritt sind *alle* Objekte in einem Cluster vereinigt, wobei das Varianzkriterium im letzten Schritt nochmals um $1/2 \cdot 109,533 = 54,767$ erhöht wurde. Die Gesamtfehlerquadratsumme beträgt somit im Endzustand $10,833 + 54,767 = 65,6$.

Der Fusionierungsprozess entsprechend dem Ward-Verfahren lässt sich zusammenfassend durch ein Dendrogramm wiedergeben, wobei nach jedem Schritt die Fehlerquadratsumme aufgeführt ist (Abbildung 8.35).

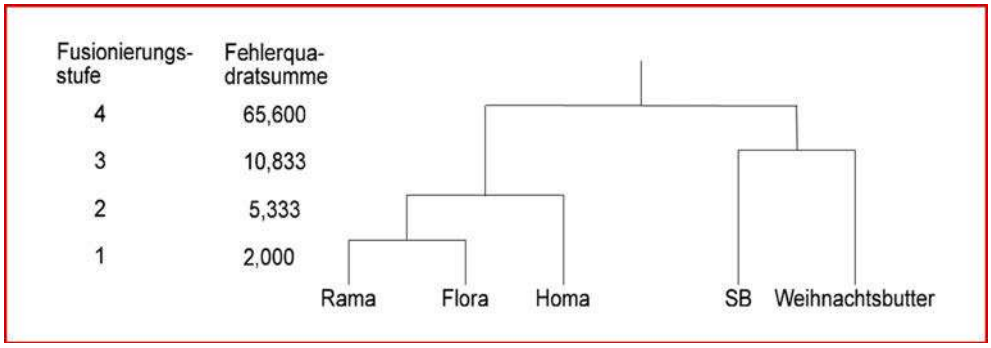


Abbildung 8.35: Dendrogramm für das Ward-Verfahren

8.2.2.3 Fusionierungseigenschaften ausgewählter Clusterverfahren

Die bisher betrachteten Clusterverfahren lassen sich bezüglich ihrer Fusionierungseigenschaften allgemein in dilatierende, kontrahierende und konservative Verfahren unterteilen.[17] *Dilatierende Verfahren* neigen dazu, die Objekte verstärkt in einzelne etwa gleich große Gruppen zusammenzufassen, während *kontrahierende Algorithmen* dazu tendieren, zunächst wenige große Gruppen zu bilden, denen viele kleine gegenüberstehen. Kontrahierende Verfahren sind damit geeignet, insbesondere „Ausreißer“ in einem Objektraum zu identifizieren. Weist ein Verfahren weder Tendenzen zur Dilatation noch zur Kontraktion auf, so wird es als *konservativ* bezeichnet. Daneben lassen sich Verfahren auch danach beurteilen, ob sie zur Kettenbildung neigen, d. h.

Dilatierende Clusterverfahren

Kontrahierende Clusterverfahren

[17] Vgl. auch Kohn (2005), S. 539 ff.; Steinhausen/Langer (1977), S. 75 ff.

Konservative Clusterverfahren

ob sie im Fusionierungsprozess primär einzelne Objekte aneinanderreihen und damit große Gruppen erzeugen. Schließlich kann noch danach gefragt werden, ob mit zunehmender Fusionierung das verwendete Heterogenitätsmaß monoton ansteigt oder ob auch ein Absinken des Heterogenitätsmaßes möglich ist. Betrachtet man die obigen Kriterien, so lassen sich die hier besprochenen Verfahren wie in Abbildung 8.36 gezeigt charakterisieren.

Fusionierungseigenschaften agglomerativer Clusterverfahren

Verfahren	Eigenschaft	Monoton?	Proximitätsmaße	Bemerkungen
Single-Linkage	kontrahierend	ja	alle	neigt zur Kettenbildung
Complete-Linkage	dilatierend	ja	alle	neigt zu kleinen Gruppen
Average-Linkage	konservativ	ja	alle	-
Centroid	konservativ	nein	Distanzmaße	-
Median	konservativ	nein	Distanzmaße	-
Ward	konservativ	ja	Distanzmaße	bildet etwa gleich große Gruppen

Abbildung 8.36: Charakterisierung agglomerativer Clusterverfahren

Vorteile des Ward-Verfahrens

Bezüglich des *Ward-Verfahrens* sei noch darauf hingewiesen, dass eine Untersuchung von Bergs gezeigt hat, dass das Ward-Verfahren im Vergleich zu anderen Algorithmen in den meisten Fällen *sehr gute Partitionen* findet und die Elemente „*richtig*“ den Gruppen zuordnet.[18] Das Ward-Verfahren kann somit als *sehr guter Fusionierungsalgorithmus* angesehen werden, wenn:[19]

- die Verwendung eines Distanzmaßes ein (inhaltlich) sinnvolles Kriterium zur Ähnlichkeitsbestimmung darstellt;
- alle Variablen auf metrischem Skalenniveau gemessen wurden;
- keine Ausreißer in einer Objektmenge enthalten sind bzw. vorher eliminiert wurden;
- die Variablen unkorreliert sind;
- zu erwarten ist, dass die Elementzahl in jeder Gruppe ungefähr gleich groß ist;
- die Gruppen in etwa die gleiche Ausdehnung besitzen.

Die drei letztgenannten Voraussetzungen beziehen sich auf die Anwendbarkeit des im Rahmen des Ward-Verfahrens verwendeten Varianzkriteriums (auch „Spur-W-Kriterium“ genannt). Allerdings neigt das Ward-Verfahren dazu, möglichst *gleich große Cluster* zu bilden und ist *nicht* in der Lage, langgestreckte Gruppen oder solche mit kleiner Elementzahl zu erkennen.

Für die Verfahren „Single-Linkage“, „Complete-Linkage“ und „Ward“ sollen abschließend deren zentrale Fusionierungseigenschaften anhand eines *fiktiven Beispiels* verdeutlicht werden.

Dabei kann der jeweilige Fusionierungsprozess der verschiedenen Verfahren mit Hilfe der zugehörigen Dendrogramme verdeutlicht werden. Bei den von SPSS erzeugten

[18] Vgl. Bergs (1981), S. 96 f.

[19] Vgl. hierzu auch die Untersuchungen von Milligan (1980), S. 332 ff. oder die in der Studienübersicht von Punj/Stewart (1983), S. 141 ff. zusammengetragenen Ergebnisse, sowie Bacher/Pöge/Wenzig (2010), S. 166 ff.

Dendrogrammen ist zu beachten, dass SPSS die Heterogenitätsentwicklung immer auf eine Skala von 0 bis 25 normiert, wobei dem Endstadium des Fusionierungsprozesses (alle Fälle befinden sich in einem Cluster) der Heterogenitätswert von 25 zugewiesen wird. Aus den so erstellten Dendrogrammen lassen sich dann bereits optisch sinnvolle Gruppentrennungen erkennen: So macht z. B. das Dendrogramm in Abbildung 8.40 deutlich, dass sich bei einem normierten Heterogenitätsmaß von 4 eine Vier-Cluster-Lösung und bei einem Heterogenitätsmaß von ca. 6 eine Drei-Cluster-Lösung herausbildet. Die Zwei-Cluster-Lösung entsteht erst bei einer Clusterdistanz von ca. 10.

In unserem fiktiven Beispiel werden 56 Fälle betrachtet, die jeweils durch zwei Variable beschrieben werden und in Abbildung 8.37 graphisch verdeutlicht sind. Die *Beispieldaten* wurden so gewählt, dass bereits optisch drei Gruppen erkennbar sind: Gruppe A besteht aus 15 Fällen, Gruppe B aus 20 Fällen und Gruppe C aus 15 Fällen. Darüber hinaus treten sechs Ausreißer auf, die jeweils durch einen Stern markiert wurden.

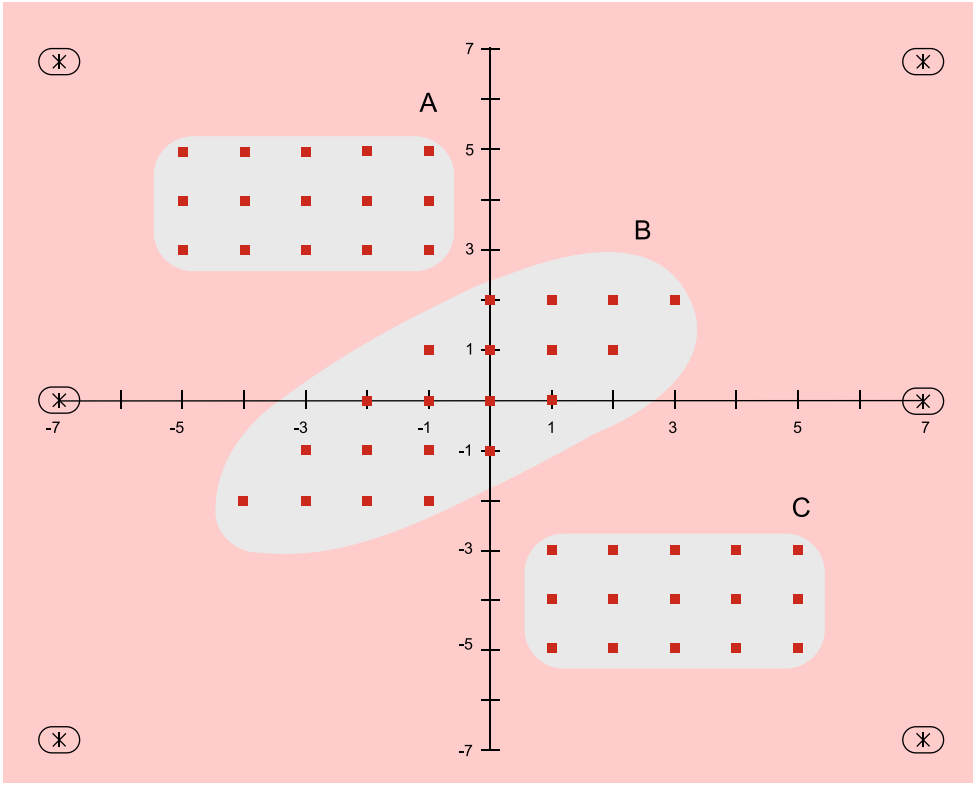


Abbildung 8.37: Beispieldaten zur Verdeutlichung der Fusionierungseigenschaften

Wird auf die Beispieldaten in Abbildung 8.37 zunächst das *Single-Linkage-Verfahren* angewandt, so lässt das entsprechende Dendrogramm in Abbildung 8.38 deutlich die Neigung dieses Verfahrens zur Kettenbildung erkennen. Während die Objekte der drei Gruppen quasi auf der gleichen Stufe zusammengefasst werden, werden die als Ausreißer gekennzeichneten Objekte erst am Ende des Prozesses fusioniert. Damit ist auch klar erkennbar, dass sich das Single-Linkage-Verfahren in besonderem Maße dazu eignet, „Ausreißer“ in einer Objektmenge zu erkennen. Wen-

Single-Linkage-Verfahren

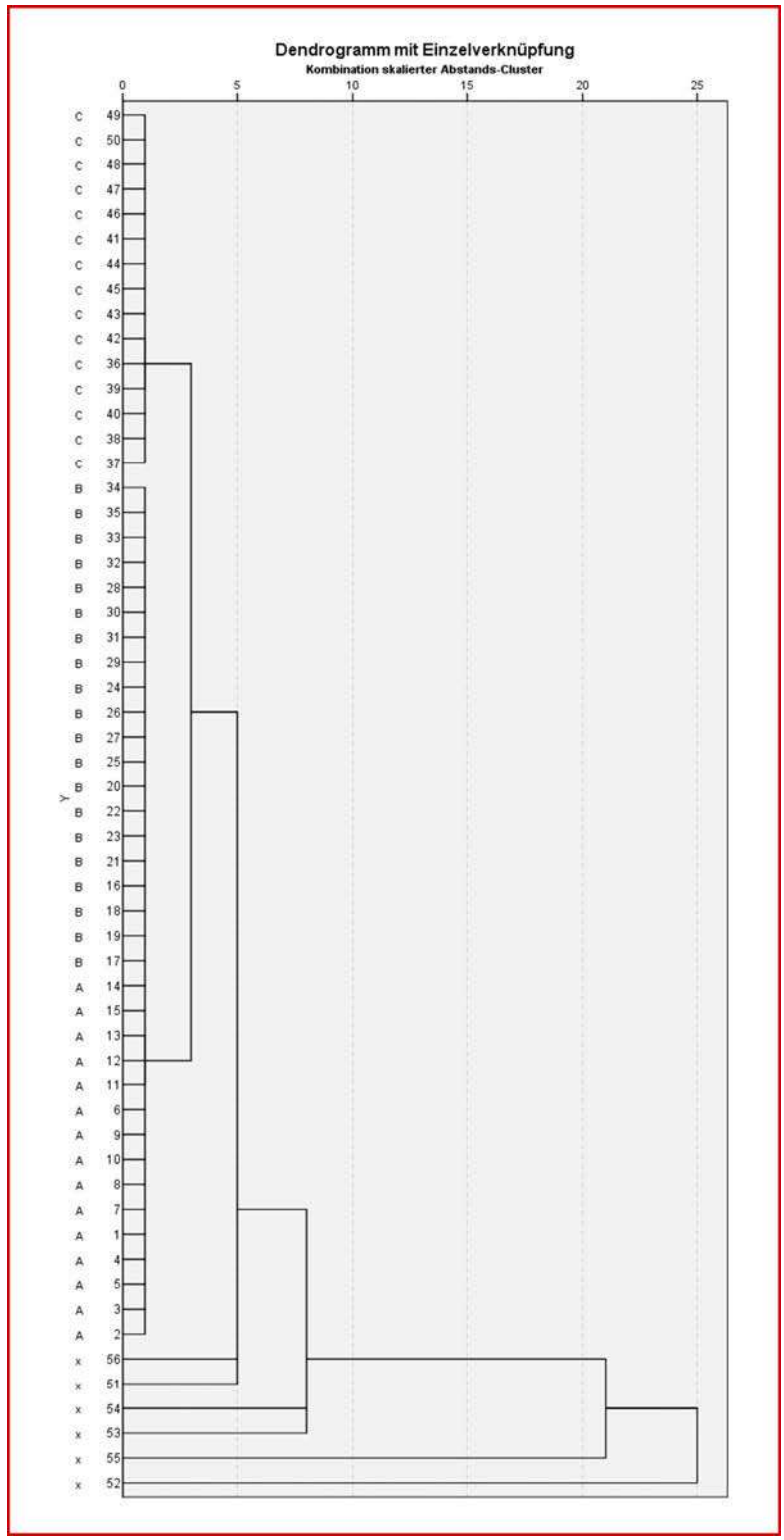


Abbildung 8.38: Dendrogramm des Single-Linkage-Verfahrens zur Verdeutlichung der Fusionierungseigenschaften

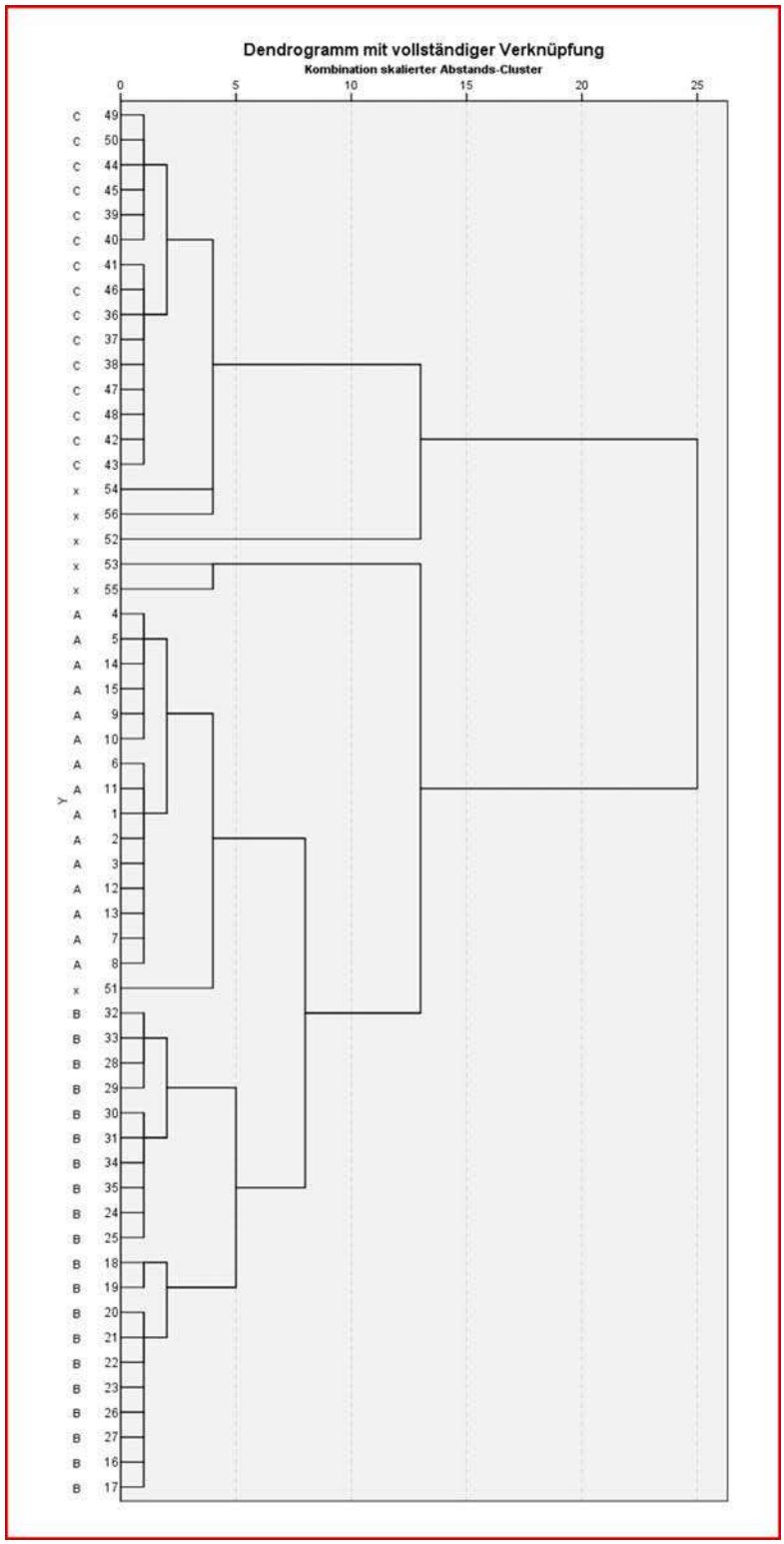


Abbildung 8.39: Dendrogramm des Complete-Linkage-Verfahrens zur Verdeutlichung der Fusionierungseigenschaften

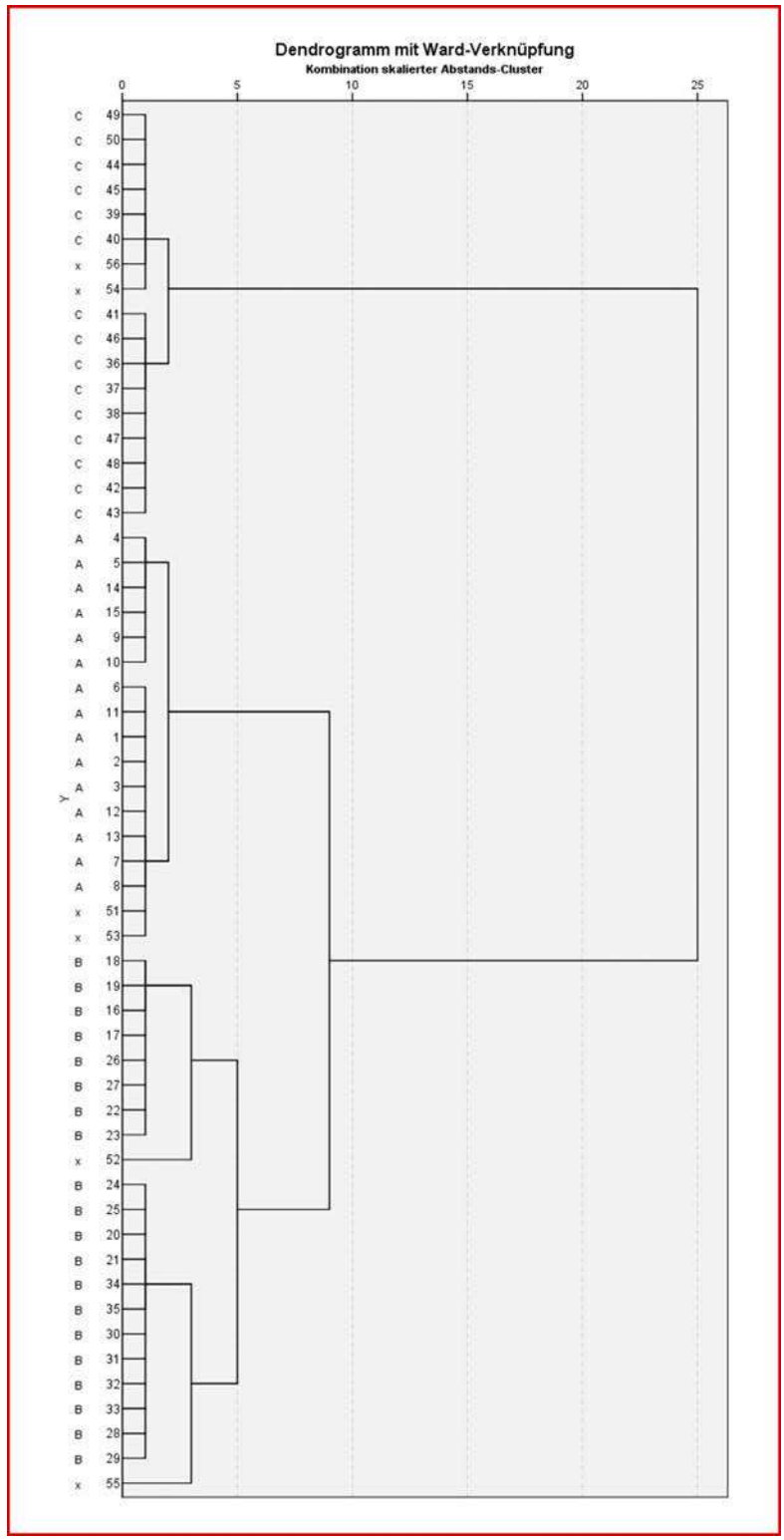


Abbildung 8.40: Dendrogramm des Ward-Verfahrens zur Verdeutlichung der Fusionierungseigenschaften

det man hingegen auf die Daten aus Abbildung 8.37 das Complete-Linkage- und das Ward-Verfahren an, so sind im Vergleich zum Single-Linkage-Verfahren deutlich unterschiedliche Fusionierungsverläufe erkennbar. Abbildung 8.39 lässt für das *Complete-Linkage-Verfahren* zwar eine klare 3-Cluster-Lösung erkennen, jedoch wird nur die Gruppe C exakt isoliert, während die Gruppe B nur teilweise separiert und die überwiegende Zahl der Elemente aus B mit den Objekten aus Gruppe A zusammengefasst wird. Das Complete-Linkage-Verfahren ist damit nicht in der Lage, die „wahre Gruppierung“ entsprechend Abbildung 8.37 zu reproduzieren.

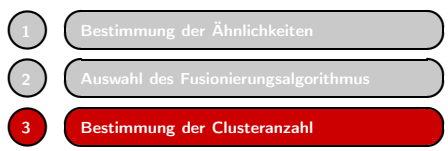
Complete-Linkage-Verfahren

Demgegenüber zeigt das Dendrogramm in Abbildung 8.40, dass das *Ward-Verfahren* die „wahre Gruppierung“ gemäß Abbildung 8.37 erzeugen kann, wobei sich die Ausreißer auf die 4-Cluster-Lösung verteilen. Damit werden auch die Untersuchungen von Bergs bestätigt, wonach das Ward-Verfahren sehr gut in der Lage ist, Objekte zu den „wahren Gruppen“ zusammenzufassen.

Ward-Verfahren

Aus den dargestellten Zusammenhängen lässt sich abschließend die Empfehlung ableiten, dass bei praktischen Anwendungen eine Objektmenge zunächst mit Hilfe des Single-Linkage-Verfahrens auf Ausreißer untersucht werden sollte. Anschließend sind die gefundenen „Ausreißer-Objekte“ zu eliminieren, und die reduzierte Objektmenge ist dann mit Hilfe eines anderen agglomerativen Verfahrens zu gruppieren, wobei die Auswahl des Verfahrens vor dem Hintergrund der jeweiligen Anwendungssituation zu erfolgen hat.

8.2.3 Bestimmung der optimalen Clusterzahl



Die bisherigen Ausführungen haben gezeigt, nach welchen Kriterien verschiedene Clusteranalysealgorithmen eine Fusionierung von Einzelobjekten zu Gruppen vornehmen. Dabei gehen alle *agglomerativen Verfahren* von der feinsten Partition (alle Objekte bilden jeweils ein eigenständiges Cluster) aus und enden mit einer Zusammenfassung aller Objekte in einer großen Gruppe. Der Anwender muss deshalb im dritten Schritt entscheiden, welche Anzahl von Gruppen (Cluster-Lösung) als die „beste“ anzusehen ist. I. d. R. hat der Anwender keine sachlogisch begründbaren Vorstellungen zur Gruppierung der Untersuchungsobjekte und versucht deshalb mit Hilfe der Clusteranalyse eine den Daten inhärente Gruppierung aufzudecken. Vor diesem Hintergrund sollte sich auch die Bestimmung der Clusterzahl an *statistische Kriterien* orientieren und *nicht* sachlogisch (im Hinblick auf den Gruppen zugeordneten Fällen) begründet werden. Bei der Entscheidung über die Clusterzahl besteht immer ein *Zielkonflikt* zwischen der „Homogenitätsanforderung an die Cluster-Lösung“ und der „Handhabbarkeit der Cluster-Lösung“. Zur Lösung dieses Konflikts können auch sachlogische Überlegungen herangezogen werden, die sich allerdings nur auf die Anzahl der zu wählenden Cluster beziehen und *nicht* an den in den Clustern zusammengefassten Fällen ausgerichtet sein sollten. Im Folgenden werden verschiedene Möglichkeiten zur *Bestimmung der optimalen Clusterzahl* besprochen.[20]

Bestimmung der optimalen Clusterzahl

[20] Da in SPSS bisher keine Kriterien zur Bestimmung der optimalen Clusterzahl verfügbar sind, wird empfohlen ggf. auf alternative Programme wie S-Plus, R oder SAS und das hier verfügbare Cubic Clustering Criterion (CCC) zurückzugreifen.

8.2.3.1 Analyse der Zuordnungsübersicht und Elbow-Kriterium

Einen ersten Anhaltspunkt zur Bestimmung der Clusterzahl liefert die (optische) Identifikation eines „Sprungs“ (Elbow) in der Veränderung des Heterogenitätsmaßes alternativer Cluster-Lösungen. So zeigt die *Zuordnungsübersicht* in Abbildung 8.41 für die insgesamt 56 Fälle unserer Beispieldaten aus Abbildung 8.37, auf welcher Fusionierungsstufe (Spalte 1) welche Fälle bzw. Cluster (Spalte 2 ‚zusammengeführte Cluster') durch das Ward-Verfahren bei welchem Heterogenitätsmaß zusammengefasst wurden. Als Heterogenitätsmaß diente dem Ward-Verfahren dabei die Fehlerquadratsumme, deren Entwicklung in der Spalte ‚Koeffizienten' aufgezeigt ist. Eine graphische Verdeutlichung des Fusionierungsprozesses liefert auch das zugehörige *Dendrogramm*, das für die Beispieldaten und das Ward-Verfahren bereits in Abbildung 8.40 dargestellt wurde.

Zuordnungsübersicht

Schritt	Zusammengeführte Cluster		Koeffizienten	Erstes Vorkommen des Clusters		Nächster Schritt
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	49	50	,500	0	0	25
2	47	48	1,000	0	0	26
3	41	46	1,500	0	0	29
4	44	45	2,000	0	0	25
5	42	43	2,500	0	0	26
...	...	...	...	...	...	...
50	36	39	229,686	40	47	55
51	16	52	288,464	42	0	53
52	20	55	349,823	45	0	53
53	16	20	487,807	51	52	54
54	1	16	723,997	49	53	55
55	1	36	1428,214	54	50	0

Abbildung 8.41: Zuordnungsübersicht des Ward-Verfahrens für die Beispieldaten (Ausschnitt)

Elbow-Kriterium

Darüber hinaus ist es aber auch hilfreich, die in der Zuordnungsübersicht aufgezeigte Heterogenitätsentwicklung gegen die zugehörige Clusterzahl in einem Koordinatensystem abzutragen. Zeigt sich in diesem Diagramm (sog. *Scree-Plot*) ein „*Ellbogen*“ (Elbow) in der Entwicklung des Heterogenitätsmaßes, so kann dieser als Entscheidungskriterium für die zu wählende Clusteranzahl verwendet werden. Wir sprechen in diesem Fall auch von dem sog. *Elbow-Kriterium* als Entscheidungshilfe. Für die 56 Beispieldaten ergibt sich der in Abbildung 8.42 dargestellte Scree-Plot, wobei hier nach dem Elbow-Kriterium eine Vier-Cluster-Lösung zu wählen wäre. Da das Elbow-Kriterium auch eine optische Unterstützung bei der Clusterentscheidung liefert, sollte bei der Konstruktion des entsprechenden Diagramms die Ein-Cluster-Lösung *nicht* berücksichtigt werden. Der Grund für diese Empfehlung ist darin zu sehen, dass beim Übergang von der Zwei- zur Ein-Cluster-Lösung immer der größte Heterogenitätssprung zu verzeichnen ist und sich bei dessen Berücksichtigung bei nahezu allen Anwendungsfällen ein Elbow herausbildet.

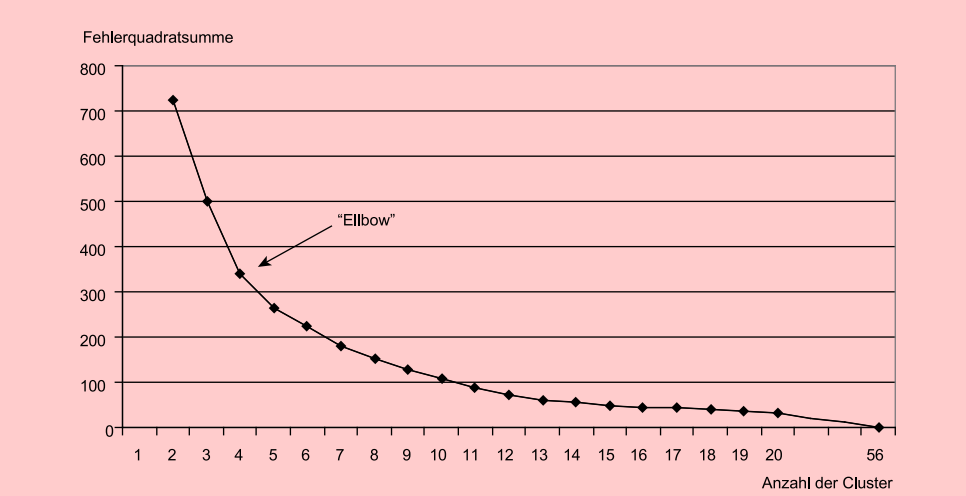


Abbildung 8.42: Elbow-Kriterium zur Bestimmung der Clusteranzahl (Scree-Plot)

8.2.3.2 Stopping Rule von Calinski/Harabasz

Da das Elbow-Kriterium stark von der subjektiven Einschätzung des Anwenders abhängt, wurde in der Literatur zusätzlich eine Vielzahl statistischer Kriterien (sog. *Stopping-Rules*) entwickelt, die statistische und damit weitgehend objektive Anhaltspunkte zur Bestimmung der optimalen Clusterzahl bei Anwendung der hierarchischen Clusteranalyse liefern.[21] Im Rahmen einer umfangreichen Simulationsstudie haben z. B. Milligan/Cooper insgesamt 30 dieser Stopping Rules im Rahmen einer umfangreichen Simulationsstudie getestet. Von den Autoren wurden unterschiedlich trennscharfe Cluster-Lösungen (mit 2 bis 5 Clustern) vorgegeben und anschließend unter Rückgriff auf das Single-, Complete-, Average-Linkage und Ward-Verfahren getestet, inwieweit die einzelnen Verfahren in der Lage sind, die „wahre“ Gruppenzahl zu identifizieren. Im Ergebnis wurde das *Kriterium von Calinski/Harabasz* als beste Stopping Rule identifiziert, da sie in über 90 % der untersuchten Fälle die wahre Gruppenstruktur aufdecken konnte.

Stopping rules

Calinski-Harabasz-Kriterium

Das Kriterium nach Calinski/Harabasz, welches für metrische Merkmale geeignet ist, nimmt in Analogie zur Varianzanalyse eine Gegenüberstellung der Inner-Gruppen-Streuung (W) und der Zwischen-Gruppen-Streuung (B) vor.[22] Dabei wird für jede Cluster-Lösung mit K Gruppen die nachfolgende CH-Statistik berechnet, wobei $\bar{y}_k$ der Mittelwert der Gruppe k und $\bar{y}$ der Mittelwert aller Beobachtungen ist. Es gilt:

$$CH(K) = \frac{tr(B)}{tr(W)} \frac{n-K}{K-1} \tag{8.10}$$

$$B = \sum_{k=1}^{K} n_k (\bar{y}_k - \bar{y})(\bar{y}_k - \bar{y})'$$

$$W = \sum_{k=1}^{K} \sum_{j=1}^{n_k} n_k (y_{kj} - \bar{y}_k)(y_{kj} - \bar{y}_k)'$$

[21] Vgl. zu einem Überblick: Milligan/Cooper (1985), 163 ff.
[22] Vgl. Calinski/Harabasz (1974), passim.

Sofern ein Maximum von CH an der Stelle K existiert, welches innerhalb der möglichen Gruppenbereiche von 2 und (N-1) liegt, so deutet dies darauf hin, dass K Gruppen innerhalb der Daten existieren. Fällt der Wert der CH-Statistik hingegen monoton mit zunehmender Gruppenzahl, so ist dies ein Indiz dafür, dass keine Gruppenstruktur vorliegt. Nimmt der Wert der Teststatistik demgegenüber mit steigendem K zu, so spricht dies für einen hierarchisch strukturierten Datensatz.

8.2.3.3 Test von Mojena

Test von Mojena

In der Studie von Milligan/Cooper wurde weiterhin der *Test von Mojena* zu den besten 10 Verfahren zur Ermittlung der Clusterzahl ermittelt. Da sich dieser Test relativ einfach mit Hilfe einer Tabellenkalkulation selbst durchführen lässt, sei er im Folgenden ebenfalls kurz dargestellt: Ausgangspunkt dieses Tests sind die standardisierten Fusionskoeffizienten ($\tilde{\alpha}$) je Fusionsstufe, die folgendermaßen berechnet werden:[23]

$$\bar{\alpha} = \frac{1}{n-1}\sum_{i=1}^{n-1}\alpha_i \quad ... \quad s_\alpha = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n-1}(\alpha_i - \bar{\alpha})^2} \quad ... \quad \widetilde{\alpha}_i = \frac{\alpha_i - \bar{\alpha}}{s_\alpha} \tag{8.11}$$

Als Indikator für eine gute Cluster-Lösung gilt die größte Gruppenzahl, bei der ein vorgegebener Wert des standardisierten Fusionskoeffizienten zum ersten Mal überschritten wird. In der Literatur bestehen hierfür verschiedene Maßgaben: So erzielt Mojena in seiner Simulationsstudie mit einem Schwellwert von 2,75 die besten Ergebnisse, wohingegen die Studien von Milligan/Cooper für einen Wert von 1,25 sprechen, wobei hier die Ergebnisgüte nur geringfügige Variationen in einem Wertebereich von 1 bis 2 aufweist. Eine endgültige Empfehlung kann jedoch nicht ausgesprochen werden, da der optimale Parameter stark von der vorliegenden Datenstruktur abhängt. Nach eigenen Studien der Autoren dieses Buches zufolge erscheinen Werte im Bereich von 1,8 bis 2,7 für die meisten Datenkonstellationen gut geeignet.

8.3 Fallbeispiel

8.3.1 Problemstellung

Ein Margarinehersteller ist an den Konsumentenbeurteilungen von elf Emulsionsfetten (Butter und Margarine) im Hinblick auf zehn Eigenschaften interessiert. Im Einzelnen handelte es sich um die in Abbildung 8.43 aufgeführten Marken und Eigenschaften.

Die Eigenschaftsbeurteilung erfolgte durch 32 *Probanden*, die gebeten wurden, jede Marke einzeln nach diesen Eigenschaften auf einer siebenstufigen Intervallskala zu beurteilen. Man erhielt somit eine dreidimensionale Matrix (32 x 11 x 10) mit 3.520 metrischen Eigenschaftsurteilen. Da die Algorithmen der Clusteranalyse lediglich *zweidimensionale Matrizen* verarbeiten können, wurde aus den 32 Urteilen pro Eigenschaft das *arithmetische Mittel* berechnet, sodass wir für die nachfolgenden Betrachtungen eine 11x10-Matrix heranziehen, mit den 11 Emulsionsfetten als Fälle und den 10 Eigenschaftsurteilen als Variablen. Bei einer solchen Durchschnittsbildung ist allerdings zu beachten, dass bestimmte Informationen (nämlich die über die Streuung der Ausprägungen zwischen den Personen) verloren gehen.

[23] Vgl. Mojena (1977), S. 359 ff.

Emulsionsfette (Butter und Margarine) M_k $(k = 1, ..., 11)$		Merkmalsvariable: x_j $(j = 1, ..., 10)$ (subjektive Beurteilungen)	
1	Sanella	1	Streichfähigkeit
2	Homa	2	Preis
3	SB	3	Haltbarkeit
4	Delicado	4	Anteil ungesättigter Fettsäuren
5	Holländische Markenbutter	5	Back- und Brateignung
6	Weihnachtsbutter	6	Geschmack
7	Du darfst	7	Kaloriengehalt
8	Becel	8	Anteil tierischer Fette
9	Botteram	9	Vitamingehalt
10	Flora	10	Natürlichkeit
11	Rama		

Abbildung 8.43: Untersuchte Marken und Variablen im Fallbeispiel

Bei den meisten Anwendungen im Rahmen der Clusteranalyse wird jedoch *keine* Durchschnittsbildung vorgenommen und im Ausgang die Rohdatenmatrix betrachtet. Dabei können Probleme insbesondere dadurch entstehen, dass einzelnen Variablen bei bestimmten Fällen kein Wert zugewiesen wurde (zum Problem der missing values siehe Abschnitt 8.4.2).

Missing Values

Der Margarinehersteller möchte im ersten Schritt die elf Emulsionsfette (Butter und Margarine) auf Ausreißer untersuchen und in Abhängigkeit der dabei erzielten Ergebnisses ggf. die Objektzahl um mögliche Ausreißer reduzieren und dann mit dem Ward-Verfahren eine hierarchische Clusteranalyse durchführen.

Analyse mit Hilfe von SPSS

Zur Analyse von Ausreißern unter den Objekten ist in besonderer Weise das Single-Linkage-Verfahren geeignet (vgl. Abschnitt 8.2.2.2.2). Zur Durchführung einer hierarchischen Clusteranalyse ist in SPSS unter dem Hauptmenü „Analysieren“ der Unterpunkt „Klassifizieren“ und dort die Prozedur „Hierarchische Clusteranalyse“ aufzurufen (vgl. Abbildung 8.44).

Hierarchische Clusteranalyse

Im erscheinenden Dialogfeld „Hierarchische Clusteranalyse“ sind die zehn metrisch skalierten Variablen zur Beschreibung der Emulsionsfette aus der Liste auszuwählen und in das Feld „Variable(n)“ zu übertragen. Die Variable „Marke“ dient der Beschreibung der elf Emulsionsfette und wird in das Feld „Fallbeschriftung“ übernommen (vgl. Abbildung 8.45).

Über den Unterpunkt „Statistiken“ können eine Zuordnungsübersicht und die Ähnlichkeitsmatrix der elf Objekte auf Basis des gewählten Proximitätsmaßes angefordert werden. Weiterhin kann hier ein Bereich von Lösungen angegeben werden (z. B. 2 bis 5 Clusterlösung), für den dann die jeweilige Clusterzugehörigkeit der Objekte ausgegeben wird. Mit dem Unterpunkt „Diagramme“ können graphische Darstellungen zum Fusionierungsverlauf angefordert werden, wobei im Fallbeispiel die Option „Dendrogramm“ gewählt wurde (vgl. Abbildung 8.46).

Mit Hilfe des Unterpunktes „Methode“ werden die Clustermethode (Fusionierungsalgorithmus) und das Proximitätsmaß (Maß) bestimmt, wobei unterschiedliche Maße

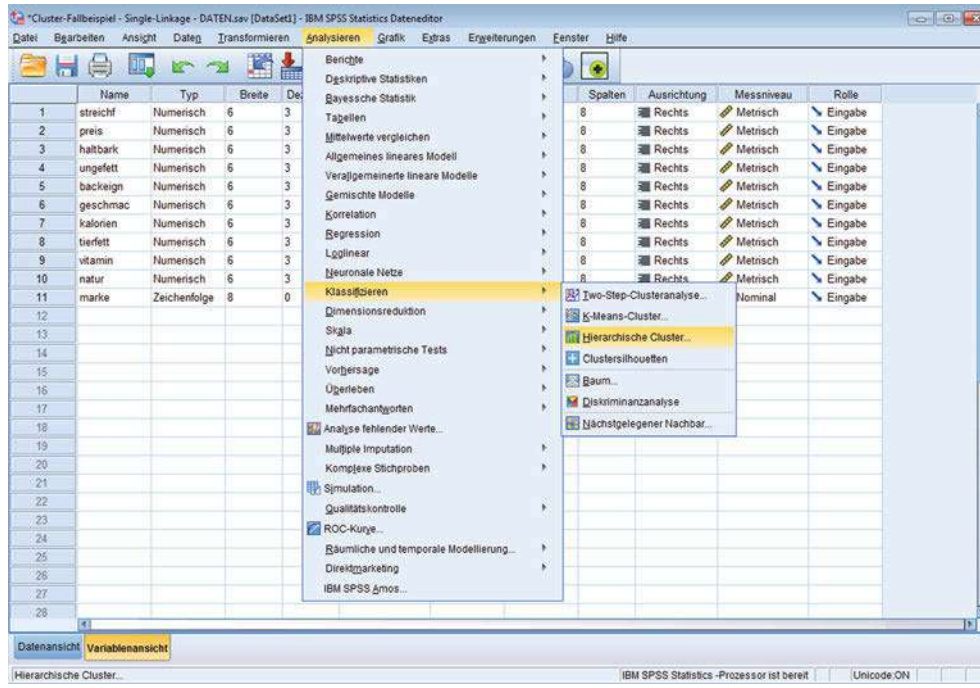


Abbildung 8.44: Daten-Editor mit Auswahl des Analyseverfahrens „Hierarchische Clusteranalyse“

Abbildung 8.45: Dialogfeld der Prozedur „Hierarchische Clusteranalyse“

für metrisch (intervall) skalierte, nominal skalierte (Häufigkeiten) und binär kodierte Variable zur Verfügung stehen (vgl. Abbildung 8.47). Im Fallbeispiel wurden zunächst die elf Objekte zum Zwecke der Ausreißeranalyse mit dem Single-Linkage-Verfahren (*Nächstgelegener Nachbar*) untersucht. Anschließend wurde das Objekt „Delicado“ eliminiert und die verbleibenden zehn Marken mit der *Ward-Methode* geclustert. In beiden Fällen wurde die *Quadrierte euklidische Distanz* als Proximitätsmaß verwendet. Abbildung 8.47 zeigt das Dialogfeld „Methode“ und die dort zur Verfügung stehenden

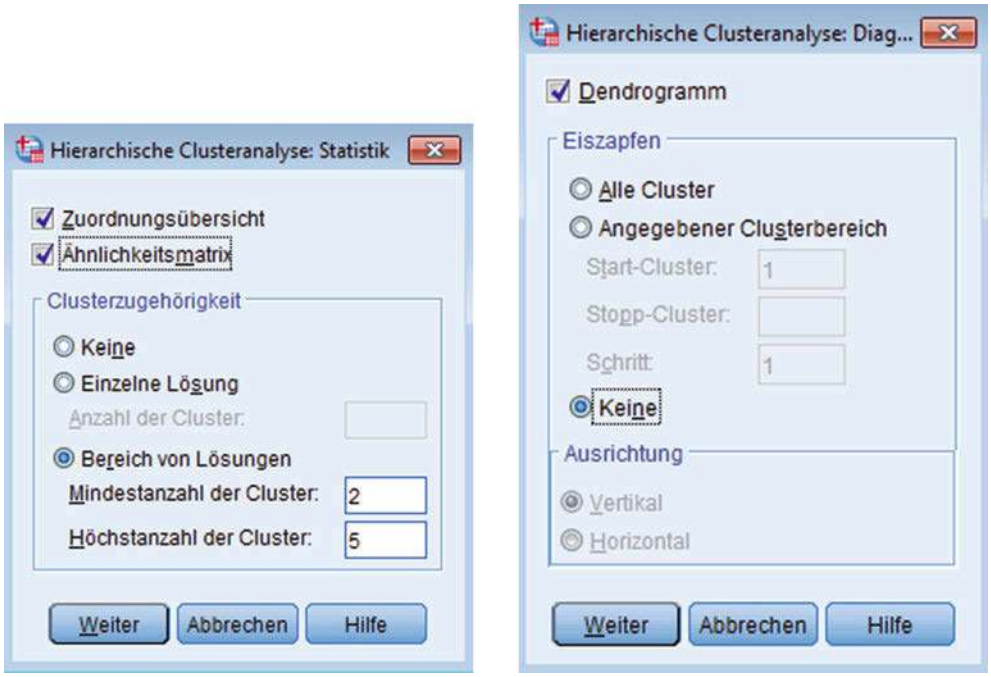


Abbildung 8.46: Dialogfelder „Statistiken“ und „Diagramme“

Abbildung 8.47: Dialogfeld „Methode“ mit Clustermethoden und Proximitätsmaßen

Clustermethoden sowie die für metrisch skalierte Variable zur Verfügung stehenden Proximitätsmaße.

Nach erfolgten Einstellungen in den Untermenüs gelangt der Anwender durch den Button „Weiter“ jeweils wieder zurück zur Prozedur „Hierarchische Clusteranalyse“, und die Durchführung der Analyse kann durch Drücken von „OK“ gestartet werden.

8.3.2 Ergebnisse

Ausreißeranalyse mittels Single Linkage-Verfahren (Nächstgelegener Nachbar)

Ausreißeranalyse mit dem Single-Linkage-Verfahren

Die sich nach Durchführung des Single-Linkage-Verfahrens ergebende Abstandsmatrix ist in Abbildung 8.48 dargestellt. Die Überschrift der Matrix lautet „Näherungsmatrix“, und in der Fußnote wird der Hinweis gegeben: „Dies ist eine Unähnlichkeitsmatrix“.

Es wird deutlich, dass die größte Unähnlichkeit zwischen „Homa“ und „Weihnachtsbutter“ besteht; der Wert der quadrierten euklidischen Distanz beträgt hier 38,621. Die geringsten Distanzen weisen „SB“, „Botteram“ und „Rama“ auf. Das zugehörige Dendrogramm ist in Abbildung 8.49 dargestellt. Im Dendrogramm wird das verwendete Heterogenitätsmaß auf das Intervall [0;25] normiert und so die Zusammenfassungen der Objekte bzw. Cluster graphisch verdeutlicht. Es ist erkennbar, dass die Marke „Delicado“ erst am Ende des Fusionierungsprozesses auf einem normierten Heterogenitätslevel von ca. 13 mit dem Cluster „Holländische Butter; Weihnachtsbutter“ zusammengefasst wird. Delicado wird deshalb hier als Ausreißer interpretiert und aus der Analyse ausgeschlossen.

Clusterung mit Hilfe der Ward-Methode

Nach Ausschluss der Marke „Delicado“ wurden die verbleibenden 10 Marken mit Hilfe der Ward-Methode analysiert. Der Verlauf des Fusionierungsprozesses wird durch die sog. *Zuordnungsübersicht* verdeutlicht, die in Abbildung 8.50 wiedergegeben ist. Die Tabelle wie folgt zu lesen:

Ward-Verfahren

Interpretation der Zuordnungsübersicht

In der Spalte „Schritt“ wird die jeweilige *Fusionierungsstufe* angegeben. Es gibt insgesamt immer genau einen Schritt weniger als Objekte existieren. Die Spalte „Zusammengeführte Cluster“ gibt unter den Überschriften „Cluster 1“ und „Cluster 2“ die Nummer der im jeweiligen Schritt fusionierten Objekte bzw. Cluster an, und in der Spalte „Koeffizienten“ steht der jeweilige *Wert des verwendeten Heterogenitätsmaßes* (hier: Varianzkriterium) am Ende eines Fusionierungsschrittes. Die zu einem Cluster zusammengefassten Objekte bzw. Cluster erhalten als neue Identifikation immer die Nummer des zuerst genannten Objektes (Clusters). In der Spalte „Erstes Vorkommen des Clusters“ wird jeweils der Fusionierungsschritt angegeben, bei dem das jeweilige Objekt (Cluster) *erstmals* in dieser Form zur Fusionierung herangezogen wurde. Die Spalte „Nächster Schritt“ zeigt schließlich an, auf welcher Stufe die gebildete Gruppe zum *nächsten Mal* in den Fusionierungsprozess einbezogen wird. So wird z. B. im 7. Schritt das Cluster 1, das in dieser Form bereits im vierten Schritt gebildet wurde, mit dem Objekt 2 bei einem Heterogenitätsmaß von 12,702 vereinigt. Die sich dabei ergebende Gruppe erhält die Kennung „1“ und wird im 8. Schritt wieder zur Fusionierung herangezogen. Insgesamt macht Abbildung 8.50 deutlich, dass bei den ersten vier Fusionierungsschritten die Marken „SB (3), Botteram (8), Rama (10), Flora (9) und Sanella (1)“ vereinigt werden, wobei die Fehlerquadratsumme nach der vierten Stufe 4,220 beträgt, d. h., dass die Varianz der Variablenwerte in dieser Gruppe also noch relativ gering ist.

Dendrogramm

Eine graphische Verdeutlichung des Fusionierungsprozesses liefert das in Abbildung 8.51 dargestellte Dendrogramm.

Die Werte in Abbildung 8.50 geben einen *ersten Anhaltspunkt*, wie viele Cluster als endgültige Lösung heranzuziehen sind. Da vom 8. zum 9. Fusionierungsschritt (bzw. von der Zwei- zur Ein-Cluster-Lösung) ein enormer Zuwachs der Fehler-

Abbildung 8.48: Abstandsmatrix der Quadrierten euklidischen Distanzen für die elf Emulsionsfette

Näherungsmatrix

Quadriertes euklidisches Distanzmaß

Fall	1:SANELLA	2:HOMA	3:SB	4:DELICADO	5:HOLLBUTT	6:WEIHBUTT	7:DUDARFST	8:BECEL	9:BOTTERAM	10:FLORA	11:RAMA
1:SANELLA	,000	3,792	3,794	15,198	21,442	25,484	4,882	6,025	2,268	2,909	2,113
2:HOMA	3,792	,000	6,322	23,871	30,458	38,621	10,881	8,063	5,325	6,194	3,396
3:SB	3,794	6,322	,000	14,151	24,971	28,933	3,998	3,471	1,099	2,361	1,725
4:DELICADO	15,198	23,871	14,151	,000	6,496	11,882	11,692	18,362	15,929	16,520	17,030
5:HOLLBUTT	21,442	30,458	24,971	6,496	,000	3,606	16,410	26,957	25,334	25,906	26,768
6:WEIHBUTT	25,484	38,621	28,933	11,882	3,606	,000	15,887	32,336	29,999	28,195	32,272
7:DUDARFST	4,882	10,881	3,998	11,692	16,410	15,887	,000	6,422	5,156	3,825	6,932
8:BECEL	6,025	8,063	3,471	18,362	26,957	32,336	6,422	,000	3,395	6,376	6,022
9:BOTTERAM	2,268	5,325	1,099	15,929	25,334	29,999	5,156	3,395	,000	1,564	1,118
10:FLORA	2,909	6,194	2,361	16,520	25,906	28,195	3,825	6,376	1,564	,000	2,152
11:RAMA	2,113	3,396	1,725	17,030	26,768	32,272	6,932	6,022	1,118	2,152	,000

Dies ist eine Unähnlichkeitsmatrix

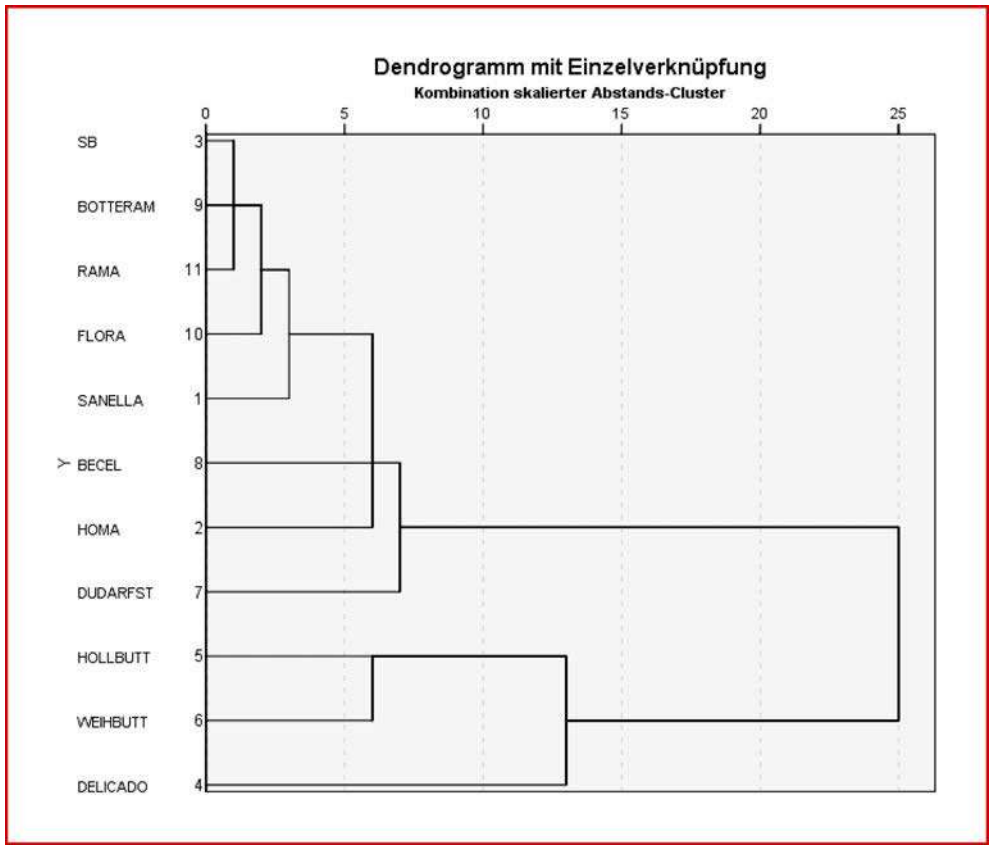


Abbildung 8.49: Dendrogramm für das Single-Linkage-Verfahren

Zuordnungsübersicht

Schritt	Zusammengeführte Cluster Cluster 1	Cluster 2	Koeffizienten	Erstes Vorkommen des Clusters Cluster 1	Cluster 2	Nächster Schritt
1	3	8	,549	0	0	2
2	3	10	1,314	1	0	3
3	3	9	2,505	2	0	4
4	1	3	4,220	0	3	7
5	4	5	6,023	0	0	9
6	6	7	9,234	0	0	8
7	1	2	12,702	4	0	8
8	1	6	17,000	7	6	9
9	1	4	55,516	8	5	0

Abbildung 8.50: Entwicklung der Fehlerquadratsumme beim Ward-Verfahren

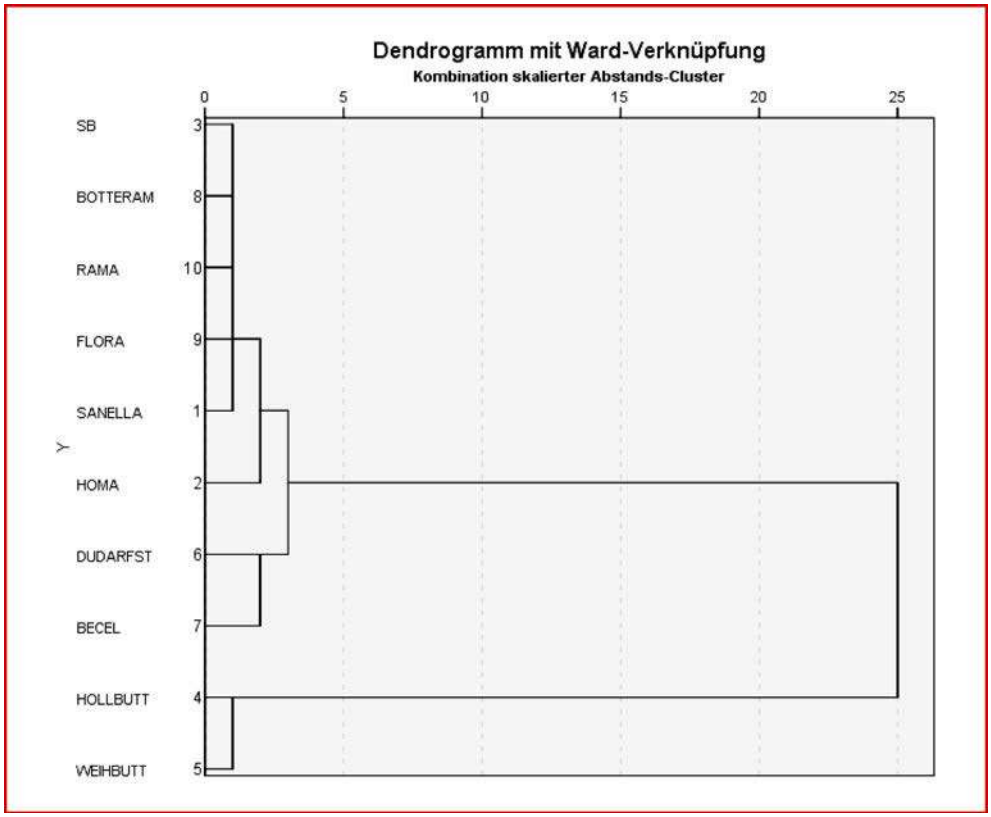


Abbildung 8.51: Dendrogramm für das Ward-Verfahren

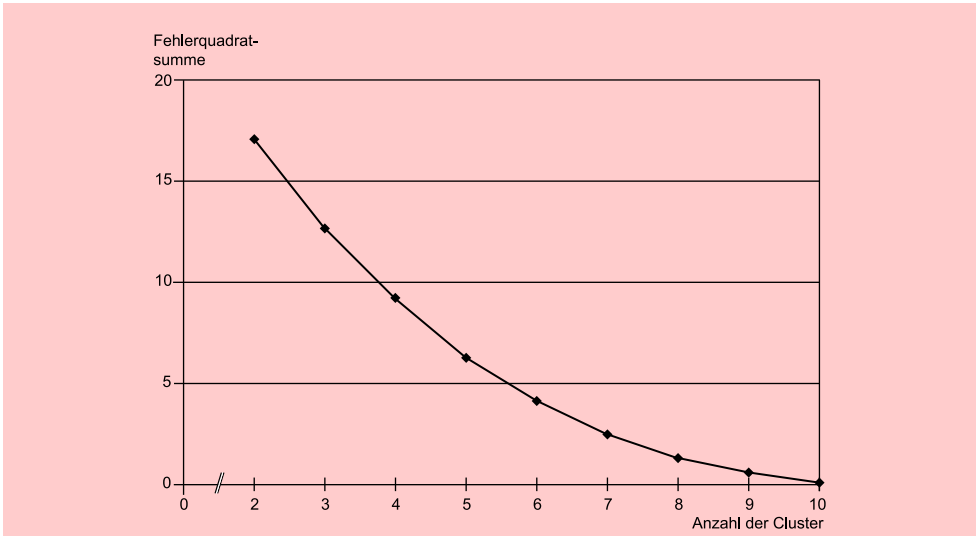


Abbildung 8.52: Entwicklung des Heterogenitätsmaßes im Fallbeispiel

quadratsumme (Vergrößerung des Heterogenitätszuwachse um $55,516 - 17,0 = 38,516$) zu beobachten ist, erscheint eine Zwei-Cluster-Lösung in diesem Fall zweckmäßig. Um die Entscheidung für eine Zwei-Cluster-Lösung abzusichern, sollten weiterhin aber auch die in Abschnitt 8.2.3 diskutierten Kriterien zur *Bestimmung der optimalen Clusterzahl* herangezogen werden. An dieser Stelle seien deshalb zusätzlich noch das Elbow-Kriterium und der Test von Mojena zur Entscheidungsstützung betrachtet:

Optimale Clusterzahl

Elbow-Kriterium

Zur Anwendung des *Elbow-Kriteriums* ist die Fehlerquadratsumme gegen die entsprechende Clusterzahl in einem Koordinatensystem abzutragen, wobei der Übergang von der Zwei- zur Ein-Cluster-Lösung nicht berücksichtigt wird, da hier immer der größte Sprung in der Heterogenitätsentwicklung liegt. Da das in Abbildung 8.52 dargestellte Diagramm für unser Fallbeispiel keinen eindeutigen „*Elbow*“ erkennen lässt, kann dies als Indikator für die Zwei-Cluster-Lösung gewertet werden.

Test von Mojena

Zur Durchführung des *Tests von Mojena* sind die Fusionslevels aus Abbildung 8.50 (Spalte „Koeffizienten“) in eine Excel-Tabelle zu übernehmen und die standardisierten Fusionskoeffizienten (α) je Fusionsstufe entsprechend Formel (8.11) zu berechnen (vgl. Abschnitt 8.2.3.3). Dabei ergibt sich ein mittlerer Fusionskoeffizient von $\bar{\alpha} = 12,118$ und eine Standardabweichung der Koeffizienten von $s_\alpha = 17,170$. Die Ergebnisse sind in Abbildung 8.53 aufgeführt. Wird *2 als kritischer Schwellenwert* herangezogen, so wird dieser nur von der Zwei-Cluster-Lösung ($\widetilde{\alpha}_i = 2,528$) überschritten, womit auch der Test von Mojena die Zwei-Cluster-Lösung bestätigt.

Fusionsschritt	1	...	6	7	8	9
Clusterzahl	10	...	5	4	3	2*
Fusions-Koeffizienten α_i	0,549	...	9,234	12,702	17,000	55,516
Standardisierte Fusionskoeffizienten $\widetilde{\alpha}_i$	-0,674	...	-0,168	0,034	0,284	2,528

Abbildung 8.53: Ergebnistabelle zum Test von Mojena im Fallbeispiel

Clusterzuordnungen

Abschließend lässt sich mit Hilfe von Abbildung 8.54 erkennen, welches Objekt sich in welchem Cluster befindet. Für unser Beispiel wurden die Clusterzuordnungen für die 2-, 3-, 4- und 5-Cluster-Lösung angegeben. Es wird deutlich, dass bei der 2-Cluster-Lösung die Objekte „Holländische Butter“ und „Weihnachtsbutter“ (*Butter-Cluster*) in der zweiten Gruppe zusammengefasst sind, während alle übrigen Objekte zu Cluster 1 (*Margarine-Cluster*) gehören.

Zum Vergleich der agglomerativen Verfahren wurde das hier betrachtete Fallbeispiel auch mit den Verfahren „Complete-Linkage“, „Average-Linkage“, „Centroid“ und „Median“ analysiert.[24] Als zentraler Unterschied zum Ward-Verfahren ist dabei vor allem zu nennen, dass diese Verfahren in der Spalte „Koeffizienten“ der „Zuordnungsübersicht“ (vgl. Abbildung 8.50) *nicht* den Zuwachs der Fehlerquadratsumme, sondern die Distanzen bzw. Ähnlichkeiten der jeweils zusammengefassten Objekte oder Gruppen enthalten. Allerdings führten im vorliegenden Fall alle Verfahren zu identischen Lösungen im 2-Cluster-Fall. Es ergab sich immer ein „Butter-Cluster“ und ein „Margarine-Cluster“.

Abschließend sei die *2-Cluster-Lösung* unseres Fallbeispiels noch einer näheren Betrachtung unterzogen: Im ersten Cluster sind die Produkte „Holländische Butter“ und „Weihnachtsbutter“ zusammengefasst, und wir bezeichnen dieses Cluster deshalb als

[24] In der deutschsprachigen Version von SPSS sind die Verfahren Single Linkage als „Nächstgelegener Nachbar“, Complete Linkage als „Entferntester Nachbar“ und Average Linkage als „Linkage zwischen den Gruppen“ bezeichnet. SPSS bietet bei dem Verfahren „Linkage zwischen den Gruppen“ jedoch nur die ungewichtete Variante.

Cluster-Zugehörigkeit

Fall	5 Cluster	4 Cluster	3 Cluster	2 Cluster
1:SANELLA	1	1	1	1
2:HOMA	2	2	1	1
3:SB	1	1	1	1
4:HOLLBUTT	3	3	2	2
5:WEIHBUTT	3	3	2	2
6:DUDARFST	4	4	3	1
7:BECEL	5	4	3	1
8:BOTTERAM	1	1	1	1
9:FLORA	1	1	1	1
10:RAMA	1	1	1	1

Abbildung 8.54: Clusterzuordnungen im Margarine-Beispiel

„*Butter-Cluster*“. Das zweite Cluster enthält alle Margarinesorten, und es wird deshalb als „*Margarine-Cluster*“ bezeichnet. Zur Beurteilung der beiden Gruppen lassen sich die Mittelwerte und Varianzen der 10 Eigenschaftsurteile über die zehn betrachteten Marken (ohne die Marke „Delicado Sahnebutter“) sowie die entsprechenden Mittelwerte und Varianzen der Variablen in dem jeweiligen Cluster heranziehen. Diese Kennzahlen wurden für die Erhebungsgesamtheit (10 Marken) mit Hilfe der SPSS-Prozedur „Deskriptive Statistiken“ (*Menüfolge: Analysieren → Deskriptive Statistiken → Deskriptive Statistik*) berechnet und sind in Abbildung 8.55 abgedruckt.

Ein erstes Kriterium zur Beurteilung der *Homogenität einer gefundenen Gruppe* stellt der F-Wert dar, der sich für jede Variable in einer Gruppe wie folgt berechnet: F-Wert

$$F = \frac{V(J,G)}{V(J)} \tag{8.12}$$

mit

$V(J,G)$ = Varianz der Variable J in Gruppe G
$V(J)$ = Varianz der Variable J in der Erhebungsgesamtheit

Je kleiner ein F-Wert ist, desto geringer ist die Streuung dieser Variable in einer Gruppe im Vergleich zur Erhebungsgesamtheit. Der F-Wert sollte 1 nicht übersteigen, da in diesem Fall die entsprechende Variable in der Gruppe eine größere Streuung aufweist als in der Erhebungsgesamtheit.

Für die Variable „Streichfähigkeit“ im „Butter-Cluster“ ergibt sich beispielsweise eine Varianz von 0,00157, womit sich der entsprechende F-Wert wie folgt berechnet:

$$F = \frac{0,00157}{0,6134} = 0,00256$$

Die F-Werte sind nun für *alle* Variablen in beiden Clustern zu berechnen. Ein Cluster ist dann als *vollkommen homogen* anzusehen, wenn alle F-Werte kleiner als 1 sind.

Deskriptive Statistik

	N	Mittelwert	Varianz
Streichfähigkeit	10	4,76330	,613
Preis	10	4,01590	,239
Haltbarkeit	10	4,29140	,185
Anteil ungesättigter Fettsäuren	10	3,86760	,032
Brat- und Backeignung	10	3,84160	,482
Geschmack	10	4,43390	,373
Kaloriengehalt	10	4,11530	,655
Anteil tierischer Fette	10	2,73810	2,870
Vitamingehalt	10	4,11060	,229
Natürlichkeit	10	4,07780	,490
Gültige Werte (Listenweise)	10		

Abbildung 8.55: Mittelwerte und Varianzen der Eigenschaftsurteile über die zehn betrachteten Marken

t-Wert

Ein weiteres Kriterium, das allerdings primär Anhaltspunkte zur Interpretation der Cluster liefern soll, stellt der t-Wert dar. Er berechnet sich für jede Variable in einer Gruppe wie folgt:

$$t = \frac{\bar{X}(J,G) - \bar{X}(J)}{S(J)}$$

mit

$\bar{X}(J,G)$ = Mittelwert der Variable J über die Objekte in Gruppe G
$\bar{X}(J)$ = Gesamtmittelwert der Variable J in der Erhebungsgesamtheit
$S(J)$ = Standardabweichung der Variable J in der Erhebungsgesamtheit

Die t-Werte stellen normierte Werte dar, wobei

- negative t-Werte anzeigen, dass eine Variable in der betrachten Gruppe im Vergleich zur Erhebungsgesamtheit unterrepräsentiert ist;
- positive t-Werte anzeigen, dass eine Variable in der betrachten Gruppe im Vergleich zur Erhebungsgesamtheit überrepräsentiert ist.

Somit dienen diese Werte nicht zur Beurteilung der Güte einer Cluster-Lösung, sondern können zur *Charakterisierung der jeweiligen Cluster* herangezogen werden. Für die Variable „Streichfähigkeit“ im „Butter-Cluster“ ergibt sich ein Mittelwert von 3,472. Der t-Wert errechnet sich somit wie folgt:

$$t = \frac{3,472 - 4,7633}{\sqrt{0,6134}} = -1,6487$$

In Abbildung 8.56 sind die F- und t-Werte für beide Cluster zusammengefasst. Da SPSS zu den Cluster-Lösungen keine Statistiken ausdruckt, empfiehlt es sich, mittels der SPSS-Prozedur AGGREGATE (*Menüpunkt: Daten - Aggregieren*) zunächst die Gruppenmittelwerte und -streuungen zu bestimmen. Daran anschließend lassen sich die F- und t-Werte am einfachsten berechnen, indem die Gruppenmittelwerte und -streuungen in ein Excel-Diagramm übertragen werden. In Excel lassen sich Ergebniszeilen bzw. -spalten festlegen, die durch bestimmte Rechenfunktionen determiniert werden. Als solche lassen sich hier die Formeln zur Berechnung der F- und t-Werte bezeichnen.

Berechnung der F- und t-Werte

	F-Werte		t-Werte	
	Margarine-Cluster	Butter-Cluster	Margarine-Cluster	Butter-Cluster
Streichfähigkeit	0,31450	0,00256	0,41219	-1,64875
Preis	0,45789	5,07480	-0,13416	0,53664
Haltbarkeit	0,86499	0,02542	0,27017	-1,08067
Ungesättigte Fettsäuren	1,15860	0,87227	-0,02094	0,08374
Back- und Brateignung	1,07148	0,48145	-0,15955	0,63819
Geschmack	0,53122	0,02589	-0,36248	1,44990
Kaloriengehalt	0,52749	0,15041	-0,35907	1,43627
Anteil tierischer Fette	0,10985	0,02582	-0,45291	1,81165
Vitamingehalt	0,97901	0,60863	-0,19611	0,78442
Natürlichkeit	0,14823	0,00960	-0,44589	1,78357

Abbildung 8.56: F- und t-Werte für die 2-Cluster-Lösung

Es wird deutlich, dass bei den F-Werten nur die Variablen „Ungesättigte Fettsäuren“ sowie „Back- und Brateignung“ im Margarine-Cluster und die Variable „Preis“ im Butter-Cluster Werte größer 1 aufweisen. Das bedeutet, dass diese Variablen in den Gruppen eine größere Heterogenität aufweisen als in der Erhebungsgesamtheit. Ansonsten sind beide Cluster durch eine relativ homogene Variablenstruktur gekennzeichnet.

Interpretation

Bezüglich der t-Werte zeigt sich für das „Margarine-Cluster“, dass die Variablen „Streichfähigkeit“ und „Haltbarkeit“ positive Werte aufweisen, d. h. überrepräsentiert sind. Im „Butter-Cluster“ hingegen sind genau diese Variablen unterrepräsentiert, denn sie weisen dort negative t-Werte auf.

Alle übrigen Variablen zeigen die umgekehrte Tendenz. Sie sind im „Margarine-Cluster“ unterrepräsentiert (negative t-Werte) und im „Butter-Cluster“ überrepräsentiert (positive t-Werte). Somit sind die Marken im „Margarine-Cluster“ vor allem durch eine hohe „Streichfähigkeit“ sowie „Haltbarkeit“ gekennzeichnet. Andererseits werden z. B. „Geschmack“, „Kaloriengehalt“ und „Natürlichkeit“ der Margarinemarken eher als gering angesehen.

Das „Butter-Cluster“ hingegen ist durch z. B. hohe Werte bei „Geschmack“, „Kaloriengehalt“ und „Natürlichkeit“ gekennzeichnet, während „Streichfähigkeit“ und „Haltbarkeit“ bei den Buttermarken nur gering ausgeprägt sind.

Eine weitere Möglichkeit zur Feststellung der Trennschärfe zwischen den gefundenen Clustern bietet auch die Anwendung einer Diskriminanzanalyse im Anschluss an die Clusteranalyse. In diesem Fall werden die gefundenen Cluster als Gruppen vorgegeben und die Eigenschaftsurteile als unabhängige Variable betrachtet. Mit Hilfe

einer schrittweisen Diskriminanzanalyse lassen sich dann diejenigen Eigenschaftsurteile ermitteln, die besonders zur Trennung der gefundenen Cluster beitragen.[25]

8.3.3 SPSS-Kommandos

In Abbildung 8.57 ist abschließend die Syntaxdatei mit den SPSS-Kommandos für das Fallbeispiel wiedergegeben.[26]

```
* MVA: Fallbeispiel Clusteranalyse.
* DATENDEFINITION.
DATA LIST FIXED / streichf preis haltbark ungefett backeign geschmac kalorien
                  tierfett vitamin natur 1-50 (3) marke 51-60 (A).

BEGIN DATA
4500 4000 4375 3875 3250 3750 4000 2000 4625 4125 Sanella
5167 4250 3833 3833 2167 3650 3273 1857 3750 3417 Homa
5059 3824 4765 3438 4235 4471 3765 1923 3529 3529 SB
.....
4500 4000 4200 3900 3700 3900 3600 1500 3500 3700 Rama
END DATA.

* PROZEDUR.
* Clusteranalyse für den Margarinemarkt (Ward).
CLUSTER streichf TO natur
  /METHOD = WARD
  /MEASURE = SEUCLID
  /ID = marke
  /PRINT SCHEDULE CLUSTER (2,5)
  /PRINT DISTANCE
  /PLOT = DENDROGRAM.
```

Abbildung 8.57: SPSS-Job zur Clusteranalyse

8.4 Anwendungsempfehlungen

8.4.1 Vorüberlegungen bei der Clusteranalyse

Bevor eine Clusteranalyse durchgeführt wird, sollte der Anwender einige Überlegungen zur Auswahl und Aufbereitung der Ausgangsdaten anstellen. Im Einzelnen sollten insbesondere folgende Punkte Beachtung finden:[27]

1. Anzahl der Objekte
2. Problem der Ausreißer
3. Anzahl zu betrachtender Merkmale (Variable)
4. Gewichtung der Merkmale
5. Vergleichbarkeit der Merkmale

Wurde eine *Clusteranalyse auf Basis einer Stichprobe* durchgeführt und sollen aufgrund der gefundenen Gruppierung Rückschlüsse auf die Grundgesamtheit gezogen werden, so muss sichergestellt werden, dass auch genügend Elemente in den einzelnen Gruppen enthalten sind, um die entsprechenden Teilgesamtheiten in der Grundgesamtheit zu repräsentieren. Da i. d. R. im Vorfeld aber nicht bekannt ist, welche Gruppen in einer Erhebungsgesamtheit vertreten sind, – denn das Auffinden solcher

[25] Vgl. zur Diskriminanzanalyse Kap. 4 in diesem Buch.
[26] Vergleiche zur SPSS-Syntax auch die Ausführungen im Kapitel „Zur Verwendung dieses Buches“.
[27] Vgl. zu speziellen Anwendungsfragen auch Bacher/Pöge/Wenzig (2010), S. 457 ff.

Gruppen ist ja gerade das Ziel der Clusteranalyse – sollte insbesondere sog. Ausreißer aus einer gegebenen Objektmenge herausgenommen werden. *Ausreißer* sind Objekte, die im Vergleich zu den übrigen Objekten eine vollkommen anders gelagerte Kombination der Merkmalsausprägungen aufweisen und sich dadurch von allen anderen Objekten stark unterscheiden. Sie führen dazu, dass der Fusionierungsprozess der übrigen Objekte stark beeinflusst wird und damit das Erkennen der Zusammenhänge zwischen den übrigen Objekten erschwert wird und Verzerrungen auftreten. Eine Möglichkeit zum Auffinden solcher Ausreißer bietet z. B. das Single-Linkage-Verfahren (vgl. Abschnitt 8.2.2.3). Mit seiner Hilfe können Ausreißer erkannt und dann aus der Untersuchung ausgeschlossen werden.

Ausreißeranalyse

Ebenso wie für die Anzahl der zu betrachtenden Objekte gibt es auch für die Zahl der in einer Clusteranalyse heranzuziehenden Variablen keine eindeutigen Vorschriften. Der Anwender sollte darauf achten, dass nur solche Merkmale im Gruppierungsprozess Berücksichtigung finden, die aus theoretischen Überlegungen als *relevant* für den zu untersuchenden Sachverhalt anzusehen sind. Merkmale, die für den Untersuchungszusammenhang bedeutungslos sind, müssen aus dem Gruppierungsprozess herausgenommen werden.

Weiterhin lässt sich im Voraus i. d. R. nicht bestimmen, ob die betrachteten Merkmale mit unterschiedlichem Gewicht zur Gruppenbildung beitragen sollen, sodass bei praktischen Anwendungen weitgehend eine *Gleichgewichtung der Merkmale* unterstellt wird. Hierbei ist darauf zu achten, dass insbesondere durch hoch korrelierende Merkmale bei der Fusionierung der Objekte bestimmte Aspekte überbetont werden, was wiederum zu einer Verzerrung der Ergebnisse führen kann. Soll eine Gleichgewichtung der Merkmale sichergestellt werden und liegen *korrelierte Ausgangsdaten* vor, so bieten sich folgende Lösungsmöglichkeiten an:

Korrelierte Ausgangsdaten

- Ausschluss korrelierter Variable:
  Weisen zwei Merkmale hohe Korrelationen (>0,9) auf, so gilt es zu überlegen, ob eines der Merkmale nicht aus den Ausgangsdaten auszuschließen ist. Die Informationen, die eine hoch korrelierte Variable liefert, werden größtenteils durch die andere Variable mit erfasst und können von daher als redundant angesehen werden. Der Ausschluss korrelierter Merkmale aus der Ausgangsdatenmatrix ist u. E. die sinnvollste Möglichkeit, eine Gleichgewichtung der Daten sicherzustellen.
- Vorschalten einer explorativen Faktorenanalyse:
  Das Ziel der explorativen Faktorenanalyse (vgl. Kapitel 7 in diesem Buch) liegt vor allem in der Reduktion hoch korrelierter Variablen auf unabhängige Faktoren. Werden die Ausgangsvariablen mit Hilfe einer Faktorenanalyse (Hauptkomponentenanalyse) auf unabhängige Faktoren verdichtet, so kann auf Basis der Faktorwerte, zwischen denen keine Korrelationen mehr auftreten, eine Clusteranalyse durchgeführt werden. Dabei ist aber darauf zu achten, dass die Faktoren und damit auch die Faktorwerte Interpretationsschwierigkeiten aufweisen können, sofern nur auf die zentralen und nicht auf alle Faktoren zurückgegriffen wird auch nur einen Teil der Ausgangsinformation widerspiegeln.
- Verwendung der *Mahalanobis-Distanz*:
  Wird zur Ermittlung der Unterschiede zwischen den Objekten die Mahalanobis-Distanz verwendet, so lassen sich dadurch bereits im Rahmen der Distanzberechnung zwischen den Objekten etwaige Korrelationen zwischen den Variablen ausschließen. Die Mahalanobis-Distanz stellt allerdings bestimmte Vorausset-

Mahalanobis-Distanz

zungen an das Datenmaterial (z. B. einheitliche Mittelwerte der Variablen in allen Gruppen), die gerade bei Clusteranalyseproblemen häufig nicht erfüllt sind.[28]

Konstante Merkmale

Schließlich sollte der Anwender darauf achten, dass in den Ausgangsdaten *keine konstanten Merkmale*, d. h. Merkmale, die bei allen Objekten dieselbe Ausprägung besitzen, auftreten, da sie zu einer Nivellierung der Unterschiede zwischen den Objekten beitragen und somit Verzerrungen bei der Fusionierung hervorrufen können. Konstante Merkmale sind nicht trennungswirksam und können von daher aus der Analyse herausgenommen werden (das gilt besonders für Merkmale, die fast überall Null-Werte aufweisen).

Ebenfalls zu einer (impliziten) Gewichtung kann es dann kommen, wenn die Ausgangsdaten auf *unterschiedlichen Skalen* erhoben wurden. So kommt es allein dadurch zu einer Vergrößerung der Differenzen zwischen den Merkmalsausprägungen, wenn ein Merkmal auf einer sehr fein dimensionierten (d. h. breiten) Skala erhoben wurde. Um eine Vergleichbarkeit zwischen den Variablen herzustellen, empfiehlt es sich, zu Beginn der Analyse z. B. eine Standardisierung der Daten vorzunehmen.[29] Durch die Transformation

$$z_{kj} = \frac{x_{kj} - \bar{x}_j}{S_j}$$

mit

x_{kj} = Ausprägung von Merkmal j bei Objekt k
$\bar{x}_j$ = Mittelwert von Merkmal j
S_j = Standardabweichung von Merkmal j

wird erreicht, dass alle Variable einen Mittelwert von Null und eine Varianz von Eins besitzen (sog. standardisierte oder normierte Variable).

8.4.2 Empfehlungen zur Durchführung einer Clusteranalyse

Erst nach den im vorangegangenen Abschnitt vorgetragenen Überlegungen beginnt die eigentliche Aufgabe der Clusteranalyse. Der Anwender muss nun entscheiden, welches Proximitätsmaß und welcher Fusionierungsalgorithmus verwendet werden soll. Diese Entscheidungen können letztlich nur vor dem Hintergrund einer konkreten Anwendungssituation getroffen werden, wobei die in Abschnitt 8.2.2.3 diskutierten Fusionierungseigenschaften der alternativen agglomerativen Clusterverfahren als Entscheidungshilfe dienen können. Dabei ist besonders das Ward-Verfahren hervorzuheben, da eine Simulationsstudie von Bergs gezeigt hat, dass nur das Ward-Verfahren „gleichzeitig sehr gute Partitionen findet und meistens die richtige Clusterzahl signalisiert“.[30] Zur „Absicherung“ der Clusteranalyse können die Ergebnisse z. B. des Ward-Verfahrens anschließend auch durch die Anwendung anderer Algorithmen überprüft werden. Dabei sollten aber die unterschiedlichen Fusionierungseigenschaften der einzelnen Algorithmen beachtet werden (vgl. Abbildung 8.36).

[28] Vgl. Hair et al. (2014), S. 522 ff.; Kline (2011), S. 54 f.; Steinhausen/Langer (1977), S. 89 ff.
[29] Weitere Möglichkeiten zur Sicherstellung der Vergleichbarkeit von Merkmalen zeigt Bergs (1981), S. 59 f.
[30] Bergs (1981), S. 97.

Die agglomerativen Verfahren führen allerdings insbesondere bei einer großen Fallzahl zu Berechnungsproblemen, da sie für jeden Fusionierungsschritt die Berechnung der Distanzmatrix zwischen allen Fällen erfordern. Bei einer *großen Anzahl von Fällen* empfiehlt sich deshalb die Verwendung eines partitionierenden Clusteralgorithmus, wie er in SPSS in Form der *K-Means-Clusteranalyse* (*Menüfolge: Analysieren → Klassifizieren → K-Means-Cluster*) implementiert ist. Das Verfahren erfordert vorab die Festlegung der gewünschten Clusterzahl und ordnet dann die Fälle entsprechend der zur Clusterung herangezogenen metrisch skalierten Variablen den Gruppen zu. Die Clusterzentrenanalyse minimiert die Streuungsquadratsumme innerhalb der Cluster mit Hilfe der einfachen euklidischen Distanz, wodurch eine optimale Zuordnung der Objekte zu den Clustern erfolgt. Im Ergebnis liefert das Verfahren eine Zuordnung der Fälle zu der vorgegebenen Clusterzahl und es kann eine F-Statistik zur Varianzanalyse (ANOVA-Tabelle) angefordert werden. Aus der relativen Größe dieser Statistik lassen sich dann auch – ähnlich einer Diskriminanzanalyse – Informationen über den Beitrag jeder Variablen zu der Trennung der Gruppen gewinnen.

K-Means-Clusteranalyse

Zur abschließenden Verdeutlichung der durchzuführenden Tätigkeiten im Rahmen einer Clusteranalyse sei auf Abbildung 8.58 verwiesen. Sie enthält auf der linken Seite die acht wesentlichen Arbeitsschritte eines Gruppierungsprozesses. Die einzelnen

Entscheidungsprobleme der Clusteranalyse

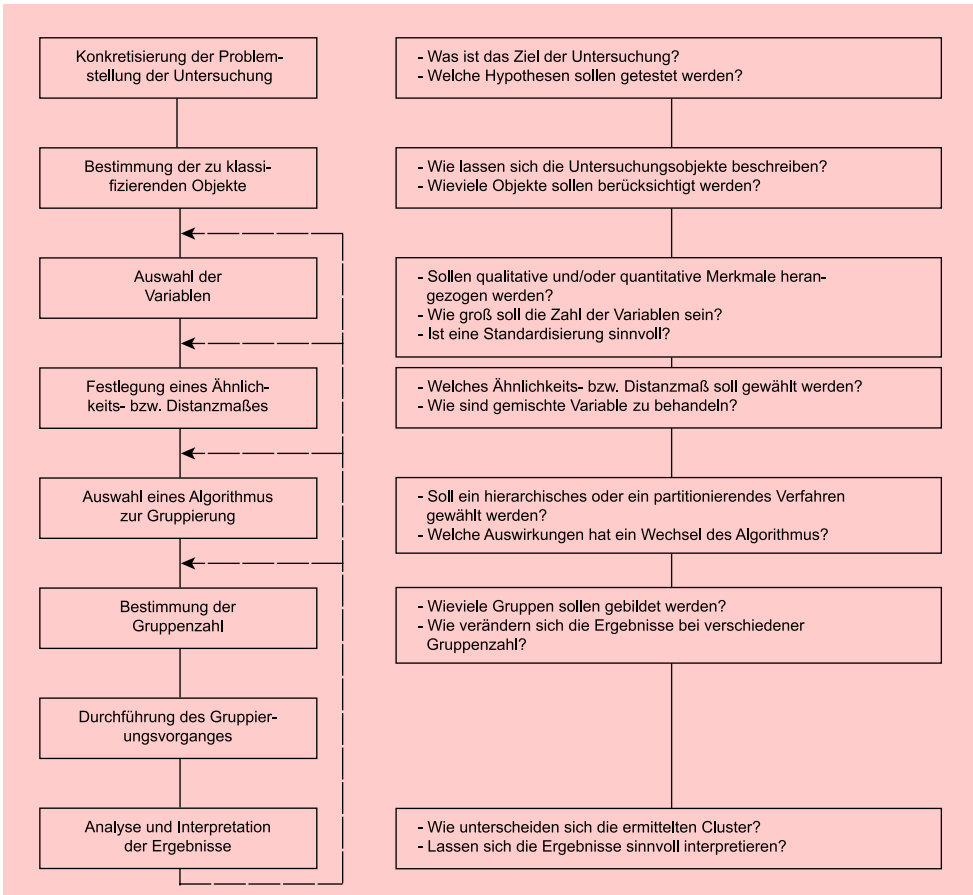


Abbildung 8.58: Ablaufschritte und Entscheidungsprobleme der Clusteranalyse

Schritte bedürfen nunmehr keiner weiteren Erläuterung, es soll allerdings vermerkt werden, dass die Analyse und Interpretation der Ergebnisse zu einem wiederholten Durchlauf einzelner Stufen führen kann. Dies wird immer dann der Fall sein, wenn die Ergebnisse keine sinnvolle Interpretation gestatten. Eine weitere Begründung für die Wiederholung erkennt man bei Betrachtung der rechten Seite der Abbildung. Dort sind für jeden Ablaufschritt beispielhaft Problemstellungen in Form von Fragen genannt, auf die bei Durchführung einer Studie Antwort gefunden werden muss. Die Überprüfung der Auswirkungen einer anderen Antwortalternative auf die Gruppierungsergebnisse kann somit ebenfalls zu einem wiederholten Durchlauf einzelner Stufen führen.

Abschließend sei noch darauf hingewiesen, dass die genannten Fragen nur die zentralen Entscheidungsprobleme einer Clusteranalyse betreffen und auf viele dieser Fragen mehr als zwei Antwortalternativen existieren. Vor diesem Hintergrund wird deutlich, dass der Anwender bei der Clusteranalyse über einen breiten Manövrier- und Einflussraum verfügt. Diese Tatsache hat zwar den *Vorteil*, dass sich hierdurch ein breites Anwendungsgebiet der Clusterverfahren ergibt. Auf der anderen Seite steht der Anwender in der *Gefahr*, die Daten der Untersuchung so zu manipulieren, dass sich die gewünschten Ergebnisse einstellen. Um Dritten einen Einblick in das Vorgehen im Rahmen der Analyse zu geben, sollte der jeweilige Anwender deshalb bei Darstellung seiner Ergebnisse wenigstens die nachstehenden Fragen begründet und eindeutig beantworten.

1. Welches Ähnlichkeitsmaß und welcher Algorithmus wurden gewählt?
2. Was waren die Gründe für die Wahl?
3. Wie stabil sind die Ergebnisse bei
   - Veränderung des Ähnlichkeitsmaßes
   - Wechsel des Algorithmus
   - Veränderung der Gruppenzahl?

Die Behandlung von Missing Values

Missing values

Als fehlende Werte (MISSING VALUES) werden Variablenwerte bezeichnet, die von den Befragten entweder außerhalb des zulässigen Beantwortungsintervalls vergeben oder überhaupt nicht eingetragen wurden. Im Datensatz können fehlende Werte der Merkmalsvariablen als Leerzeichen kodiert werden. Sie werden dann vom Programm automatisch durch einen sog. System-missing-value ersetzt. Alternativ kann man die fehlenden Werte im Datensatz auch durch eine 0 (oder durch einen anderen Wert, der unter den beobachteten Werten nicht vorkommt), ersetzen. Mit Hilfe der Anweisung

MISSING VALUES streichf to natur (00000)

kann man dem Programm sodann mitteilen, dass der Wert 00000 für einen fehlenden Wert steht. Derartige vom Benutzer bestimmte fehlende Werte werden von SPSS als User-missing-values bezeichnet. Für eine Variable lassen sich mehrere Missing Values angeben, z. B. 0 für „Ich weiß nicht“ und 9 für „Antwort verweigert“. Im Rahmen der hier aufgezeigten Clusteranalyse treten allerdings keine fehlenden Werte auf.

Die Clusteranalyse selbst verfügt über keine Optionen zur Behandlung von fehlenden Werten. Allerdings stellt SPSS vor allem im Modul „SPSS Missing Value Analysis“

spezielle Verfahren zur Analyse fehlender Werte zur Verfügung. Da die Clusteranalyse „vollständige Datensätze“ voraussetzt, empfiehlt es sich, den Datensatz zunächst um fehlende Werte zu bereinigen. Dabei sei hier vor allem auf die folgenden Möglichkeiten hingewiesen:

- Variable mit großer Anzahl fehlender Werte aus der Analyse ausschließen.
- Fälle mit fehlenden Werten für Variablen vollständig aus der weiteren Analyse ausschließen (sog. Listenweiser Fallausschluss). Problem: Reduktion der Fallzahl.
- Fehlende Werte z. B. durch den Mittelwert der Ausprägungen einer Variablen bei den gültigen Fällen ersetzen. Problem: Ergebnisverzerrung und Verringerung der Varianz innerhalb der Datenstruktur bei zu häufigem Auftreten von fehlenden Werten bei einer Variablen.

Literaturhinweise

A. Basisliteratur zur Clusteranalyse

Bacher, J./Pöge, A./Wenzig, K. (2010), Clusteranalyse: Anwendungsorientierte Einführung in Klassifikationsverfahren, 3. Auflage, München.

Everitt, B./Landau, S./Leese, M./Stahl, D. (2011), Cluster Analysis, 5. Auflage, New York.

Hair, J./Black, W./Babin, B./Anderson, R. (2014), Multivariate Data Analysis, 7. Auflage, Englewood Cliffs (N.J.), Kapitel 9.

Jensen, O. (2008), Clusteranalyse, in: Herrmann A./Homburg C./Klarmann, M. (Hrsg.): Handbuch Marktforschung - Methoden, Anwendungen und Praxisbeispiele, 3. Auflage, Wiesbaden, S. 335–372.

Schendera, C. (2010), Clusteranalyse in SPSS, München.

B. Zitierte Literatur

Bacher, J./Pöge, A./Wenzig, K. (2010), Clusteranalyse: Anwendungsorientierte Einführung in Klassifikationsverfahren, 3. Auflage, München.

Bergs, S. (1981), Optimalität bei Cluster-Analysen, Diss. Münster.

Calinski, T./Harabasz, J. (1974), A dendrite method for cluster analysis, in: *Communications in Statistics – Theory and Methods A*, Vol. 3, Nr. 1, S. 1–27.

Fett, K. (2008), Clusteranalyse in CRM, Sales und Marketing: Grundlagen und praktische Anwendung, Norderstedt.

Gower, J. C. (1971), A general coefficient of similarity and some of its properties, in: *Biometrics*, Vol. 27, Nr. 4, S. 857–872.

Hair, J./Black, W./Babin, B./Anderson, R. (2014), Multivariate Data Analysis, 7. Auflage, Englewood Cliffs (N.J.), Kapitel 3.

Handl, A. (2010), Multivariate Analysemethoden, 2. Auflage, Berlin.

Herink, M./Petersen, V. (2004), Kurzbeitrag – Clusteranalyse als Instrument zur Gruppierung von spezialisierten Marktfruchtunternehmen, in: *Agrarwirtschaft*, Vol. 53, Nr. 2, S. 289–294.

Kaufman, L./Rousseeuw, P. (2008), Finding Groups in Data: An Introduction to Cluster Analysis, New York.

Kaufmann, H./Pape, H. (1996), Clusteranalyse, in: Fahrmeir, L./Hamerle, A./Tutz, G. (Hrsg.): Multivariate statistische Verfahren, 2. Auflage, Berlin u. a.

Kline, R. (2011), Principles and Practice of Structural Equation Modeling, 3. Auflage, New York.

Kohn, W. (2005), Statistik: Datenanalysis und Wahrscheinlichkeitsrechnung, 3. Auflage, Berlin/Heidelberg.

Meyer, J. (2004), Mundpropaganda im Internet: Bezugsrahmen und empirische Fundierung des Einsatzes von Virtual Communities im Marketing, Hamburg.

Milligan, G./Cooper, M. (1985), An Examination of Procedures for Determining the Number of Clusters in a Data Set, in: *Psychometrika*, Vol. 50, Nr. 2, S. 159–179.

Milligan, G. W. (1980), An Examination of the Effect of Six Types of Error Pertubation on Fifteen Clustering Algorithms, in: *Psychometrika*, Vol. 45, Nr. 3, S. 325–342.

Mojena, R. (1977), Hierarchical clustering methods and stopping rules: A evaluation, in: *The Computer Journal*, Vol. 20, Nr. 4, S. 359–363.

Punj, G./Stewart, D. (1983), Cluster Analysis in Marketing Research: Review and Suggestions for Application, in: *Journal of Marketing Research*, Vol. 20, Nr. 2, S. 134–148.

Steinhausen, D./Langer, K. (1977), Clusteranalyse, Berlin u. a.

Swoboda, B./Hälsig, F./Morschett, D. (2007), Einfluss von Einkaufsmotiven auf den Aufbau einer Händlermarke – Eine Mehrgruppenkausalbetrachtung, in: Ahlert, D./Olbrich, R./Schröder, H. (Hrsg.): Shopper Research – Kundenverhalten im Handel, Jahrbuch Vertriebs- und Handelsmanagement 2007, Frankfurt am Main, S. 19–38.

Weiber, R./Fälsch, H. (2007), Ubiquitous Computing – Eine neue Dimension in der Gestaltung von Interaktionsbeziehungen im Direktmarketing, in: *Zeitschrift für Betriebswirtschaft – Special Issue*, Vol. 3, S. 83–116.

Wiesener, O. (2014), Mit mehrstufigem Wissenserwerb zu mehr Innovationserfolg, Wiesbaden.

9 Conjoint-Analyse

K. Backhaus, B. Erichson, W. Plinke, R. Weiber, *Multivariate Analysemethoden*,
https://doi.org/10.1007/978-3-662-56655-8_10

9.1 Problemstellung

Allgemein bezeichnet der Begriff „Conjoint-Analyse“, für den in der Literatur insbesondere auch die Bezeichnungen „Conjoint Measurement“, „konjunkte Analyse“ oder „Verbundmessung“ gebräuchlich sind, eine Vorgehensweise zur Abbildung der Beurteilung einer Menge von Objekten durch eine einzelne Person. Conjoint-Analysen sind damit zunächst einmal *Individualanalysen*, durch die sich das Beurteilungsverhalten einer konkreten Person nachvollziehen lässt. Die dabei betrachteten Objekte sind zwar durch bestimmte Merkmale oder Eigenschaften beschrieben (sog. experimentelles Design), es wird aber unterstellt, dass sie von einer Person ganzheitlich betrachtet (CONsidered JOINTly) und beurteilt werden. In Abhängigkeit des mit der Beurteilung verfolgten Ziels lassen sich zwei grundsätzliche Verfahrensgruppen von Conjoint-Analysen unterscheiden:

Traditionelle Conjoint-Analysen

(1) Traditionelle Conjoint-Analysen:
Bei Conjoint-Analysen werden traditionell die betrachteten Objekte in eine *Rangordnung* gebracht, die den persönlichen Objekt-Präferenzen einer Person entspricht. Sie unterstellen damit, dass der Beurteiler über ein *vollständig determiniertes Präferenzmodell* verfügt, das ihm die Aufstellung einer solchen vollständigen Rangordnung ermöglicht. Die hier bestehenden Verfahrensvarianten versuchen bestimmte Verfahrensprobleme (z. B. Beurteilung umfangreicher Erhebungsdesigns, individuelle Anpassungen der Merkmalsstruktur) durch geeignete methodische Variationen zu beseitigen oder zumindest abzumildern. Diese Gruppe von Conjoint-Analysen ist Gegenstand dieses Kapitels.

Auswahlbasierte Conjoint-Analysen

(2) Auswahlbasierte Conjoint-Analysen:
Bei auswahlbasierten Conjoint-Analysen besteht das Ziel darin, eine konkrete *Auswahlentscheidung* aus der betrachteten Objekt-Menge vorzunehmen (sog. Choice Based Conjoint). Der Beurteiler ist hier also nicht aufgefordert eine Präferenzrangfolge über alle Objekte zu erstellen, sondern jeweils aus einem Set von Alternativen eine ihm als geeignet erscheinende Auswahl zu treffen, wobei auch die Option besteht, keine der Alternativen zu wählen (sog. *probabilistisches Präferenzmodell*). In Abhängigkeit des unterstellten Auswahlmodus ergeben sich dann wiederum unterschiedliche Verfahrensvarianten der auswahlbasierten Conjoint-Analysen.

> Die Ansätze der auswahlbasierten Conjoint-Analyse werden den *„Fortgeschrittenen Verfahren der multivariaten Analyse“* zugeordnet und sind als eigenständiges Kapitel in dem Buch *„Backhaus, K./Erichson, B./Weiber, R.: Fortgeschrittene Multivariate Analysemethoden, 3. Auflage, Berlin 2015.“* enthalten.

Beiden Verfahrensgruppen ist gemeinsam, dass aus der erhobenen ordinalen Rangordnung der Objekte bzw. der vorgenommenen Auswahlentscheidungen sog. metrische Teilnutzenwerte für die einzelnen Eigenschaftsausprägungen der Objekte geschätzt werden, mit deren Hilfe sich dann durch Addition ein Gesamtnutzenwert pro Objekt bestimmen lässt. Die Teilnutzenwerte pro Eigenschaftsausprägung sind dabei so zu bestimmen, dass sie zu Gesamtnutzenwerten pro Objekt führen, die es erlauben, die vom Befragten aufgestellte ordinale Rangordnung bzw. die vorgenommenen Auswahlentscheidungen wieder abzubilden. Charakteristisch für Conjoint-Analysen ist weiterhin ihr *dekompositioneller Ansatz*. Dieser ist dadurch gekennzeichnet,

Dekompositioneller Ansatz

dass sich die erhobenen Urteile einer Person zwar auf die Objekte in ihrer Gesamtheit beziehen, diese Gesamturteile dann aber von der Conjoint-Analyse dazu verwendet werden, die sog. Teilnutzenwerte pro Eigenschaftsausprägung zu errechnen.

Die Ausführungen in diesem Kapitel fokussieren die Grundidee der klassischen oder traditionellen Conjoint-Analyse. Auf Verfahrensvarianten der traditionellen Conjoint-Analyse wird im Abschnitt „Anwendungsempfehlungen“ am Ende dieses Kapitels hingewiesen.

Problemstellung der traditionellen Conjoint-Analyse

Die grundlegende Problemstellung der klassischen Conjoint-Analyse sei hier zunächst am Beispiel der Neuproduktplanung verdeutlicht: Im Rahmen der Neuproduktplanung stellt sich u. a. die Frage, wie ein neues Produkt (oder eine Dienstleistung) in Hinsicht auf die Bedürfnisse des Marktes optimal zu gestalten ist. Dabei muss vom Untersucher vorab festgelegt werden, welche Objekteigenschaften und welche Ausprägungen dieser Eigenschaften für das Neuprodukt relevant sind und in die Untersuchung einbezogen werden sollen. Wir gehen von folgender Problemsituation aus:

Festlegung der Eigenschaften

Ein Hersteller von Margarine plant die Neueinführung eines Produktes, das sich in zwei Eigenschaften von bestehenden Produkten abheben soll: Kaloriengehalt und Verpackung. Als Eigenschaftsausprägung betrachtet er:

- Kaloriengehalt: hoch/niedrig
- Verpackung: Becher/Papier

Durch die Festlegung von zwei Eigenschaften, mit jeweils zwei Eigenschaftsausprägungen, können vier Kombinationen von Eigenschaftsausprägungen, d. h. vier fiktive Produkte, gebildet werden:

Fiktive Produkte

Produkt I	*Produkt II*	*Produkt III*	*Produkt IV*
wenig Kalorien	wenig Kalorien	viel Kalorien	viel Kalorien
im Becher	in Papier	im Becher	in Papier

Diese vier fiktiven Produkte werden einer Auskunftsperson zur Beurteilung vorgelegt, um deren Nutzenstruktur zu ermitteln. Hierbei ist man allerdings nicht auf eine rein verbale Beschreibung der Eigenschaften und ihrer Ausprägungen beschränkt, wie es bei der Beschreibung der alternativen Produkte mittels sog. Produktkarten der Fall ist. Es lassen sich vielmehr auch reale Darstellungen oder Computeranimationen in das Erhebungsdesign integrieren. So können die verschiedenen Verpackungsformen in obigem Beispiel durchaus mittels realer Verpackungen dargestellt werden. Die Auskunftsperson wird dabei aufgefordert, über die Produkte entsprechend ihrer subjektiven Nutzenvorstellung eine Rangordnung zu bilden. Beispielsweise möge sich folgende Rangordnung ergeben haben:

Rang	*Produkt*	*Eigenschaftsausprägungen*
1	III	viel Kalorien, im Becher
2	IV	viel Kalorien, in Papier
3	I	wenig Kalorien, im Becher
4	II	wenig Kalorien, in Papier

Diese Rangreihe bildet die Grundlage zur Ableitung von Teilnutzenwerten für die einzelnen Eigenschaftsausprägungen. Die Auskunftsperson gibt also *ordinale* Gesamt-

nutzenurteile ab, aus denen durch die Conjoint-Analyse dann *metrische* Teilnutzenwerte errechnet werden. Damit wird es außerdem möglich, durch Addition der Teilnutzenwerte auch metrische Gesamtnutzenwerte zu ermitteln.

Problemstellung	Eigenschaften	Eigenschaftsausprägungen
Dehnungspotential einer Dachmarke[1]	Leistungskonzept	allg. Erfahrungen, aus Modulen, individuelle Lösungsstrategie
	vergleichbare Referenzprojekte	keine, wenige, viele
	Leistungsniveau	Abweichung bei max. 2 Kriterien, alle Kriterien erfüllt, auch künftig alle Kriterien erfüllt
	Servicegrad	bis 90 %, ca. 95 %, über 98 %
	Marke	Einzelmarke, Dachmarke, Konkurrenzmarke
Präferenzen bzgl. des Mobilen Internets[2]	Vermittlungsart	Always-On, Dial-Up
	Priorität	Priorisierter Zugang, Wartezeit bei Auslastung
	Zugriffsbeschränkung	mit, ohne
	Office Lösung	mit, ohne
	Operative Systeme	mit, ohne
	Transaktion	ohne M-Commerce, mit M-Commerce, M-Commerce und M-Payment
	Location Based Services	mit, ohne
	Messaging	SMS, Multimedia Nachrichten, Videotelefon
	Kosten	100 %, 125 %, 150 %, 175 % pro Mitarbeiter und Monat
Nachfragerpräferenzen zum ,,Intelligenten Haus “[3]	Anwendungsschwerpunkt	Komfort, Energie, Sicherheit, Kommunikation
	Bedienung	mobil, stationär, integriert, Sprache
	Installation	Fachmann, eigenständig
	Kundenservice	vor Ort, E-Mail, Call Center Hotline
Glaubwürdigkeit von Produktvorankündigungen[4]	Innovationsgrad	revolutionär, Weiterentwicklung
	Detaillierungsgrad	detailliert, undetailliert
	Zeithorizont	4 Monate, 1 Jahr
	Unternehmen	Marktführer, mittelgroßer Anbieter
Präferenzen für Online Video-Dienste[5]	Bildqualität	hoch, mittel, gering
	Wartezeit	keine, 1-2 Stunden, 12-24 Stunden, 4-6 Tage
	Nutzungsdauer	48 Stunden, 1 Monat, unbegrenzt
	Preis	3,60€, 1,80€, 0,60€, 0,00€

Abbildung 9.1: Anwendungsbeispiele zur Conjoint-Analyse

Durch die Conjoint-Analyse werden Produkte als gebündelte Menge von Eigenschaftsausprägungen aufgefasst, wobei die Objekteigenschaften die unabhängigen Variablen darstellen. Die Eigenschaftsausprägungen sind dann konkrete Werte der unabhängigen Variablen. Die abhängige Variable ist die Präferenz der Auskunftsperson für die fiktiven Produkte. In Abbildung 9.1 sind einige Anwendungsbeispiele der tra-

[1] Weiber/Billen (2004), S. 82 ff.
[2] Wirtz/Olderog/Heithecker (2003), S. 81 ff.
[3] Szuppa (2009), S. 273 ff.
[4] Schirm (1995).
[5] Nitschke/Völckner (2006), S. 755 ff.

ditionellen Conjoint-Analyse zusammengestellt. Sie vermitteln einen Einblick in die Problemstellung, die Zahl und Art der Eigenschaften sowie die betrachteten Eigenschaftsausprägungen.[6]

Die Conjoint-Analyse ist in ihrem Kern eine Analyse *individueller* Nutzenvorstellungen. Häufig interessiert darüber hinaus die Nutzenstruktur einer Mehrzahl von Personen. So möchte z. B. der Margarinehersteller nicht primär die Nutzenstruktur eines einzelnen Konsumenten ermitteln, sondern die seiner Käufer insgesamt. Zu diesem Zwecke ist eine Aggregation der individuellen Ergebnisse notwendig.

Die Planung und Durchführung einer Conjoint-Analyse erfordert die in Abbildung 9.2 dargestellten Ablaufschritte:

Ablaufschritte der Conjoint-Analyse

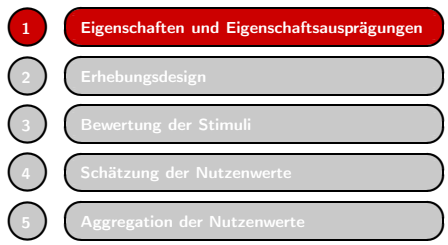
Abbildung 9.2: Ablaufschritte der Conjoint-Analyse

Zunächst müssen vom Untersucher die Eigenschaften und Eigenschaftsausprägungen ausgewählt und sodann ein Erhebungsdesign entwickelt werden. Im dritten Schritt erfolgt die Erhebung der Daten durch Befragung, wobei die fiktiven Produkte (Stimuli) von den Auskunftspersonen bewertet werden. Aus diesen Daten werden mit Hilfe der Conjoint-Analyse die Teilnutzenwerte geschätzt. Evtl. wird anschließend eine Aggregation der individuellen Nutzenwerte vorgenommen. Während die ersten drei Schritte die *Datenerhebung* betreffen, beziehen sich die Schritte vier und fünf auf die *Datenauswertung*. Die nachfolgenden Betrachtungen sind entsprechend der Darstellung in Abbildung 9.2 aufgebaut.

Stimuli

9.2 Vorgehensweise

9.2.1 Eigenschaften und Eigenschaftsausprägungen



Die durch die Conjoint-Analyse zu ermittelnden Teilnutzenwerte beziehen sich auf einzelne Ausprägungen von Eigenschaften, die der Untersucher für die Analyse vorgeben muss. Bei der Auswahl der Eigenschaften bzw. Eigenschaftsausprägungen sollten folgende Gesichtspunkte beachtet werden:[7]

[6] Einen guten Überblick zur conjointanalytischen Forschung geben Teichert/Shehu (2009), S. 19 ff.
[7] Vgl. zu einer ausführlichen Darstellung Weiber/Mühlhaus (2009), S. 43 f.

Auswahl von Eigenschaften

1. Die Eigenschaften müssen *präferenzrelevant* sein.
   Das bedeutet, dass der Untersucher größte Sorgfalt darauf verwenden muss, nur solche Eigenschaften auszuwählen, von denen zu vermuten ist, dass sie für die Gesamtnutzenbewertung der Befragten von Bedeutung für die Kaufentscheidung sind.

2. Die Eigenschaften müssen durch den Untersucher *beeinflussbar* sein.
   Wenn die Ergebnisse der Conjoint-Analyse z. B. für Produktentscheidungen nutzbar gemacht werden sollen, muss eine Variation der betreffenden Eigenschaften auch im Rahmen der Produktgestaltung möglich sein.

3. Die ausgewählten Eigenschaften sollten *unabhängig* sein.
   Eine Verletzung dieser Bedingung widerspricht dem additiven Modell der Conjoint-Analyse. Präferenzunabhängigkeit der Eigenschaften bedeutet, dass der empfundene Nutzen einer Eigenschaftsausprägung nicht durch die Ausprägungen anderer Eigenschaften beeinflusst wird. Darüber hinaus sollte sichergestellt werden, dass die Eigenschaftsausprägungen auch empirisch unabhängig sind. Darunter wird verstanden, dass die Ausprägungen realiter auch gemeinsam auftreten können bzw. nicht vom Befragten als abhängig voneinander wahrgenommen werden. Insbesondere bei der gemeinsamen Betrachtung von Eigenschaften wie Marke und Preis ist darauf zu achten, dass keine unplausiblen Stimuli-Konstellationen ins Erhebungsdesign integriert werden.

Bestimmung von Eigenschaftsausprägungen

4. Die Eigenschaftsausprägungen müssen *realisierbar* sein.
   Die Nutzbarkeit der Ergebnisse für die Produktgestaltung erfordert, dass die untersuchten Eigenschaftsausprägungen vom Hersteller technisch durchführbar sind.

5. Die einzelnen Eigenschaftsausprägungen müssen in einer *kompensatorischen Beziehung* zueinander stehen.
   Kompensatorische Conjoint-Modelle gehen von der Annahme aus, dass sich die Gesamtbeurteilung eines Objektes durch Summation aller Einzelurteile der als gegenseitig substituierbar angesehenen Eigenschaftsausprägungen ergibt. Das bedeutet, dass in der subjektiven Wahrnehmung der Befragten z. B. eine Verringerung des Kaloriengehaltes einer Margarine durch eine Verbesserung des Geschmacks kompensiert werden kann. Damit wird ein einstufiger Entscheidungsprozess unterstellt, bei dem alle Eigenschaftsausprägungen simultan in die Beurteilung eingehen.[8]

6. Die betrachteten Eigenschaften bzw. Eigenschaftsausprägungen dürfen *keine Ausschlusskriterien* (K.O.-Kriterien) darstellen.
   Ausschlusskriterien liegen vor, wenn bestimmte Eigenschaftsausprägungen für die Auskunftspersonen auf jeden Fall gegeben sein müssen. Im Fall des Vorhandenseins von K.O.-Kriterien wäre das kompensatorische Verhältnis der Eigenschaftsausprägungen untereinander nicht mehr gegeben.

[8] Darüber hinaus existieren auch nicht-kompensatorische Conjoint-Modelle, die eine Kompensation einer negativ beurteilten Eigenschaftsausprägung durch eine positive Bewertung einer anderen Ausprägung nicht zulassen. Da den kompensatorischen Modellen in der Praxis jedoch die größere Bedeutung zukommt, beschränken sich die Betrachtungen im Folgenden auf diesen Modelltyp.

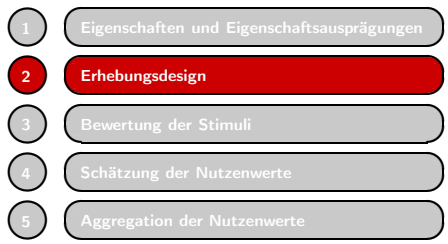
7. Die Anzahl der Eigenschaften und ihrer Ausprägungen muss *begrenzt* werden. Der Befragungsaufwand wächst exponentiell mit der Zahl der Eigenschaftsausprägungen. Deshalb ist es aus erhebungstechnischen Gründen notwendig, sich auf relativ wenige Eigenschaften und je Eigenschaft auf wenige Ausprägungen zu beschränken.

In Erweiterung des Ausgangsbeispiels gehen wir im Folgenden davon aus, dass sich der Margarinehersteller für die in Abbildung 9.3 dargestellten Eigenschaften und Eigenschaftsausprägungen entschieden hat, wobei er vermutet, dass die gewählten Eigenschaften obige Kriterien erfüllen.

Eigenschaften	Eigenschaftsausprägungen
A: Verwendung	1: Brotaufstrich 2: Kochen, Backen, Braten 3: universell
B: Kaloriengehalt	1: kalorienarm 2: normaler Kaloriengehalt
C: Verpackung	1: Becherverpackung 2: Papierverpackung

Abbildung 9.3: Eigenschaften und Eigenschaftsausprägungen

9.2.2 Erhebungsdesign



Im Rahmen der Festlegung des Erhebungsdesigns sind zwei Entscheidungen zu treffen:
1. Definition der Stimuli: Profil- oder Zwei-Faktor-Methode?
2. Zahl der Stimuli: Vollständiges oder reduziertes Design?

9.2.2.1 Definition der Stimuli

Als Stimulus wird eine Kombination von Eigenschaftsausprägungen verstanden, die den Auskunftspersonen zur Beurteilung vorgelegt wird. Bei der *Profilmethode*, die auch als „Full-Profile-Method“ bezeichnet wird, besteht ein Stimulus aus der Kombination je einer Ausprägung aller Eigenschaften. Dadurch können sich in unserem Beispiel in Abbildung 9.3 für die drei Eigenschaften mit jeweils zwei bzw. drei Ausprägungen maximal (2 x 2 x 3 =) 12 Stimuli ergeben, die in Abbildung 9.4 als Übersicht dargestellt sind.

Profilmethode

Margerine I	Margerine II	Margerine III
kalorienarm	kalorienarm	kalorienarm
Becherverpackung	Becherverpackung	Becherverpackung
als Brotaufstrich geeignet	zum Kochen, Backen, Braten	universell anwendbar
Margerine IV	**Margerine V**	**Margerine VI**
normale Kalorien	normale Kalorien	normale Kalorien
Becherverpackung	Becherverpackung	Becherverpackung
als Brotaufstrich geeignet	zum Kochen, Backen, Braten	universell anwendbar
Margerine VII	**Margerine VIII**	**Margerine IX**
kalorienarm	kalorienarm	kalorienarm
Papierverpackung	Papierverpackung	Papierverpackung
als Brotaufstrich geeignet	zum Kochen, Backen, Braten	universell anwendbar
Margerine X	**Margerine XI**	**Margerine XII**
normale Kalorien	normale Kalorien	normale Kalorien
Papierverpackung	Papierverpackung	Papierverpackung
als Brotaufstrich geeignet	zum Kochen, Backen, Braten	universell anwendbar

Abbildung 9.4: Stimuli nach der Profilmethode

Zwei-Faktor-Methode
Trade-Off-Matrix

Bei der *Zwei-Faktor-Methode*, die auch als „Trade-Off-Analyse“ bezeichnet wird, werden zur Bildung eines Stimulus jeweils nur zwei Eigenschaften (Faktoren) herangezogen.[9] Für jedes mögliche Paar von Eigenschaften wird eine Trade-Off-Matrix gebildet. Diese enthält die Kombinationen der Ausprägungen der beiden Eigenschaften. Man erhält damit bei n Eigenschaften insgesamt $\binom{n}{2}$ Trade-Off-Matrizen. In unserem Beispiel ergeben sich damit $\binom{3}{2}$, also 3 Trade-Off-Matrizen, die in Abbildung 9.5 wiedergegeben sind. Jede Zelle einer Trade-Off-Matrix bildet damit einen Stimulus. Die Wahl zwischen Profil- und Zwei-Faktor-Methode sollte im Hinblick auf folgende drei Gesichtspunkte erfolgen:

1. *Ansprüche an die Auskunftsperson:* Da bei der Zwei-Faktor-Methode die Auskunftsperson nur jeweils zwei Faktoren gleichzeitig betrachtet und gegeneinander abwägen muss („trade off“), besteht gegenüber der Profilmethode eine leichter zu bewältigende Bewertungsaufgabe. Die Zwei-Faktor-Methode kann daher auch ohne Interviewereinsatz (z. B. in Form einer schriftlichen Befragung) angewendet werden, während dem mit der Profilmethode verbundenen Erklärungsaufwand nur äußerst schwer in einem Fragebogen Rechnung zu tragen ist.

2. *Realitätsbezug:* Da beim realen Beurteilungsprozess i. d. R. komplette Produkte und nicht isolierte Eigenschaften miteinander verglichen werden, liefert die Profilmethode ein realitätsnäheres Design. Außerdem können die Stimuli nicht nur in schriftlicher Form, sondern auch als anschauliche Abbildungen oder Objekte vorgegeben werden.

3. *Zeitaufwand:* Mit zunehmender Anzahl der Eigenschaften und ihrer Ausprägungen steigt die Zahl möglicher Stimuli bei der Profilmethode wesentlich schneller als bei der Zwei-Faktor-Methode, wodurch eine sinnvolle Bewertung aller Stimuli durch die Auskunftsperson u. U. unmöglich werden kann.

[9] Die Zwei-Faktor-Methode geht zurück auf Johnson (1974), S. 121 ff.

	B: Kaloriengehalt	
A: Verwendung	1: kalorienarm	2: normale Kalorien
1: Brotaufstrich	A1B1	A1B2
2: Kochen, Backen, Braten	A2B1	A2B2
3: universell	A3B1	A3B2

	C: Verpackung	
A: Verwendung	1: Becherverpackung	2: Papierverpackung
1: Brotaufstrich	A1C1	A1C2
2: Kochen, Backen, Braten	A2C1	A2C2
3: universell	A3C1	A3C2

	C: Verpackung	
B: Kaloriengehalt	1: Becherverpackung	2: Papierverpackung
1: kalorienarm	B1C1	B1C2
2: normale Kalorien	B2C1	B2C2

Abbildung 9.5: Trade-Off-Matrizen

In der Regel steht bei Anwendungen der Conjoint-Analyse der Realitätsbezug im Vordergrund, sodass meist der Profilmethode der Vorzug gegeben wird. Der Gesichtspunkt des Zeitaufwandes, der tendenziell für die Zwei-Faktor-Methode spricht, wird allerdings durch die Tatsache relativiert, dass die Möglichkeit existiert, bei der Profilmethode aus allen möglichen Stimuli eine repräsentative Teilmenge auszuwählen, wobei sich der Zeitaufwand bei der Profilmethode durch die Anwendung eines sog. reduzierten Designs wesentlich reduzieren lässt. Im Folgenden steht daher die *Profilmethode* im Vordergrund der Betrachtungen.

9.2.2.2 Zahl der Stimuli

In vielen empirischen Untersuchungen besteht der Wunsch, mehr Eigenschaften und/oder Ausprägungen zu analysieren als erhebungstechnisch realisierbar sind. Dies ist insbesondere bei der Profilmethode der Fall. Bereits bei sechs Eigenschaften mit jeweils nur drei Ausprägungen ergeben sich $(3^6 =)729$ Stimuli, was erhebungstechnisch nicht sinnvoll zu bewältigen ist. Daraus erwächst die Notwendigkeit, aus der Menge der theoretisch möglichen Stimuli (*vollständiges Design*) eine zweckmäßige Teilmenge (*reduziertes Design*) auszuwählen.

Vollständiges Design

Die Grundidee eines reduzierten Designs besteht darin, eine Teilmenge von Stimuli zu finden, die das vollständige Design möglichst gut repräsentiert. Beispielsweise könnte eine Zufallsstichprobe gezogen werden. Davon wird jedoch in der Regel nicht Gebrauch gemacht, sondern es wird eine systematische Auswahl der Stimuli vorgenommen. In der experimentellen Forschung ist eine Reihe von Verfahren entwickelt worden, die zur Lösung dieses Problems herangezogen werden kann. Dabei wird zwischen symmetrischen und asymmetrischen Designs unterschieden:

Reduziertes Design

Ein *symmetrisches Design* liegt vor, wenn alle Eigenschaften die gleiche Anzahl von Ausprägungen aufweisen. Ein spezielles reduziertes symmetrisches Design ist das lateinische Quadrat. Seine Anwendung ist auf den Fall von genau drei Eigenschaften beschränkt. Das vollständige Design, das dem lateinischen Quadrat zugrunde liegt,

Symmetrisches Design

Lateinisches Quadrat

umfasst z. B. im Fall von drei Ausprägungen je Eigenschaft (3 x 3 x 3 =) 27 Stimuli, die in Abbildung 9.6 dargestellt sind.

A1B1C1	A2B1C1	A3B1C1
A1B2C1	A2B2C1	A3B2C1
A1B3C1	A2B3C1	A3B3C1
A1B1C2	A2B1C2	A3B1C2
A1B2C2	A2B2C2	A3B2C2
A1B3C2	A2B3C2	A3B3C2
A1B1C3	A2B1C3	A3B1C3
A1B2C3	A2B2C3	A3B2C3
A1B3C3	A2B3C3	A3B3C3

Abbildung 9.6: Vollständiges faktorielles Design

Von den 27 Stimuli des vollständigen Designs werden 9 derart ausgewählt, dass jede Ausprägung einer Eigenschaft genau einmal mit jeder Ausprägung einer anderen Eigenschaft berücksichtigt wird. Damit ergibt sich, dass jede Eigenschaftsausprägung genau dreimal (statt neunmal) im Design vertreten ist. Abbildung 9.7 zeigt das entsprechende Design.

	A1	A2	A3
B1	A1 B1 C1	A2 B1 C2	A3 B1 C3
B2	A1 B2 C2	A2 B2 C3	A3 B2 C1
B3	A1 B3 C3	A2 B3 C1	A3 B3 C2

Abbildung 9.7: Lateinisches Quadrat

Asymmetrisches Design

Die Reduzierung von *asymmetrischen Designs*, in denen die verschiedenen Eigenschaften eine unterschiedliche Anzahl von Ausprägungen aufweisen, wie das (2 x 2 x 3)-faktorielle Design des Margarinebeispiels, ist wesentlich komplizierter.[10] Auch hier wurden Pläne zur Konstruktion reduzierter Designs entwickelt, die auch in SPSS implementiert sind, sodass eine aufwändige Konstruktion per Hand heutzutage entfällt.

Vereinfachung des Beispiels

Da im Folgenden die konkreten Rechenschritte der Conjoint-Analyse im Einzelnen aufgezeigt werden sollen, nehmen wir nochmals eine Modifikation unseres Margarinebeispiels vor und beschränken die nachfolgenden Betrachtungen auf die Eigenschaften „Verwendung“ und „Kaloriengehalt“ aus Abbildung 9.3. Durch Kombination aller Eigenschaftsausprägungen erhält man dann die in Abbildung 9.8 aufgeführten sechs Stimuli (fiktiven Produkte):

I	A1, B1	Brotaufstrich	Kalorienarm
II	A1, B2	Brotaufstrich	normale Kalorien
III	A2, B1	Kochen, Backen, Braten	Kalorienarm
IV	A2, B2	Kochen, Backen, Braten	normale Kalorien
V	A3, B1	universell verwendbar	Kalorienarm
VI	A3, B2	universell verwendbar	normale Kalorien

Abbildung 9.8: Stimuli im vollständigen Design für das Margarinebeispiel

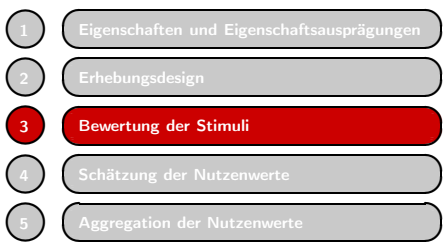
[10] Vgl. zur Konstruktion reduzierter asymmetrischer Designs Addelman (1962), S. 21 ff. oder Street/Burgess/Louviere (2005), S. 460 ff.

Die sechs fiktiven Produkte bilden ein vollständiges Design, wodurch auf eine Reduktion dieses Designs verzichtet werden kann, da davon auszugehen ist, dass sechs Stimuli von den Auskunftspersonen ohne Probleme in eine Präferenzrangfolge gebracht werden können. Damit folgt ein vollständiges, zweistufiges Untersuchungsdesign, das in Abbildung 9.9 dargestellt ist. Durch p sind dabei die empirischen Rangwerte der jeweiligen Stimuli bezeichnet, die im Rahmen der Untersuchung erhoben werden müssen.

		Eigenschaft B	
		1	2
Eigenschaft A	1	p_I	p_{II}
	2	p_{III}	p_{IV}
	3	p_V	p_{VI}

Abbildung 9.9: Vollständiges Untersuchungsdesign für das Beispiel

9.2.3 Bewertung der Stimuli



Die Conjoint-Analyse erfordert, dass eine Präferenzordnung der Stimuli entsprechend den Nutzenvorstellungen der Auskunftsperson ermittelt wird. Dazu bieten sich verschiedene Vorgehensweisen an. Üblich ist die Erhebung über *Rangreihung*. Dabei werden die Stimuli nach empfundenem Nutzen mit Rangwerten versehen. Bei einer größeren Anzahl von Stimuli empfiehlt sich eine indirekte Vorgehensweise. Es erfolgt zunächst eine Grobeinteilung in Gruppen unterschiedlichen Nutzens (z. B. niedriger, mittlerer, hoher Nutzen).

Rangreihung

Innerhalb der Gruppen werden Rangfolgen der einzelnen Stimuli ermittelt, die dann zur Gesamtrangordnung zusammengefasst werden. Weitere Möglichkeiten bestehen darin, die Rangwerte über Rating-Skalen oder Paarvergleiche zu ermitteln.[11]

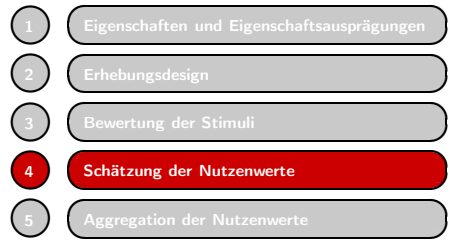
Für unser Beispiel (vgl. Abbildung 9.8) wurde eine Person gebeten, die sechs möglichen Margarinesorten mit Rangwerten von 1 bis 6 zu versehen, wobei 1 der am wenigsten und 6 der am stärksten präferierte Stimulus sein sollte. Das Ergebnis der Rangreihung zeigt Abbildung 9.10:

[11] Die Vorgehensweise der Ermittlung einer Rangfolge durch Paarvergleiche findet insbesondere im Rahmen verschiedener computergestützter Conjointanalyseverfahren Anwendung. Eine ausführliche Darstellung der verschiedenen Möglichkeiten der Abfrage liefert Hensel-Börner (2000), Kap. 3.

		Eigenschaft B	
		1	2
Eigenschaft A	1	2	1
	2	3	4
	3	6	5

Abbildung 9.10: Rangwerte für eine Auskunftsperson im Beispiel

9.2.4 Schätzung der Nutzenwerte



Teilnutzenwerte

Auf Basis der empirisch ermittelten Rangdaten einer Menge von Stimuli werden mit Hilfe der Conjoint-Analyse zunächst Teilnutzenwerte (partworths) für alle Eigenschaftsausprägungen ermittelt. Aus diesen Teilnutzenwerten lassen sich dann folgende Größen ableiten:

- metrische *Gesamtnutzenwerte* für alle Stimuli
- relative *Wichtigkeiten* der einzelnen Eigenschaften

Die Schätzung der Teilnutzenwerte wird nachfolgend anhand unseres Beispiels aus Abbildung 9.8 dargestellt. Den Berechnungen legen wir die Beurteilungen der Auskunftsperson entsprechend Abbildung 9.10 zugrunde.

Für jede der insgesamt fünf Eigenschaftsausprägungen ist jetzt ein Teilnutzenwert β zu schätzen. Aus der Verknüpfung der Teilnutzenwerte ergibt sich dann der Gesamtnutzenwert y eines Stimulus. Im einfachsten Fall wird daher das folgende additive Modell zugrunde gelegt:

$$y = \beta_A + \beta_B \tag{9.1}$$

Additives Modell der Conjoint-Analyse

In allgemeiner Form lässt sich das *additive Modell der Conjoint-Analyse* wie folgt formulieren:

$$y_k = \sum_{j=1}^{J} \sum_{m=1}^{M_j} \beta_{jm} \cdot x_{jmk} \tag{9.2}$$

mit

y_k = geschätzter Gesamtnutzenwert für Stimulus k

β_{jm} = Teilnutzenwert für Ausprägung m von Eigenschaft j

$$x_{jmk} = \begin{cases} 1 & \text{falls bei Stimulus k die Eigenschaft j in Ausprägung m vorliegt} \\ 0 & \text{sonst} \end{cases}$$

Das additive Modell, das in der Conjoint-Analyse vornehmlich Anwendung findet, besagt, dass die Summe der Teilnutzen den Gesamtnutzen ergibt. Durch Anwendung

dieses Modells ergeben sich im Beispiel die folgenden Gesamtnutzenwerte (vgl. Abbildung 9.9):

$$\begin{aligned} y_I &= \beta_{A1} + \beta_{B1} \\ y_{II} &= \beta_{A1} + \beta_{B2} \\ y_{III} &= \beta_{A2} + \beta_{B1} \\ y_{IV} &= \beta_{A2} + \beta_{B2} \\ y_V &= \beta_{A3} + \beta_{B1} \\ y_{VI} &= \beta_{A3} + \beta_{B2} \end{aligned}$$

Zielkriterium zur Bestimmung der Teilnutzenwerte

Das zur Bestimmung der Teilnutzenwerte verwendete *Zielkriterium* lässt sich wie folgt formulieren:

Die Teilnutzenwerte β_{jm} sollen so bestimmt werden, dass die resultierenden Gesamtnutzenwerte y_k „möglichst gut“ den empirischen Rangwerten p_k entsprechen. Das Zielkriterium wird im Folgenden näher spezifiziert.

Das zur Ermittlung der Teilnutzenwerte üblicherweise verwendete Rechenverfahren wird als monotone Varianzanalyse bezeichnet. Es bildet eine Weiterentwicklung der gewöhnlichen (metrischen) Varianzanalyse, die in Kapitel 3 dieses Buches behandelt wird.

9.2.4.1 Metrische Lösung

Metrische Varianzanalyse

Das Problem der Conjoint-Analyse soll zunächst durch Anwendung der metrischen Varianzanalyse gelöst werden. Dabei wird unterstellt, dass die Befragten die Abstände zwischen den vergebenen Rangwerten jeweils als gleich groß (äquidistant) einschätzen, womit die empirisch ermittelten p-Werte nicht mehr ordinales Skalenniveau besitzen, sondern metrisch interpretiert werden können. Das Modell (9.1) muss dabei durch Einbeziehung eines konstanten Terms μ wie folgt modifiziert werden:

$$y = \mu + \beta_A + \beta_B \tag{9.3}$$

Basisnutzen

Die Konstante μ spiegelt dabei den „Durchschnittsrang“ über alle vergebenen (metrischen) Rangwerte wider. Die Konstante μ kann auch als Basisnutzen interpretiert werden, von dem sich die Eigenschaftsausprägungen positiv oder negativ abheben. Für unser Beispiel ergibt sich als Summe über alle sechs empirischen Rangdaten (vgl. Abbildung 9.10) $1 + 2 + 3 + 4 + 5 + 6 = 21$ und damit ein „Durchschnittsrang“ von $21/6 = 3,5$. Zur Bestimmung der einzelnen Teilnutzenwerte wird im zweiten Schritt für jede Eigenschaftsausprägung der durchschnittliche empirische Rangwert ermittelt. Zu diesem Zweck wird für jede Eigenschaftsausprägung geprüft, welche Rangdaten der Befragte in Verbindung mit dieser Eigenschaft vergeben hat und daraus der Durchschnitt gebildet. Betrachtet man Abbildung 9.10 so hat die Auskunftsperson z. B. bei Eigenschaftsausprägung A1 die Rangwerte 2 und 1 vergeben, woraus sich eine Durchschnittseinschätzung von $3/2 = 1,5$ ergibt. Damit bleibt die durchschnittliche Einschätzung der Eigenschaftsausprägung A1 aber hinter dem „Durchschnittsrang“ von 3,5 zurück, d. h. sie liefert einen geringeren Teilnutzenwert als der Durchschnitt. Das Ausmaß, in dem Eigenschaftsausprägung A1 hinter dem Durchschnittsrang zurückbleibt, ergibt sich durch einfache Differenzbildung und beträgt $(1,5 - 3,5) = -2,0$. Dieser Differenzwert stellt den Teilnutzenwert der Eigenschaftsausprägung A1 dar. Entsprechend wird mit allen anderen Eigenschaftsausprägungen verfahren. Abbildung 9.11 zeigt das entsprechende Berechnungstableau auf.

		Eigenschaft B		$\bar{p}_A$	$\bar{p}_A - \bar{p}$
		1	2		
Eigenschaft A	1	2	1	1,5	-2,0
	2	3	4	3,5	0,0
	3	6	5	5,5	2,0
$\bar{p}_B$		3,6667	3,3333	3,5	
$\bar{p}_B - \bar{p}$		0,1667	-0,1667		

Abbildung 9.11: Berechnungstableau der metrischen Varianzanalyse

Anmerkung: Ein Teilnutzenwert ergibt sich allgemein durch $\beta_j = \bar{p}_j - \bar{p}$, wobei $\bar{p}_j$ den Mittelwert einer Zeile oder Spalte und $\bar{p}$ das Gesamtmittel der p-Werte bezeichnet.

Abbildung 9.11 enthält in der letzten Spalte und Zeile die empirischen Schätzwerte (Teilnutzenwerte), die nachfolgend nochmals zusammengefasst sind.

$$\mu = 3,5 \qquad \begin{aligned} \beta_{A1} &= -2,000 \\ \beta_{A2} &= 0,000 \\ \beta_{A3} &= 2,000 \end{aligned} \qquad \begin{aligned} \beta_{B1} &= 0,1667 \\ \beta_{B2} &= -0,1667 \end{aligned}$$

Damit ergibt sich beispielsweise für Stimulus I ein Gesamtnutzenwert von:

$$Y_I = 3,5 + (-2,0) + 0,1667 = 1,1667$$

In Abbildung 9.12 sind die empirischen und geschätzten Nutzenwerte sowie deren einfache und quadrierte Abweichungen zusammengefasst:

Stimulus	p	y	$p - y$	$(p - y)^2$
I	2	1,6667	0,333	0,1111
II	1	1,3333	-0,333	0,1111
III	3	3,6667	-0,667	0,4444
IV	4	3,3333	0,667	0,4444
V	6	5,6667	0,333	0,1111
VI	5	5,3333	-0,333	0,1111
	21	21,0000	0,000	1,3333

Abbildung 9.12: Ermittlung der quadratischen Abweichungen zwischen den empirischen und geschätzten Nutzenwerten

Kleinst-Quadrate-Schätzungen

Die durch Anwendung der Varianzanalyse ermittelten Teilnutzenwerte β sind Kleinst-Quadrate-Schätzungen, d. h. sie wurden so ermittelt, dass die Summe der quadratischen Abweichungen zwischen den empirischen und geschätzten Nutzenwerten minimal ist:

$$\underset{\beta}{\text{Min}} \sum_{k=1}^{K} (p_k - y_k)^2 \tag{9.4}$$

Zu der gleichen Lösung gelangt man auch durch Anwendung einer Regressionsanalyse (vgl. Kapitel 1 in diesem Buch) der p-Werte auf die 0/1-Variablen (Dummy-Variablen) x_{jm} in Formel (9.2). Eine derartige Dummy-Regression wird im Rahmen der Conjoint-Analyse häufig angewendet.[12]

Dummy-Regression

9.2.4.2 Nichtmetrische Lösung

Lässt man die Annahme metrisch skalierter Ausgangswerte fallen und beschränkt sich auf die Annahme ordinal skalierter p-Werte, so gewinnt man größeren Spielraum für die Lösung des Problems einer optimalen Schätzung der Teilnutzenwerte. Dieser Spielraum kann durch Anwendung der *monotonen Varianzanalyse* genutzt werden. Die Art der Ergebnisse und deren Interpretation ändern sich dabei nicht.

Monotone Varianzanalyse

Die von Kruskal entwickelte monotone Varianzanalyse bildet ein iteratives Verfahren und ist somit bedeutend rechenaufwändiger als die metrische Varianzanalyse.[13] Die metrische Lösung kann als Ausgangspunkt für den Iterationsprozess verwendet werden.[14] Das Prinzip der *monotonen Varianzanalyse* lässt sich wie folgt darstellen:

Monotone Varianzanalyse

$$p_k \overset{f_M}{\rightarrow} z_k \cong y_k = \sum_{j=1}^{J} \sum_{m=1}^{M_j} \beta_{jm} \cdot x_{jm} \tag{9.5}$$

mit

p_k	=	empirische Rangwerte der Stimuli (k=1,...,K)
z_k	=	monoton angepasste Rangwerte
y_k	=	metrische Gesamtnutzenwerte, die durch das additive Modell 9.2 gewonnen wurden.
f_M	=	monotone Transformation zur Anpassung der z-Werte an die y- Werte
$\cong$	=	bedeutet möglichst gute Anpassung im Sinne des Kleinst-Quadrate-Kriteriums

Die monotone Varianzanalyse unterscheidet sich von der metrischen Varianzanalyse dadurch, dass die Anpassung der y-Werte (durch Schätzung der Teilnutzenwerte (β)) nicht direkt an die empirischen p-Werte erfolgt, sondern indirekt über die z-Werte. Diese müssen der nachstehenden Monotoniebedingung folgen:

$$z_k \leq z_{k'} \qquad \text{für} \qquad p_k < p_{k'} \quad \text{(schwache Monotonie)} \tag{9.6}$$

[12] Vgl. dazu auch die Ausführungen in den Anwendungsempfehlungen dieses Kapitels.

[13] Zur monotonen Varianzanalyse, die auch der in Kapitel 15 behandelten Multidimensionalen Skalierung zugrunde liegt, vgl. insbesondere Kruskal (1965), S. 251 ff.

[14] Da das Verfahren gegen suboptimale Lösungen (lokale Optima) konvergieren kann, ist es von Vorteil, den Iterationsprozess wiederholt mit verschiedenen Ausgangslösungen zu starten. Während das Programm MONANOVA mit einer metrischen Ausgangslösung beginnt, enthält das Programm UNICON eine Option zur Generierung von unterschiedlichen Ausgangslösungen durch einen Zufallsgenerator.

Das Zielkriterium der monotonen Varianzanalyse beinhaltet daher im Unterschied zu Formel (9.4) eine Minimierung der Abweichungen zwischen z und y. Es lautet wie folgt:

STRESS-Maß

Zielkriterium der monotonen Varianzanalyse (STRESS-Maß)

$$\min_{f_M} \min_{\beta} \text{STRESS} = \min_{f_M} \min_{\beta} \sqrt{\frac{\sum_{k=1}^{K}(z_k - y_k)^2}{\sum_{k=1}^{K}(y_k - \bar{y})^2}} \tag{9.7}$$

Das Zentrum des STRESS-Maßes bildet das Kleinst-Quadrate-Kriterium im Zähler der Wurzel. Der Nenner dient lediglich als Skalierungsfaktor und bewirkt, dass lineare Transformationen der z-Werte (und damit der angepassten y-Werte) keinen Einfluss auf die Größe „STRESS“ haben. Die Wurzel selbst soll nur der besseren Interpretation dienen und hat keinen Einfluss auf die Lösung.

Das Zielkriterium erfordert eine zweifache Optimierung, nämlich über die Transformation f_m, die die Bedingung in Formel (9.6) erfüllen muss und über die Teilnutzenwerte β. Es kommen daher auch zwei verschiedene Rechenverfahren zur Anwendung.

Wechselseitig erfolgt für eine

Gradientenverfahren

- gegebene Transformation f_M:
  Anpassung von y an z durch Auffindung von Teilnutzenwerten β (*Gradientenverfahren*).
- gegebene Menge von β-Werten:
  Anpassung von z an y durch Auffinden einer monotonen Transformation f_M (*monotone Regression*).

Das zur Optimierung über β herangezogene Gradientenverfahren (Methode des steilsten Anstiegs) ist ein iteratives Verfahren.[15] Bei jedem Schritt dieses Verfahrens werden für die gefundenen Teilnutzenwerte β die resultierenden Gesamtnutzenwerte y_k berechnet und sodann die Werte z_k durch monotone Regression (von p auf y) optimal angepasst. Abbildung 9.13 veranschaulicht den Ablauf.

[15] Vgl. Kruskal (1965), S. 261 f.; Kruskal (1964b), S. 119 ff.

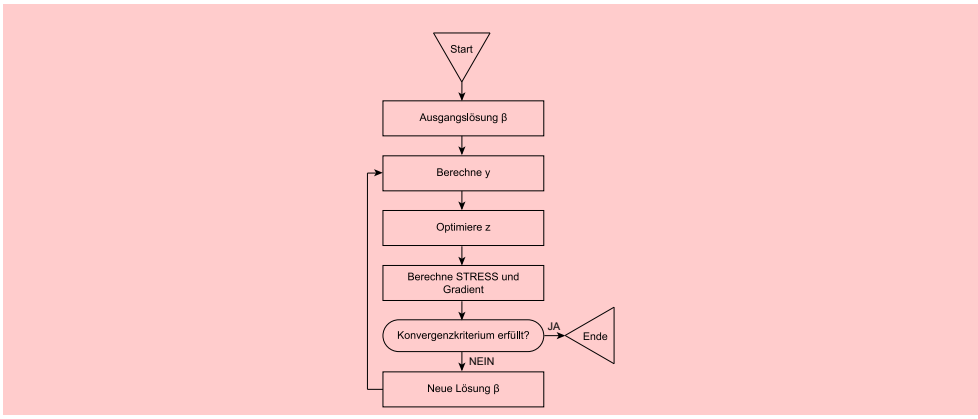


Abbildung 9.13: Ablauf der monotonen Varianzanalyse

9.2.4.3 Monotone Regression

Monotone Regression

Unter dem Begriff der monotonen Regression, die als Baustein der monotonen Varianzanalyse dient, verbirgt sich ein im Prinzip sehr einfaches Verfahren. Die Abbildung 9.14 und Abbildung 9.15 dienen zur Veranschaulichung.

In Abbildung 9.14 sind die in Abschnitt 9.2.4.1 durch metrische Varianzanalyse ermittelten Gesamtnutzenwerte y_k über den empirischen Rangwerten der sechs Stimuli eingetragen (vgl. Abbildung 9.12).

Wie man sieht, ist der sich ergebende Verlauf nicht monoton. Die y-Werte für Stimulus III und IV verletzen die Monotoniebedingung in Formel (9.6); denn es gilt:

$$y_{III} > y_{IV} \qquad \text{aber} \quad p_{III} < p_{IV}$$

Durch monotone Regression von y über p werden jetzt monoton angepasste Werte z, die optimal im Sinne des Kleinst-Quadrate-Kriteriums sind, wie folgt angepasst:

- Es wird $z_k = y_k$ gesetzt, wenn y_k die Monotoniebedingung (bezüglich aller übrigen y-Werte) erfüllt.
- Verletzten zwei Werte y_k und $y_{k'}$ die Monotoniebedingung, so wird deren Mittelwert gebildet und den z-Werten zugeordnet:

  $z_k = z_{k'} = \frac{y_k + y_{k'}}{2}$

Analog wird verfahren, wenn mehr als zwei y-Werte die Monotoniebedingung verletzen.

Abbildung 9.15 zeigt das Ergebnis der monotonen Regression. Die erhaltenen z-Werte sind nicht nur optimal im Sinne des Kleinst-Quadrate-Kriteriums, sondern sie minimieren auch das STRESS-Maß in Formel (9.7), da der Nenner unter der Wurzel bei der monotonen Anpassung konstant bleibt. Wenn alle y-Werte die Monotoniebedingung erfüllen, ergibt sich für den STRESS der Wert Null („perfekte Lösung“). In diesem Fall erübrigt sich eine monotone Regression.

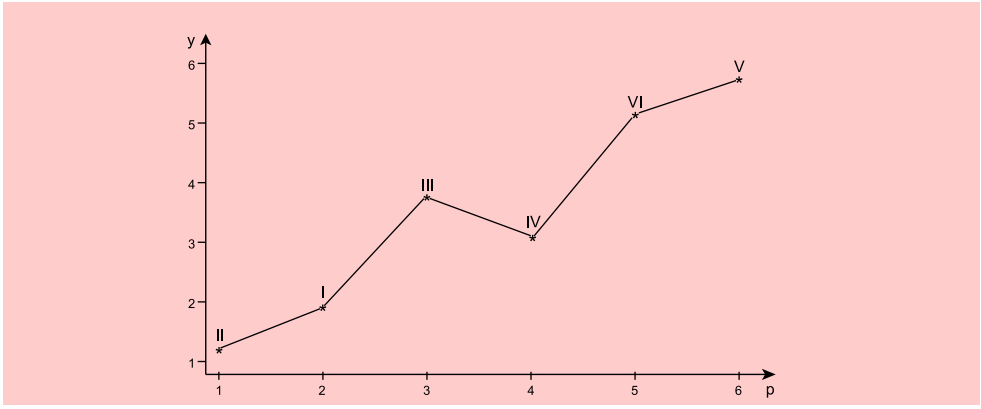


Abbildung 9.14: Verlauf der geschätzten y-Werte über den empirischen Rangdaten

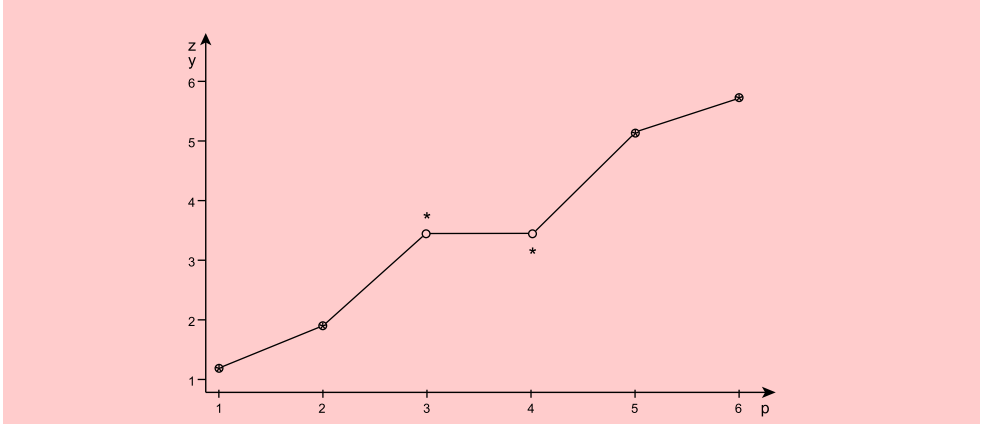


Abbildung 9.15: Verlauf der monoton angepassten z-Werte über den empirischen Rangdaten

Ties

Wenn sogenannte *Ties* unter den empirischen Rangwerten auftreten, d. h. wenn gleiche Rangwerte vorkommen, sind bei der monotonen Regression zwei alternative Vorgehensweisen möglich. Kruskal, von dem diese Einteilung stammt, erscheint der Primary Approach als die geeignetere Vorgehensweise.[16]

- *Primary Approach:*
  Aus $p_k = p_{k'}$ folgt keine Einschränkung für z_k und $z_{k'}$.
- *Secondary Approach:*
  Aus $p_k = p_{k'}$ folgt die Bedingung $z_k = z_{k'}$.

9.2.4.4 Fehlende Rangdaten

Reduziertes Design

Es wurde bereits darauf hingewiesen, dass bei größerer Anzahl von Eigenschaften und Eigenschaftsausprägungen *unvollständige (reduzierte) Untersuchungsdesigns* angewendet werden müssen, um den Erhebungsaufwand in Grenzen zu halten und

[16] Vgl. Kruskal (1964a), S. 21 ff.

eine Überforderung der Versuchspersonen zu vermeiden. Bei unvollständigen Untersuchungsdesigns werden nur für eine systematisch gebildete Teilmenge aus der Gesamtmenge der Stimuli des vollständigen Designs Rangdaten erhoben.

Bei empirischen Untersuchungen ist es weiterhin unvermeidbar, dass ungewollt fehlende Daten, sog. *Missing Values* auftreten, z. B. als Folge von Erhebungsfehlern oder weil die Auskunftspersonen nicht antworten können oder wollen. Auch aus diesen Gründen können bei der Durchführung einer Conjoint-Analyse Rangdaten fehlen.

Missing Values

Das Prinzip der Behandlung fehlender Rangdaten ist sehr einfach: Bei der Berechnung der STRESS-Formel, wie auch bei Durchführung der monotonen Regression, werden nur diejenigen Stimuli berücksichtigt, für die empirische Rangdaten vorliegen. Daher ist es gleichgültig, ob die Rangdaten als Missing Values oder infolge eines unvollständigen Designs fehlen.

Bei der Dateneingabe in ein Programm müssen für fehlende Daten *Füllwerte* eingegeben werden.

Beispiel:

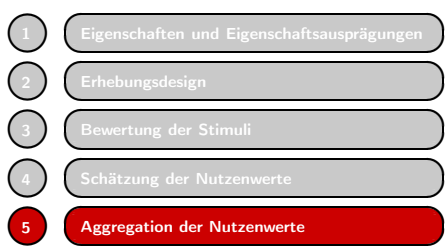
vollständige Rangdaten:	2, 1, 3, 4, 6, 5
unvollständige Rangdaten:	2, 0, 3, 4, 0, 5

Die fehlenden Daten werden jeweils durch eine Null ersetzt. Die Null kann dabei als Füllwert durch das Programm vorgegeben oder vom Benutzer (durch Spezifizierung eines Cut-off-Wertes) gewählt werden.

Natürlich dürfen nicht zu viele Rangdaten fehlen, damit eine Ermittlung der zugrundeliegenden Nutzenstruktur möglich ist. Andernfalls kann es sein, dass das Verfahren „zusammenbricht“. Man erhält dann einen minimalen STRESS-Wert von Null, obgleich die ermittelten Teilnutzenwerte bedeutungslos sind und eine degenerierte Lösung darstellen.

Degenerierte Lösung

9.2.5 Aggregation der Nutzenwerte



Die bisherigen Betrachtungen haben verdeutlicht, wie mit Hilfe der Conjoint-Analyse die Nutzenstruktur einer einzelnen Person analysiert werden kann. Sollen jedoch die Individualanalysen der einzelnen Auskunftspersonen miteinander verglichen werden, so ist dies nur möglich, wenn zunächst über eine entsprechende Normierung eine Vergleichbarkeit herbeigeführt wird. Durch die Normierung muss sichergestellt werden, dass die errechneten Teilnutzenwerte für alle Befragten jeweils auf dem gleichen „Nullpunkt“ und gleichen Skaleneinheiten basieren.

Normierung

Bezüglich des Nullpunktes ist es sinnvoll, diejenige Eigenschaftsausprägung, die den geringsten Nutzenbeitrag liefert, auf Null zu setzen.

Für die Normierungsvorschrift folgt daraus, dass im ersten Schritt jeweils die Differenz zwischen den einzelnen Teilnutzenwerten und dem kleinsten Teilnutzenwert der entsprechenden Eigenschaft zu bilden ist, was sich formal durch folgende

Transformation beschreiben lässt:

$$\beta^*_{jm} = \beta_{jm} - \beta^{Min}_j \tag{9.8}$$

mit

β_{jm} = Teilnutzenwert für Ausprägung m von Eigenschaft j
β^{Min}_j = minimaler Teilnutzenwert bei Eigenschaft j

Für die in unserem Beispiel errechneten Werte (vgl. Abschnitt 9.2.4.1) ergeben sich damit folgende transformierte Teilnutzenwerte:

$\beta^*_{A1} = (-2,000 - (-2,000)) = 0,000$ $\quad$ $\beta^*_{B1} = (0,1667 - (-0,1667)) = 0,3334$
$\beta^*_{A2} = (0,000 - (-2,000)) = 2,000$ $\quad$ $\beta^*_{B2} = (-0,1667 - (-0,1667)) = 0,0000$
$\beta^*_{A3} = (2,000 - (-2,000)) = 4,000$

Normierte Teilnutzenwerte

Für die *Justierung der Skaleneinheit* ist entscheidend, welche Größe den Maximalwert des Wertebereichs beschreiben soll. Da die Conjoint-Analyse je Eigenschaft versucht, die Nutzenbeiträge der einzelnen, sich gegenseitig ausschließenden Eigenschaftsausprägungen zu schätzen, ergibt sich für einen Befragten der am stärksten präferierte Stimulus aus der Summe der höchsten Teilnutzenwerte je Eigenschaft. Die Summe der maximalen Teilnutzenwerte je Eigenschaft ist damit gleich dem Maximalwert des Wertebereichs. Alle anderen Kombinationen von Eigenschaftsausprägungen (Stimuli) führen zu kleineren Gesamtnutzenwerten. Es ist deshalb zweckmäßig, den Gesamtnutzenwert des am stärksten präferierten Stimulus bei allen Auskunftspersonen auf 1 zu setzen. Damit ergeben sich die *normierten Teilnutzenwerte* wie folgt:

$$\hat{\beta}_{jm} = \frac{\beta^*_{jm}}{\sum_{j=1}^{J} \max_m \{\beta^*_{jm}\}} \tag{9.9}$$

Für das Margarinebeispiel ergeben sich folgende normierte Teilnutzenwerte:

$\hat{\beta}_{A1} = 0,000/4,3334 = 0,000$ $\quad$ $\hat{\beta}_{B1} = 0,3334/4,3334 = 0,077$
$\hat{\beta}_{A2} = 2,000/4,3334 = 0,462$ $\quad$ $\hat{\beta}_{B2} = 0,0000/4,3334 = 0,000$
$\hat{\beta}_{A3} = 4,000/4,3334 = 0,923$

Es wird deutlich, dass das am stärksten präferierte Produkt einen Gesamtnutzenwert von 1 erhält und hier in der Kombination aus universeller Verwendbarkeit (A3) und armem Kaloriengehalt (B1) besteht, was Stimulus V aus Abbildung 9.8 entspricht.

Relative Wichtigkeit einer Eigenschaft zur Präferenzänderung

An dieser Stelle sei darauf hingewiesen, dass sich aus der absoluten Höhe der Teilnutzenwerte zwar auf die Bedeutsamkeit einer Eigenschaftsausprägung für den Gesamtnutzenwert eines Stimulus schließen lässt, *nicht* aber auf die *relative Wichtigkeit*. Hat beispielsweise eine Eigenschaft im Vergleich zu einer anderen durchgängig hohe Teilnutzenwerte für alle Eigenschaftsausprägungen, dann lässt sich daraus *nicht* schließen, dass diese Eigenschaft für die Präferenz*veränderung* wichtiger ist als die andere. Es gehen zwar hohe Nutzenwerte in den Gesamtnutzenwert ein, jedoch tragen diese hohen Werte *für jede Eigenschaftsausprägung gleichermaßen* zum Gesamtnutzenwert bei, sodass eine Variation der Ausprägung dieser Eigenschaft keinen bedeutsamen Einfluss auf die Höhe des Gesamtnutzenwertes ausübt. Entscheidend für die Bedeutung einer Eigenschaft zur Präferenz*veränderung* ist vielmehr die *Spannweite*, d. h. die

Differenz zwischen dem höchsten und dem niedrigsten Teilnutzenwert der verschiedenen Ausprägungen jeweils einer Eigenschaft. Ist die Spannweite groß, dann kann durch eine Variation der betreffenden Eigenschaft eine bedeutsame Veränderung des Gesamtnutzenwertes erfolgen. Gewichtet man die Spannweite einzelner Eigenschaften an der Summe der Spannweiten, so erhält man die Bedeutung einzelner Eigenschaften für die Präferenzvariation. Die *relative Wichtigkeit* einer Eigenschaft lässt sich damit entsprechend Formel (9.10) bestimmen:

Relative Wichtigkeit und Spannweite

$$w_j = \frac{\max\limits_m\{\beta_{jm}\} - \min\limits_m\{\beta_{jm}\}}{\sum_{j=1}^{J}\left(\max\limits_m\{\beta_{jm}\} - \min\limits_m\{\beta_{jm}\}\right)} \tag{9.10}$$

Wird Formel (9.10) bei normierten Teilnutzenwerten verwendet (vgl. Formel (9.9)), so ist der Ausdruck $\min\limits_m\{\beta_{jm}\}$ in Zähler und Nenner der Formel (9.10) *immer* gleich Null. In diesem Fall sind Formel (9.9) und (9.10) mithin identisch. Damit liefern die *größten normierten* Teilnutzenwerte je Eigenschaft gleichzeitig auch eine Aussage über die relative Wichtigkeit der Eigenschaften. Für die in unserem Beispiel betrachtete Auskunftsperson besitzt die Eigenschaft A (Verwendbarkeit) mit 92,3 % gegenüber der Eigenschaft B (Kaloriengehalt) mit nur 7,7 % ein weit stärkeres Gewicht für die Präferenzbildung.

Durch die Normierung gemäß Formel (9.9) ist nun auch eine *Vergleichbarkeit* der Ergebnisse aus verschiedenen Individualanalysen sichergestellt. In vielen Fällen interessieren den Untersucher nämlich vor allem die aggregierten Nutzenwerte für eine Mehrzahl von Individuen. So ist es z. B. für einen Anbieter in der Regel ausreichend, wenn er die mittlere Nutzenstruktur seiner potenziellen Käufer oder für Segmente von Käufern kennt. Es existieren zwei grundsätzliche Möglichkeiten, aggregierte Ergebnisse der Conjoint-Analyse zu gewinnen:

- Durchführung von *Individualanalysen* für jede Auskunftsperson und anschließende Aggregation der gewonnenen Teilnutzenwerte.
- Durchführung einer *gemeinsamen Conjoint-Analyse* für eine Mehrzahl von Auskunftspersonen, die aggregierte Teilnutzenwerte liefert.

Wird für jede Auskunftsperson eine *Individualanalyse* durchgeführt, so lassen sich anschließend die individuellen Teilnutzenwerte je Eigenschaftsausprägung durch *Mittelwertbildung* über die Personen aggregieren. Voraussetzung ist dabei, dass zuvor eine Normierung der Teilnutzenwerte für jede Person entsprechend Formel (9.9) vorgenommen wurde.

Individualanalyse und Aggregation

Eine *gemeinsame Conjoint-Analyse* über eine Mehrzahl von Auskunftspersonen lässt sich durchführen, indem die Auskunftspersonen als Wiederholungen (Replikationen) des Untersuchungsdesigns aufgefasst werden. Die in Abschnitt 9.2.4 vorgestellten Berechnungsformeln können dabei unverändert übernommen werden, wenn man die Bedeutung des Laufindex k, der zur Identifizierung der Stimuli diente, verändert. Betrachtet man anstelle der Stimuli jetzt Punkte (wie in Abbildung 9.14 und Abbildung 9.15 dargestellt), so vervielfacht sich bei einer Gesamtanalyse die Anzahl der Punkte entsprechend der Anzahl der Personen. Bei N Personen erhält man

Gemeinsame Conjoint-Analyse

$$K = N \cdot \text{Anzahl der Stimuli}$$

$$K = N \cdot \prod_{j=1}^{J} M_j \tag{9.11}$$

Punkte, wobei J wiederum die Anzahl der Eigenschaften und M_j die Anzahl der Ausprägungen von Eigenschaft j bezeichnet. Da die aggregierten Teilnutzenwerte die empirischen Rangdaten jeder einzelnen Person nicht mehr so gut reproduzieren können, wie es bei Individualanalysen der Fall ist, fällt der STRESS-Wert der Gesamtanalysen tendenziell höher aus.

Heterogenität

Jede Aggregation ist objektiv mit einem Verlust an Informationen verbunden. Es muss daher geprüft werden, ob die aggregierten Nutzenstrukturen nicht allzu *heterogen* sind, da ansonsten wesentliche Informationen durch die Aggregation verloren gehen würden. Bei starker Heterogenität lassen sich durch Anwendung einer *Clusteranalyse* (vgl. dazu Kapitel 8 in diesem Buch) homogene(re) Teilgruppen bilden.

Die Clusterung kann auf Basis der empirischen Rangdaten wie auch auf Basis der durch die Einzelanalysen gewonnenen *normierten* Teilnutzenwerte vorgenommen werden. Dabei ist jedoch zu beachten, dass bei der Durchführung einer Clusteranalyse als Proximitätsmaß immer ein *Ähnlichkeitsmaß* (Korrelationskoeffizient) verwendet werden sollte. Der Grund hierfür ist darin zu sehen, dass es bei der Conjoint-Analyse *nicht* darauf ankommt, Niveauunterschiede zwischen den Befragten aufzudecken, sondern die Entwicklung der Teilnutzenwerte in ihrer Relation zu betrachten. Das bedeutet, dass es bei einem Vergleich von Teilnutzenwerten zwischen verschiedenen Personen nicht auf deren *absolute* Höhe ankommt, sondern darauf, wie diese Personen die Eigenschaftsausprägungen in Relation gesehen haben; denn erst durch die relative Betrachtung lässt sich feststellen, ob zwei Personen einer bestimmten Eigenschaftsausprägung im Vergleich zu einer anderen (oder allen anderen) Ausprägung(en) einen höheren bzw. geringeren Nutzenbeitrag beimessen. Soll dennoch ein Distanzmaß als Proximitätsmaß verwendet werden, z. B. weil der Anwender das Ward-Verfahren zur Clusterung heranziehen möchte, so müssen die entsprechenden Ähnlichkeitsmaße in Distanzmaße transformiert werden.

9.3 Fallbeispiel

9.3.1 Problemstellung

Im Rahmen einer empirischen Erhebung wurden 40 Personen gebeten, insgesamt 11 Margarinebeschreibungen entsprechend ihrer individuellen Präferenzen in eine Rangordnung zu bringen. Den Margarinebeschreibungen lagen folgende vier Margarine-Eigenschaften zugrunde:

A: Preis
B: Verwendung
C: Geschmack
D: Kaloriengehalt

Dabei wurde unterstellt, dass diese Eigenschaften *voneinander unabhängig* sind und für die Kaufentscheidung als *relevant* angesehen werden können. Für die vier Eigenschaften wurde von den in Abbildung 9.16 dargestellten Eigenschaftsausprägungen ausgegangen.

Da für die Eigenschaften A und B die Zahl der Ausprägungen drei und für die Eigenschaften C und D nur zwei beträgt, liegt hier ein *asymmetrisches* (3 x 3 x 2 x 2) Design vor. Das Erhebungsdesign wird nach der Profilmethode erstellt. Bei einem

Eigenschaften	Eigenschaftsausprägungen
A: Preis	1: 1,50 €
	2: 1,00 €
	3: 0,50 €
B: Verwendung	1: als Brotaufstrich geeignet
	2: zum Kochen, Backen, Braten geeignet
	3: universell verwendbar
C: Geschmack	1: Buttergeschmack
	2: pflanzlich schmeckend
D: Kaloriengehalt	1: kalorienarm (400 kcal/100 g)
	2: normale Kalorien (700 kcal/100 g)

Abbildung 9.16: Eigenschaften und Eigenschaftsausprägungen in der Margarinestudie

vollständigen Design, d. h. bei Berücksichtigung aller möglichen Kombinationen der Eigenschaftsausprägungen würden wir (3 x 3 x 2 x 2 =) 36 fiktive Produkte (Stimuli) erhalten. Allerdings dürfte die Bewertung dieser 36 Alternativen eine Überforderung für die Auskunftspersonen bedeuten, sodass hier ein *reduziertes Design* gebildet wird.

Analyse mit Hilfe von SPSS

SPSS-Prozedur „Orthoplan“

Mit SPSS können durch die Prozedur ORTHOPLAN reduzierte Designs (Orthogonal arrays) erstellt werden. Die Prozedur ORTHOPLAN ist – im Gegensatz zur eigentlichen Conjoint-Analyse – in die Menüstruktur von SPSS integriert und kann wie in Abbildung 9.17 dargestellt aktiviert werden:

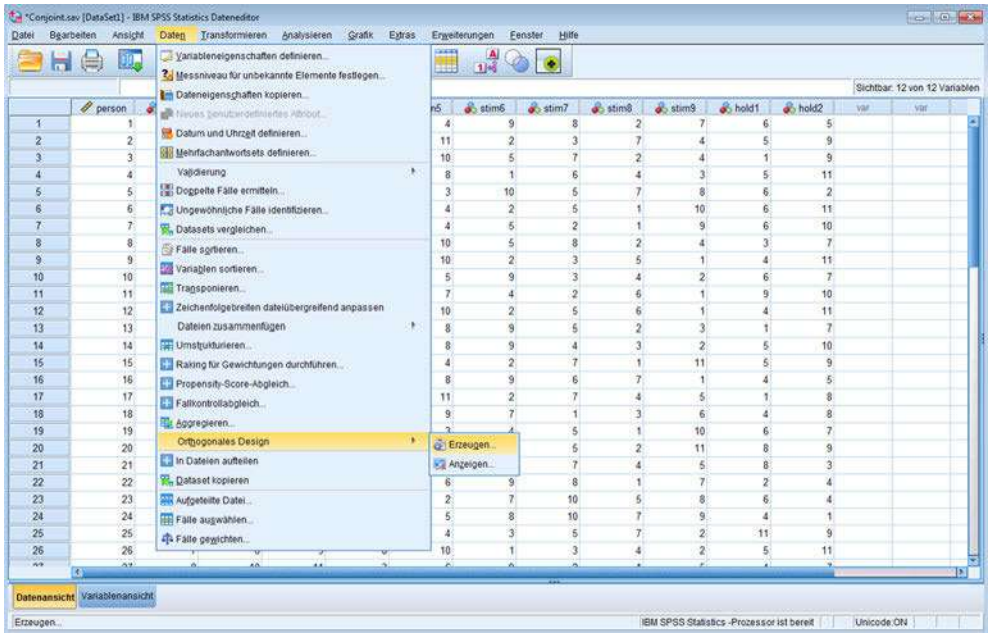


Abbildung 9.17: Aufruf der Prozedur ORTHOPLAN

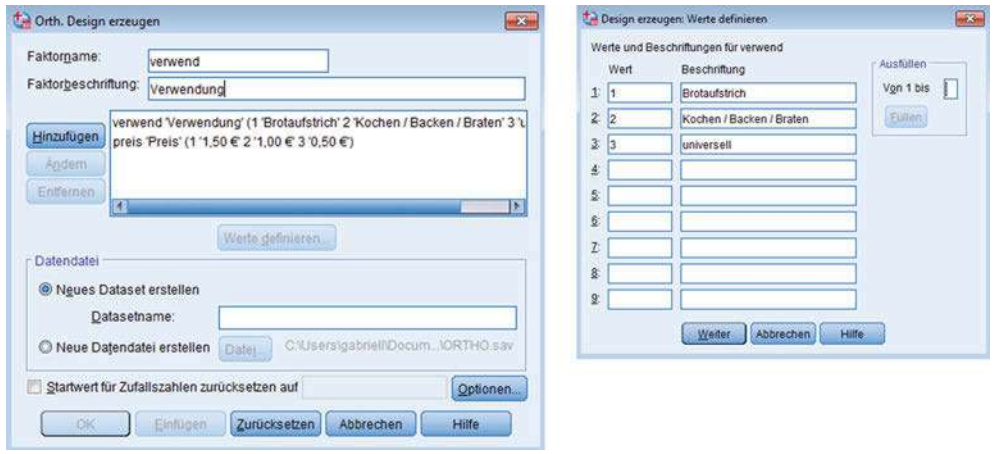


Abbildung 9.18: Erzeugung eines orthogonalen Designs mit SPSS

Orthogonales Design

Nach dem Aufruf erscheint das Fenster „Orthogonales Design erzeugen“. Hier muss für jeden Faktor (Eigenschaft) in der Analyse ein (maximal 8 Zeichen langer) Faktorname und ein dazugehöriges Faktorlabel eingegeben werden. Der Faktor ist jeweils durch anklicken der Taste „Hinzufügen“ zum Design der Conjoint-Analyse hinzuzufügen. In einem nächsten Schritt sind im Fenster „Design erzeugen: Werte definieren“ die Werte und Werte-Labels der jeweiligen Variable zu vergeben. Diese Prozedur ist für jede Variable in der Analyse zu wiederholen (vgl. Abbildung 9.17). Anschließend kann im Fenster „Orthogonales Design erzeugen“ die Taste OK angeklickt werden und durch SPSS wird ein passendes orthogonales Design erzeugt, welches im Daten-Editor angezeigt wird.

Das von ORTHOPLAN erzeugte reduzierte Design für die Margarinestudie ist nachfolgend in Abbildung 9.19 dargestellt.[17]

Preis	Verwend	Geschmac	Kalorien	STATUS_	CARD_
1,00	3,00	1,00	2,00	0	1
1,00	2,00	2,00	1,00	0	2
2,00	1,00	2,00	2,00	0	3
3,00	1,00	1,00	1,00	0	4
1,00	1,00	1,00	1,00	0	5
3,00	3,00	2,00	1,00	0	6
2,00	2,00	1,00	1,00	0	7
2,00	3,00	1,00	1,00	0	8
3,00	2,00	1,00	2,00	0	9
2,00	3,00	1,00	2,00	1	10
1,00	1,00	1,00	2,00	1	11
3,00	3,00	2,00	2,00	2	1
1,00	2,00	1,00	1,00	2	2

Number of cases read: 13 Number of cases listed: 13

Abbildung 9.19: Mit ORTHOPLAN erzeugtes reduziertes Design der Margarinestudie

[17] SPSS lässt auch reduzierte Designs zu, die durch den Anwender vorgegeben werden. In diesem Fall ist die Prozedur ORTHOPLAN überflüssig.

In Abbildung 9.19 sind in den ersten vier Spalten die jeweiligen Ausprägungen der vier Variablen aufgeführt. Darüber hinaus existieren zwei weitere Spalten, die mit „STATUS_“ und „CARD_“ überschrieben sind. Die Spalte „CARD_“ enthält dabei die *Nummerierung* der Karten.

In der Spalte „STATUS_“ sind ausschließlich die Ziffern 0, 1 und 2 vorhanden. Dabei werden die Stimuli, die dem *reduzierten Design* angehören, von SPSS mit einem STATUS_ von 0 versehen. In Abbildung 9.19 gehören mithin die ersten neun Stimuli zum reduzierten Design. Ein STATUS_ von 1 zeigt die sog. *Holdout-Karten* („holdout cards“) an. Holdout-Karten – oder Prüffälle – sind ebenfalls Stimuli, die den Auskunftspersonen zur Beurteilung vorgelegt werden. Sie werden allerdings *nicht* von SPSS zur Schätzung der Nutzenwerte verwendet, sondern zur Validitätsprüfung herangezogen. Sie werden mit den Stimuli des reduzierten Designs durchnummeriert (vgl. Spalte CARD_), um direkt erkennen zu können, wie viele Stimuli den Auskunftspersonen *insgesamt* zur Beurteilung vorgelegt werden müssen. In unserem Beispiel sind zwei Holdout-Karten vorhanden. Die vom Experimentator gewünschte Zahl an Prüffällen kann im Fenster „Orthogonales Design erzeugen“ im Unterpunkt „Optionen“ festgelegt werden. Diese Prüffälle bekommen in der Spalte CARD_ die Nummern 10 und 11. Insgesamt sind mithin elf Stimuli von den Befragten in eine Rangfolge zu bringen.

Holdout-Karte

Ein STATUS_ von 2 bedeutet, dass es sich um eine sog. *Simulations-Karte* („simulation card“) handelt. Diese werden den Auskunftspersonen *nicht* zur Bewertung vorgelegt (die Nummerierung beginnt wieder bei 1). SPSS errechnet mittels der auf Basis der Rangreihung der Stimuli geschätzten Teilnutzenwerte die Gesamtnutzenwerte der Simulations-Karten. Im vorliegenden Beispiel sind zwei Simulations-Karten vorhanden, die im Gegensatz zu den Stimuli des reduzierten Designs und den Holdout-Karten, vom Anwender selbst vorgegeben werden können. Bei der Wahl der Simulations-Karten ist es dem Anwender z. B. möglich, fiktive Produkte festzulegen, die für ihn von besonderem Interesse sind. Für diese Produkte werden dann ebenfalls Gesamtnutzenwerte berechnet sowie die Wahrscheinlichkeit ermittelt, dass ein Befragter einen durch die Simulations-Karte dargestellten Stimulus präferiert.

Simulations-Karte

Im nächsten Schritt kann den erstellten Stimuli, die bisher nur als Zahlenkombinationen zum Ausdruck kommen, die jeweils *inhaltliche Bedeutung* zugeordnet werden. Durch die Prozedur PLANCARDS bietet SPSS die Möglichkeit, sog. Produktkarten zu erzeugen. Beispielsweise bedeutet Stimulus 1 mit der Zahlenkombination (1,3,1,2), dass es sich um eine (fiktive) Margarine mit folgenden Eigenschaftsausprägungen handelt:

SPSS-Prozedur „Plancards“

Preis	1,50€
Verwendung	universell
Geschmack	Buttergeschmack
Kaloriengehalt	normale Kalorien

Abbildung 9.20 zeigt den entsprechenden Computer-Ausdruck, wobei die Karten 1 bis 9 den Stimuli des reduzierten Designs entsprechen und die Karten 10 und 11 die Holdout-Karten repräsentieren.

Margarine PLANCARDS

	Karten-ID	Preis	Verwendung	Geschmack	Kaloriengehalt
1	1	1,50€	universell	Buttergeschmack	normale Kalorien
2	2	1,50€	Kochen/Backen/ Braten	pflanzlich schmeckend	kalorienarm
3	3	1,00€	Brotaufstrich	pflanzlich schmeckend	normale Kalorien
4	4	0,50€	Brotaufstrich	Buttergeschmack	kalorienarm
5	5	1,50€	Brotaufstrich	Buttergeschmack	kalorienarm
6	6	0,50€	universell	pflanzlich schmeckend	kalorienarm
7	7	1,00€	Kochen/Backen/ Braten	Buttergeschmack	kalorienarm
8	8	1,00€	universell	Buttergeschmack	kalorienarm
9	9	0,50€	Kochen/Backen/ Braten	Buttergeschmack	normale Kalorien
10[a]	10	1,00€	universell	Buttergeschmack	normale Kalorien
11[a]	11	1,50€	Brotaufstrich	Buttergeschmack	normale Kalorien
12[b]	1	0,50€	universell	pflanzlich schmeckend	normale Kalorien
13[b]	2	1,50€	Kochen/Backen/ Braten	Buttergeschmack	kalorienarm

a. Prüfung

b. Simulation

Abbildung 9.20: Durch PLANCARDS erzeugte Produktkarten der Margarinestudie

Die Produktkarten aus Abbildung 9.20 können nun zur Befragung verwendet werden. Die Präferenzeinschätzung durch die Befragten kann dabei über verschiedene Wege erfolgen:

Methode der Rangverteilung

- Bei der *Methode der Rangverteilung* werden die Befragten gebeten, jede Produktkarte mit einem Rangwert zu versehen, wobei die Rangwerte die Produktpräferenzen der Befragten widerspiegeln. Je kleiner der Rangwert, desto größer ist die Präferenz des Befragten für die jeweilige Produktkarte.

Präferenzwertmethode

- Bei der *Präferenzwertmethode* wird jede einzelne Produktkarte z. B. mit Hilfe einer Rating-Skala durch einen (metrischen) Präferenzwert beurteilt. Je größer der Präferenzwert, desto größer ist auch die Präferenz des Befragten für diese Produktkarte.

Methode des Rangordnens

- Bei der *Methode des Rangordnens* müssen die Befragten die Produktkarten nach ihrer Präferenz sortieren, und eine Beurteilung in Form von Rang- oder Präferenzwerten wird nicht vorgenommen.

Im Rahmen der Margarinestudie wurden 40 Personen befragt und gebeten, entsprechend der *Methode der Rangverteilung*, den jeweiligen Produktkarten Rangwerte von 1 bis 11 zuzuordnen. Nach der „Eignung für den persönlichen Bedarf“ sollten die elf Produktkarten mit Rang 1, für die „am stärksten präferierte Produktalternative“, bis Rang 11, für die „am wenigsten präferierte Produktalternative“, versehen werden.

Die Befragungsergebnisse werden pro Person im Dateneditor von SPSS eingegeben. Dabei gehen die im Fallbeispiel definierten 11 Stimuli als Variable (stim1 bis stim9; hold1, hold2) ein, wobei die von den Befragten vergebenen Ränge (1 bis 11) die Variablenwerte darstellen.

Erstellung der Syntax-Datei und Start der Prozedur CONJOINT

Die Prozedur CONJOINT ist noch *nicht* in die graphische Menüstruktur von SPSS eingebunden, sodass zunächst eine Syntaxdatei erstellt werden muss. Eine neue Syntaxdatei wird über das Hauptmenü „Datei“ und den Unterpunkt „Neu“ →„Syntax“ geöffnet. Für das Fallbeispiel ist die Befehlsstruktur der Prozedur CONJOINT in Kapitel 9.3.3 ausführlich beschrieben (vgl. Abbildung 9.31), wobei in diesem Kapitel auch die Prozeduren ORTHOPLAN (Abbildung 9.29) und PLANCARDS (Abbildung 9.30) erläutert werden. An dieser Stelle sei nur auf den Unterbefehl „FACTORS“ eingegangen, durch den der Zusammenhang zwischen den jeweiligen Ausprägungen der Stimuli-Merkmale bestimmt wird. Die Spezifikationen zum Unterbefehl FACTORS erfolgen in Abhängigkeit der vorgenommenen Kodierung der Stimuli-Eigenschaften (vgl. Abbildung 9.16). Für das Fallbeispiel wird unterstellt, dass das Merkmal „Preis“ linear ansteigend (Anweisung: LINEAR MORE in Abbildung 9.31) kodiert ist. Das bedeutet, dass die drei Preisausprägungen (1=1,50€; 2=1,00€; 3=0,50€) mit einem *steigenden* Nutzen einhergehen bzw. mit steigender Ausprägungsnummer auch eine steigende Präferenz vermutet wird. Die Margarine-Eigenschaft „Kalorien“ hingegen wurde mit LINEAR LESS kodiert und damit *unterstellt*, dass mit steigender Merkmalsausprägung (im Fallbeispiel: 1=kalorienarm; 2=normale Kalorien) der Nutzen für den Anwender sinkt bzw. mit steigender Ausprägungsnummer eine fallende Präferenz einhergeht. Für die Margarine-Eigenschaften „Verwendung“ und „Geschmack“ werden keine sachlogischen Annahmen getroffen, und sie wurden durch die Angabe von „DISCRETE“ (vgl. Abbildung 9.31) als kategoriale Eigenschaften angenommen. Nach neu erstellter bzw. Aufruf einer hinterlegten Syntaxdatei kann die Prozedur CONJOINT direkt über den Hauptmenüpunkt „Ausführen“ des SPSS Syntaxeditors gestartet werden. Für das Fallbeispiel sind die Syntaxdateien ausführlich in Abschnitt 9.3.3 beschrieben und auch über den Support zum Buch erhältlich.

Beziehung zwischen Eigenschaftsausprägungen und Präferenzdaten

9.3.2 Ergebnisse

9.3.2.1 Individuelle Ergebnisse

Die Conjoint-Analyse stellt eine Individualanalyse dar. Entsprechend werden für jede befragte Person die erhobenen Rangwerte für die neun fiktiven Produkte des reduzierten Designs isoliert ausgewertet. Beispielhaft seien im Folgenden die individuellen Ergebnisse von Person 33 betrachtet, die in Abbildung 9.21 zugefasst dargestellt sind.

Unter der Überschrift „Nutzen“ werden die Nutzenschätzungen (Teilnutzenwerte) für jede Eigenschafts-Ausprägung mit ihren jeweiligen *Standardfehlern* ausgegeben. Für die Eigenschaft „Verwendung“ wurden für Person 33 folgende Teilnutzenwerte geschätzt:

SPSS liefert hierzu die einzelnen Ergebnistabellen separat aus, die der besseren Übersicht halber hier jedoch zusammengefasst dargestellt sind. Die *geschätzten Teilnutzenwerte* für jede Eigenschaftsausprägung werden mit ihren jeweiligen *Standardfehlern* ausgegeben. Betrachtet man die geschätzten Teilnutzenwerte, so betragen diese beispielsweise für die Eigenschaft „Verwendung“:

-1,667	(Ausprägung: als Brotaufstrich geeignet)
0,667	(Ausprägung: zum Kochen, Backen, Braten geeignet)
1,000	(Ausprägung: universell verwendbar)

Der Standardfehler beträgt bei allen drei Eigenschaftsausprägungen 0,598. Er liefert einen ersten Anhaltspunkt für die Güte der Conjoint-Ergebnisse. Je geringer die

Nutzen

		Nutzen-schätzung	Std.-Fehler
verwend	Brotaufstrich	-1,667	,598
	Kochen/Backen/Braten	,667	,598
	universell	1,000	,598
geschmac	Buttergeschmack	2,250	,449
	pflanzlich schmeckend	-2,250	,449
preis	1,50€	,500	,518
	1,00€	1,000	1,036
	0,50€	1,500	1,555
kalorien	kalorienarm	-,500	,898
	normale Kalorien	-1,000	1,795
(Konstante)		3,917	1,645

Wichtigkeitswerte

verwend	30,769
geschmac	51,923
preis	11,538
kalorien	5,769

Korrelationen[a]

	Wert	Sig.
Pearson-r	,959	,000
Kendall-Tau	,986	,000
Kendall-Tau für Prüfkarten	1,000	.

a. Korrelationen zwischen beobachteten und geschätzten Bevorzugungen

Koeffizienten

	B-Koeffizient	
	Schätzwert	Std.-Fehler
preis	,500	,518
kalorien	-,500	,898

Bevorzugungswerte für Simulationen

Kartennummer	ID	Wert
1	1	3,167
2	2	6,833

Abbildung 9.21: Ergebnisse der individuellen Conjoint-Analyse für Person 33

Standardfehler, desto eher lässt sich die empirische Rangfolge durch die ermittelten Rangwerte abbilden. Entsprechend sind die übrigen Werte dieser Spalte zu interpretieren.

Hingewiesen sei an dieser Stelle auf das Ergebnis, dass die Teilnutzenwerte für die Ausprägungen der Eigenschaft „Preis“ positive Werte aufweisen, d. h. der Preis wirkt nutzensteigernd, wenngleich eine höhere Preisstufe mit einem geringeren positiven Teilnutzen einhergeht. Eine Analyse der Ergebnisse für alle befragten Personen zeigt, dass sich für fast alle Befragten positive Teilnutzenwerte bei der Eigenschaft „Preis“ ergeben. Daraus kann geschlossen werden, dass dem Preis im vorliegenden Anwendungsfall die Rolle eines Qualitätsindikators zukommt.

Gesamtnutzenwerte

Die Teilnutzenwerte ermöglichen die Berechnung von *metrischen Gesamtnutzenwerten* für beliebig konstruierbare Produkte, wobei sich die Gesamtnutzenwerte für

unser Beispiel nach Maßgabe von Formel (9.12) berechnen:

$$G_k = \mu + \beta_{Am} + \beta_{Bm} + \beta_{Cm} + \beta_{Dm} \tag{9.12}$$

mit

G_k = Gesamtnutzenwert für Stimulus k
μ = konstanter Term der Nutzenschätzung
β_{Am} = Teilnutzenwert für die Ausprägung m der Eigenschaft A
β_{Bm} = Teilnutzenwert für die Ausprägung m der Eigenschaft B
β_{Cm} = Teilnutzenwert für die Ausprägung m der Eigenschaft C
β_{Dm} = Teilnutzenwert für die Ausprägung m der Eigenschaft D

Die Konstante μ kann dabei als Basisnutzen interpretiert werden, von dem sich die übrigen Eigenschaftsausprägungen positiv oder negativ abheben. Beispielhaft für Stimulus 1 (Margarine 1 in Abbildung 9.20), bei dem es sich um eine Margarine mit universeller Verwendungsmöglichkeit, Buttergeschmack, einem Preis von 1,50 € und normalem Kaloriengehalt handelt, lässt sich der Gesamtnutzenwert wie folgt berechnen:

$$G_1 = 3,917 + 1,000 + 2,250 + 0,500 + (-1,000) = 6,667$$

Entsprechend können die Gesamtnutzenwerte für die neun Stimuli des reduzierten Designs und für die beiden Holdout-Karten (Stimulus 10 und 11) berechnet werden (vgl. Abbildung 9.22).

Stimulus	Gesamtnutzenwert	resultierender Rang	tatsächlicher Rang
1	6,67	5	5
2	2,33	10	10
3	0,00	11	11
4	5,50	6	6
5	4,50	7	8
6	3,67	9	9
7	7,33	2a	2
8	7,67	1	1
9	7,33	2b	3
10	7,17	4	4
11	4,00	8	7

Abbildung 9.22: Gesamtnutzenwert, Rang und tatsächlicher Rang der Auskunftsperson 33

Aus Abbildung 9.22 wird deutlich, dass die tatsächlichen Rangwerte (Spalte 4) der Auskunftsperson 33 sehr gut durch die aus den metrischen Gesamtnutzenwerten resultierenden Rangwerte (Spalte 3) reproduziert werden. Lediglich die Rangfolge von Stimulus 5 und Stimulus 11 (Holdout-Karte) sind nicht korrekt abgebildet. Ein Maß für die Güte der Abbildung der empirischen Rangdaten auf die aus den Gesamtnutzenwerten resultierenden Ränge liefern die in Abbildung 9.21 ausgegebenen Korrelationskoeffizienten. Während der *Pearson'sche Korrelationskoeffizient* die Korrelationen zwischen den metrischen Gesamtnutzenwerten und den tatsächlichen (empirischen)

Korrelationskoeffizient von Pearson

Rängen berechnet, misst *Kendall's Tau* die Korrelation zwischen tatsächlichen und aus den Conjoint-Ergebnissen resultierenden Rängen. Je mehr sich die Korrelationskoeffizienten absolut dem Wert 1 nähern, desto besser können die empirischen Daten durch die Conjoint-Ergebnisse abgebildet werden. Allerdings ist zu beachten, dass im Falle von Pearson's R die empirischen Rangdaten als metrisch skaliert unterstellt werden müssen, was nur dann der Fall ist, wenn bei der Befragung die Präferenzwertmethode zur Anwendung kam. Darüber hinaus werden Pearson's R und Kendall's Tau auch für die Holdout-Karten berechnet und beziehen sich in diesem Fall auf die tatsächliche und geschätzte Rangfolge dieser Karten. Da Holdout-Karten bei der Schätzung der Teilnutzenwerte nicht berücksichtigt, real aber abgefragt wurden, stellen die auf die Holdout-Karten bezogenen Korrelationskoeffizienten ein Maß für die Validität der Ergebnisse dar.

Kendall's Tau

Mit Hilfe der Teilnutzenwerte aus Abbildung 9.21 lassen sich für Person 33 nun auch die Gesamtnutzenwerte für das *vollständige Design* berechnen, obwohl in der *Befragung* nur ein *reduziertes Design* erhoben wurde. Abbildung 9.23 zeigt unter der Überschrift „Gesamtnutzenwerte" die einzelnen Gesamtnutzenwerte auf, die in SPSS jedoch nicht angefordert werden können und deshalb durch den Anwender errechnet werden müssen. Mit Hilfe der „Stimuli-Anordnungen" lassen sich die Positionen der einzelnen Gesamtnutzenwerte identifizieren. So entspricht z. B. der fett gedruckte Gesamtnutzenwert von 8,1667 dem Stimulus P3311, wobei die Ziffernreihenfolge hinter dem P der Eigenschaftsreihenfolge „Preis", „Verwendung", „Geschmack", „Kaloriengehalt" entspricht und die Ziffern selbst die jeweilige Eigenschaftsausprägungen entsprechend Abbildung 9.16 angeben.

Stimulus-Anordnung (Gesamtnutzenwert)

P1111(4,5000)	P1121(0,0000)	P1211(6,8334)	P1221(2,3334)	P1311(7,1667)	P1321(2,6667)
P1112(4,0000)	P1122(-0,5000)	P1212(6,3334)	P1222(1,8334)	P1312(6,6667)	P1322(2,1667)
P2111(5,0000)	P2121(0,5000)	P2211(7,3334)	P2221(2,8334)	P2311(7,6667)	P2321(3,1667)
P2112(4,5000)	P2122(0,0000)	P2212(6,8334)	P2222(2,3334)	P2312(7,1667)	P2322(2,6667)
P3111(5,5000)	P3121(1,0000)	P3211(7,8334)	P3221(3,3334)	**P3311(8,1667)**	P3321(3,6667)
P3112(5,0000)	P3122(0,5000)	P3212(7,3334)	P3222(2,8334)	P3312(7,6667)	P3322(3,1667)

Abbildung 9.23: Gesamtnutzenwert des vollständigen Designs für Auskunftsperson 33

Die in Abbildung 9.23 unterstrichenen Werte kennzeichnen die Gesamtnutzenwerte der neun *Produktalternativen* im reduzierten Design. Allerdings wird deutlich, dass die am stärksten präferierte Produktalternative (vgl. den fett gesetzten Wert) in der Befragung selbst nicht erhoben wurde. Damit ist die Conjoint-Analyse in der Lage, *Gesamtnutzenwerte für alle Produktalternativen* zu ermitteln, auch wenn der Befragung nur ein reduziertes Design zugrunde lag.

Die bisherigen Ausführungen bezogen sich jeweils auf den Nutzenbeitrag einzelner Eigenschaftsausprägungen. Der Tabelle „Wichtigkeitswerte" in Abbildung 9.21 lässt sich aber darüber hinaus noch entnehmen, welche Bedeutung den *einzelnen Eigenschaften* bei der Präferenzbildung von Person 33 zukommt. Diese Prozentwerte spiegeln die *relativen Wichtigkeiten* der einzelnen Eigenschaften wider. An dieser Stelle sei nochmals daran erinnert, dass sich die relative Wichtigkeit einer Eigenschaft auf die Wichtigkeit zur Präferenzveränderung bezieht, die sich *nicht* aus den absoluten Werten der Teilnutzenwerte ableiten lässt. Für die relative Wichtigkeit ist die Spannweite der Teilnutzenwerte je Eigenschaft entscheidend (vgl. Abschnitt 9.2.5). Zur Verdeutlichung ist in Abbildung 9.24 die Berechnung der relativen Wichtigkeiten der Eigenschaften für Person 33 gem. Formel (9.10) aufgezeigt:

Relative Wichtigkeiten

Eigenschaft	Spannweite	relative Wichtigkeit
Verwendung	$1,0000 - (-1,6667) = 2,67$	$2,67/8,67 = \mathbf{0,3074}$
Geschmack	$2,2500 - (-2,2500) = 4,50$	$4,5/8,67 = \mathbf{0,5190}$
Preis	$1,5000 - (0,5000) = 1,00$	$1,00/8,67 = \mathbf{0,1153}$
Kaloriengehalt	$-0,5000 - (-1,0000) = 0,50$	$0,50/8,67 = \mathbf{0,0576}$
	Summe: $8,67$	Summe: $1,00$

Abbildung 9.24: Berechnung der relativen Wichtigkeiten je Eigenschaft

Die in Abbildung 9.24 fett hervorgehobenen Anteilswerte entsprechen den in Abbildung 9.21 abgedruckten Prozentwerten in der Tabelle „Wichtigkeitswerte“. Es wird deutlich, dass der Geschmack der Margarine die Gesamtpräferenz der Auskunftsperson 33 am stärksten beeinflusst (51,90 %). Danach folgen Verwendung und Preis. Der Eigenschaft Kaloriengehalt kommt mit 5,77 % die geringste Bedeutung zur Präferenzveränderung zu.

Bei der Spezifikation des Conjoint-Modells hatten wir für die Eigenschaften „Preis“ und „Kaloriengehalt“ bestimmte Beziehungszusammenhänge zwischen Eigenschaftsausprägungen und den empirischen Rangdaten unterstellt. Dabei sind wir davon ausgegangen, dass mit geringer werdendem Preis und sinkendem Kaloriengehalt der Nutzen steigen wird. Aufgrund dieser Vorgaben können für diese beiden Eigenschaften Regressionskoeffizienten berechnet werden, die in Abbildung 9.21 in der Tabelle „Koeffizienten“ ausgewiesen sind. Mit Hilfe der Regressionskoeffizienten B lassen sich die Teilnutzenwerte der Ausprägungen dieser Variablen berechnen. Diese ergeben sich dabei durch das Produkt aus der Nummer der Eigenschaftsausprägung (also 1 für die erste Ausprägung, 2 für die zweite Ausprägung usw.) und dem Regressionskoeffizienten. Für die Eigenschaft „Kaloriengehalt“ lässt sich die Höhe der Teilnutzenwerte beispielsweise wie folgt berechnen:

Erster Teilnutzenwert	$(B = -0,5)$:	$1 \cdot (-0,5) = -0,5$
Zweiter Teilnutzenwert	$(B = -1)$:	$2 \cdot (-0,5) = -1,0$

Wird ein vermuteter Zusammenhang *nicht* bestätigt, werden also beispielsweise bei einer Auskunftsperson für geringe Preise auch geringere Teilnutzenwerte ermittelt als für hohe Preise, so wird eine *Verletzung der getroffenen Annahme* als *Reversal* oder Umkehrung bezeichnet.

Umkehrungen (Reversals)

Liegen Umkehrungen vor, so werden diese in der Tabelle „Wichtigkeitswerte“ des Ergebnisoutputs durch einen Index a bei der Überschrift und einen Index b bei der betroffenen Variablen gekennzeichnet. Bei Person 33 sind keine Umkehrungen aufgetreten. Schließlich werden noch für jede Person die Gesamtnutzenwerte der Simulationskarten, wenn solche spezifiziert wurden, ausgegeben. Im Fallbeispiel wurden im Rahmen der Modellspezifizierung (Daten des Orthogonalen Designs) zwei Simulationskarten mit folgenden Ausprägungen angegeben:

Simulations-Karten

Card_1:	Preis: 3=0,50; verwend: 3=universell; geschmac: 2=pflanzlich schmeckend; kalorien: 2=normale Kalorien
Card_2:	Preis: 1=1,50; verwend: 2=Kochen/Backen/Braten; geschmac: 1=Buttergeschmack; kalorien: 1=kalorienarm

Für Person 33 sind die Gesamtnutzenwerte der beiden Simulationskarten unter der Tabellenüberschrift „Bevorzugungswerte für Simulationen“ ausgegeben. Für Simulationskarte 1 beträgt der Gesamtnutzenwert 3,167 und Simulationskarte 2 erreicht bei Person 33 einen Gesamtnutzerwert von 6,833 (vgl. Abbildung 9.21). Im Ergebnis-Output zur Conjoint-Analyse werden die hier für Person 33 besprochenen Ergebnisse für alle befragten Personen individuell ausgegeben.

9.3.2.2 Aggregierte Ergebnisse

Für die Neuprodukteinführung einer Margarinemarke sind die individuellen Auswertungen im Vergleich zu einer aggregierten Auswertung nur von untergeordnetem Interesse. In vielen Fällen möchte der Anbieter einer Margarine vor allem wissen, ob es *Gruppen von potenziellen Nachfragern* gibt, die in Bezug auf die *Teilnutzenbewertungen* ähnliche *Präferenzen* besitzen und welche *Produkteigenschaften* insgesamt als besonders *präferenzrelevant* eingestuft werden müssen. Zu diesem Zweck ist es notwendig, eine *Aggregation der individuellen* Daten vorzunehmen. Dies kann auf zwei Arten erfolgen:

Aggregation individueller Daten

- Aggregation der Individualanalysen
- Durchführung einer gemeinsamen Conjoint-Analyse

9.3.2.2.1 Aggregation der Individualanalysen

Eine Aggregation der Individualanalysen ist nur möglich, wenn zuvor eine *Normierung der ermittelten Teilnutzenwerte* vorgenommen wird. Zu diesem Zweck greifen wir auf die Normierungsvorschrift aus Abschnitt (9.2.4) zurück. Mit Hilfe von Formel (9.9) lassen sich aus den Teilnutzenwerten der Individualanalysen normierte Teilnutzenwerte errechnen, die eine Vergleichbarkeit der einzelnen Individualanalysen ermöglichen. Normierte Teilnutzenwerte werden durch SPSS nicht automatisch bereitgestellt und müssen mit Hilfe von COMPUTE-Befehlen errechnet werden. Mit Hilfe der SPSS-Prozedur DESCRIPTIVES lassen sich dann durchschnittliche normierte Teilnutzenwerte über alle Befragten ermitteln. Abbildung 9.25 zeigt die entsprechenden Ergebnisse für die Margarinestudie, wobei die relativen Gewichte der Eigenschaften gem. Formel (9.10) berechnet wurden.

Die Durchschnittswerte der normierten Teilnutzenwerte in Abbildung 9.25 sind analog zu den individuellen Teilnutzenwerten der Auskunftspersonen zu interpretieren. Es wird deutlich, dass die Befragten im Durchschnitt eine kalorienarme, nach Butter schmeckende und universell verwendbare Margarine zu einem Preis von 0,50 € präferieren. Allerdings ist zu beachten, dass im vorliegenden Beispiel unterschiedlich große *Streuungsbreiten* der Teilnutzenwerte auftreten. Die Streuungen (Standardabweichungen in Abbildung 9.25) sind dafür verantwortlich, dass trotz der Betrachtung normierter Teilnutzenwerte der Gesamtnutzen der am meisten präferierten Margarine nicht mehr genau 1 beträgt, sondern in unserem Fall nur noch (0,2756 + 0,2099 + 0,2263 + 0,1436 =) 0,86. Bei der Aggregation der Individualanalysen muss sich der Anwender deshalb bewusst sein, dass ihm bei der Errechnung von Gesamtnutzenwerten für die fiktiven Produkte die Informationen über die Streuungen verloren gehen.

Ein solcher Informationsverlust wird vermieden, wenn statt der Mittelwertbildung auf der Basis der normierten Teilnutzenwerte eine Clusteranalyse (vgl. Kapitel 8) durchgeführt wird, die Gruppen von Personen mit ähnlichen Teilnutzenprofilen ermittelt. Dabei ist allerdings zu beachten, dass als Proximitätsmaß ein *Ähnlichkeitsmaß*

	Mittelwert	Standardabweichung
Verwendung (Gewicht: 29,26 %)		
als Brotaufstrich	0,1586	0,1867
Kochen, Backen, Braten	0,1005	0,1479
universell verwendbar	0,2099	0,1887
Geschmack (Gewicht: 25,58 %)		
Buttergeschmack	0,2263	0,1990
pflanzlich schmeckend	0,0367	0,1218
Preis (Gewicht: 28,16 %)		
1,50 €	0,0131	0,0516
1,00 €	0,1443	0,1009
0,50 €	0,2756	0,2127
Kaloriengehalt (Gewicht: 16,99 %)		
kalorienarm	0,1436	0,1666
normale Kalorien	0,0332	0,0880

Abbildung 9.25: Durchschnittlich normierte Teilnutzenwerte in der Margarinestudie

(z. B. der Korrelationskoeffizient) zugrunde gelegt wird (vgl. Abschnitt 9.2.5). Im Gegensatz zur Mittelwertbildung liefert die Clusteranalyse jedoch keinen Repräsentativwert für alle Personen. Es kann aber davon ausgegangen werden, dass die Durchschnittswerte der normierten Teilnutzenwerte je Cluster eine geringere Streuung als in der Erhebungsgesamtheit aufweisen.

9.3.2.2.2 Gemeinsame Conjoint-Analyse

Gemeinsame Conjoint-Analyse

Bei der Durchführung einer gemeinsamen Conjoint-Analyse werden die Befragten als Replikationen in die Analyse einbezogen, wodurch alle Befragungswerte der Auskunftspersonen *gleichzeitig* zur Schätzung der Teilnutzenwerte herangezogen werden (vgl. Abschnitt 9.2.5). Dadurch bleiben die in den Streuungen enthaltenen Informationen erhalten, wodurch ein geringerer Informationsverlust als bei der Durchschnittsbildung entsteht. Durch SPSS wird am Ende des Ergebnis-Outputs auch eine *Gesamtstatistik* erstellt, die die Ergebnisse der gemeinsamen Conjoint-Analyse enthält (vgl. Abbildung 9.26).

Die Ergebnisse der gemeinsamen Conjoint-Analyse können analog zu den Ausführungen in Abschnitt 9.3.2.1 interpretiert werden. Werden die Ergebnisse der aggregierten Analyse (Abbildung 9.25) mit denen der gemeinsamen Conjoint-Analyse (Abbildung 9.26) verglichen, so wird deutlich, dass die ermittelten Teilnutzenwerte zwar stark unterschiedlich ausgeprägt sind, jedoch die relativen Wichtigkeiten der einzelnen Eigenschaften nicht differieren. Auch nach der gemeinsamen Conjoint-Analyse ist das fiktive Produkt mit dem höchsten Gesamtnutzen eine kalorienarme, nach Butter schmeckende, universell verwendbare Margarine, die zu einem Preis von 0,50 € erworben werden kann. Zusätzlich zu den aggregierten Ergebnissen der Individualanalysen weist die Gesamtstatistik auch die Bevorzugungswahrscheinlichkeiten für die Simulationskarten aus (vgl. Abbildung 9.27).

Die Bevorzugungswahrscheinlichkeiten geben die Wahrscheinlichkeit dafür an, dass eine Simulationskarte von den Befragten mit der höchsten Präferenz versehen wird und dann auch zum Kauf ausgewählt wird. Dabei werden Wahrscheinlichkeiten für die

Nutzen

		Nutzen-schätzung	Std.-Fehler
verwend	Brotaufstrich	,017	,213
	Kochen/Backen/Braten	-,467	,213
	universell	,450	,213
geschmac	Buttergeschmack	,825	,160
	pflanzlich schmeckend	-,825	,160
preis	1,50€	1,179	,185
	1,00€	2,358	,369
	0,50€	3,537	,554
kalorien	kalorienarm	-1,075	,320
	normale Kalorien	-2,150	,639
(Konstante)		3,800	,586

Wichtigkeitswerte

verwend	29,265
geschmac	25,582
preis	28,164
kalorien	16,990

Durchschnittlicher Wichtigkeitswert

Korrelationen[a]

	Wert	Sig.
Pearson-r	,983	,000
Kendall-Tau	1,000	,000
Kendall-Tau für Prüfkarten	1,000	.

a. Korrelationen zwischen beobachteten und geschätzten Bevorzugungen

Koeffizienten

	B-Koeffizient Schätzwert
preis	1,179
kalorien	-1,075

Bevorzugungswerte für Simulationen

Kartennummer	ID	Wert
1	1	4,812
2	2	4,263

Abbildung 9.26: Gesamtstatistik mit den Ergebnissen der gemeinsamen Conjoint-Analyse

Simulationskarten nach drei verschiedenen Auswahlmodellen (probability-of-choice-models) berechnet:

Auswahlmodell „Maximaler Nutzen“

- Das *Max Utility-Modell* weist pro Person der Simulations-Karte mit dem höchsten Gesamtnutzen eine Wahlwahrscheinlichkeit von 1 zu, während alle anderen Simulations-Karten eine Wahlwahrscheinlichkeit von 0 erhalten. In der Gesamtstatistik wird unter der Überschrift „Bevorzugungswahrscheinlichkeiten für Simulationen“ in der Spalte „Maximaler Nutzen“ der Durchschnittswert dieser Wahrscheinlichkeiten über alle Personen ausgewiesen. Falls der höchste Gesamt-

Bevorzugungswahrscheinlichkeiten für Simulationen[b]

Kartennummer	ID	Maximaler Nutzen[a]	Bradley-Terry-Luce	Logit
1	1	48,8%	50,7%	51,4%
2	2	51,3%	49,3%	48,6%

a.Einschließlich gebundener Simulationen

b.Y aus X Personen werden in der Bradley-Terry-Luce- und der Logit-Methode verwendet, da diese Personen nur nicht-negative Werte aufweisen.

Abbildung 9.27: Präferenzwahrscheinlichkeiten für Simulationskarten

nutzenwert für mehrere Simulations-Karten identisch ist, wird die Wahrscheinlichkeit von 1 auf die entsprechenden Simulations-Karten gleich verteilt.

- Das *BTL-Modell* geht auf die Überlegungen von Bradley, Terry und Luce zurück und errechnet pro Person die Wahlwahrscheinlichkeit für eine bestimmte Simulations-Karte, indem es den Gesamtnutzenwert dieser Simulations-Karte durch die Summe der Gesamtnutzenwerte aller Simulations-Karten dividiert. In der Gesamtstatistik wird unter der Überschrift „Bevorzugungswahrscheinlichkeiten für Simulationen“ in der Spalte „Bradley-Terry-Luce“ der Durchschnittswert dieser Wahrscheinlichkeiten über alle Personen ausgewiesen. Besitzt eine Simulations-Karte für eine bestimmte Person einen negativen oder Null-Gesamtnutzenwert, so wird für diese Person keine BTL-Wahlwahrscheinlichkeit berechnet.

Auswahlmodell „Bradley-Terry-Luce“

- Das *Logit-Modell* verfährt analog zum BTL-Modell, wobei jedoch nicht die absoluten Gesamtnutzenwerte betrachtet werden, sondern für jede Simulations-Karte die Gesamtnutzenwerte unter Verwendung der e-Funktion bestimmt werden. Die Wahlwahrscheinlichkeit für eine bestimmte Simulations-Karte errechnet sich damit wie folgt:

Auswahlmodell „Logit“

$$P_{Si} = \frac{e^{G_i}}{\sum_{i=1}^{I} e^{G_i}} \tag{9.13}$$

mit

P_{Si} = Wahlwahrscheinlichkeit für Simulations-Karte i
G_i = Gesamtnutzenwert der Simulations-Karte i
e = Euler'sche Zahl (e = 2,71828...)

In der Gesamtstatistik wird unter der Überschrift „Bevorzugungswahrscheinlichkeiten für Simulationen“ in der Spalte „Logit“ der Durchschnittswert dieser Wahrscheinlichkeiten über alle Personen ausgewiesen. Besitzt eine Simulations-Karte für eine bestimmte Person einen negativen oder Null-Gesamtnutzenwert, so wird für diese Person keine Logit-Wahlwahrscheinlichkeit berechnet.

Abbildung 9.27 macht deutlich, dass für unser Fallbeispiel alle drei Wahrscheinlichkeits-Modelle zu nahezu identischen Ergebnissen führen. Im vorliegenden Fall muss der Anwender davon ausgehen, dass die Wahlwahrscheinlichkeit

für beide in den Simulations-Karten vorgegebenen Margarinesorten im Durchschnitt bei nur 50% liegt. Damit ist keine eindeutige Präferenz der Befragten für eine der Simulations-Karten erkennbar.

Gesamtstatistik zu Umkehrungen (Reversals)

Schließlich wird am Ende der Gesamtstatistik eine Übersicht der vorhandenen Reversals unter der Überschrift „Anzahl der Umkehrungen“ ausgegeben. Dabei wird neben einem Ausweis der Gesamtzahl an Umkehrungen je Eigenschaft (FAKTOR) auch fallweise die Zahl an Umkehrungen bei jeder befragten Person ausgewiesen. Hieraus lassen sich Konzentrationen von Reversals oder Umkehrungen auf bestimmte Personen erkennen.

9.3.3 SPSS-Kommandos

Bei der Durchführung einer Conjoint-Analyse mit SPSS empfiehlt es sich, ebenfalls nach den Schritten „Datenerhebung“ und „Datenauswertung“ zu unterscheiden. Die Prozeduren ORTHOPLAN und PLANCARDS lassen sich dabei der Phase der Datenerhebung und die Prozedur CONJOINT der Phase der Datenauswertung zuordnen (vgl. Abbildung 9.28):

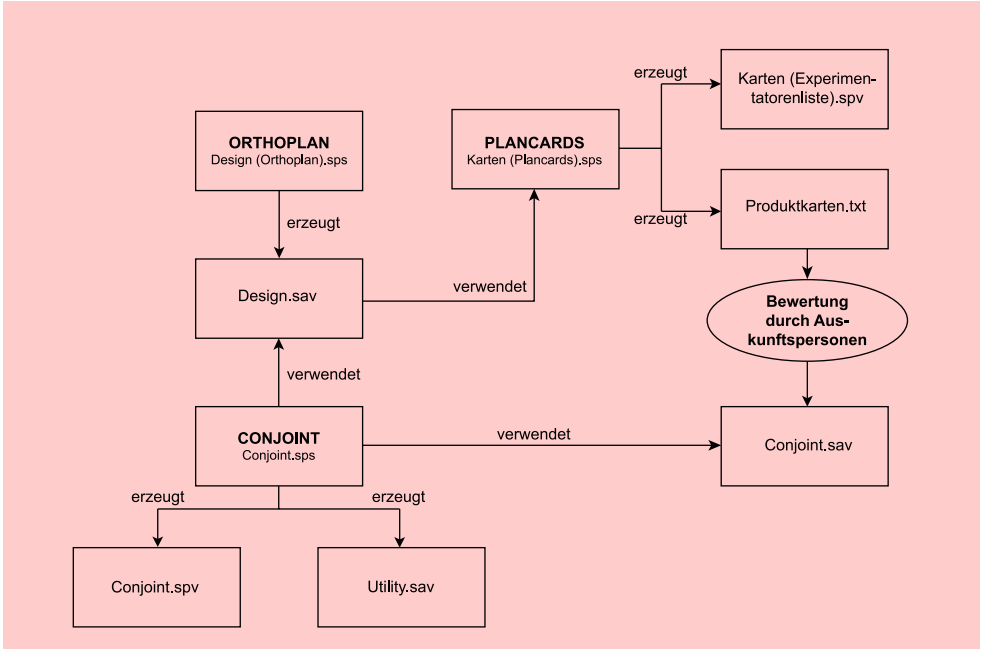


Abbildung 9.28: Zusammenwirken der SPSS-Prozeduren im Rahmen der Conjoint-Analyse

Im Folgenden werden alle in Abbildung 9.28 aufgeführten Prozeduren im Zusammenhang mit den für das Fallbeispiel verwendeten SPSS-Jobs besprochen.

Datenerhebung

Erstellung reduzierter Designs mit Hilfe der Prozedur ORTHOPLAN

Es wurde der in Abbildung 9.29 dargestellte SPSS-Job zur Erstellung des reduzierten Designs verwendet.

```
DATA LIST Free /Preis Verwend Geschmac Kalorien.

VARIABLE LABELS  Preis "Preis"
  /Verwend  "Verwendung"
  /Geschmac "Geschmack"
  /Kalorien "Kaloriengehalt".

VALUE LABELS
   Preis 1 "1,50€" 2 "1,00€" 3 "0,50€"
  /Verwend 1 "Brotaufstrich" 2 "Kochen/Backen/Braten" 3 "universell"
  /Geschmac 1 "Buttergeschmack" 2 "pflanzlich schmeckend"
  /Kalorien 1 "kalorienarm"     2 "normale Kalorien".

BEGIN DATA.
3 3 2 2
1 2 1 1
END DATA.

ORTHOPLAN   holdout = 2.

LIST VARIABLES = ALL.

SAVE OUTFILE = "a:\Conjoint - Orthogonales Design - DATEN.sav".
```

Abbildung 9.29: SPSS-Job zur Erstellung eines reduzierten Designs

SPSS-Prozedur „Orthoplan“

Durch die Prozedur ORTHOPLAN werden zur Vorbereitung der Conjoint-Analyse reduzierte Erhebungsdesigns (orthogonal arrays) ermittelt. ORTHOPLAN benötigt dabei keinen Datensatz. Auf Basis der spezifizierten Variablen errechnet ORTHOPLAN das reduzierte Erhebungsdesign und liefert im Ergebnis eine Aufstellung der notwendigen Anzahl der fiktiven Produkte (als Fälle) mit den jeweiligen Ausprägungen der Eigenschaften (als Variable). Durch den Unterbefehl MINIMUM kann die Anzahl der Stimuli angegeben werden, die durch ORTHOPLAN mindestens erzeugt werden sollen. Wird dieser Befehl nicht verwendet, erstellt ORTHOPLAN zumindest so viele Stimuli, wie sie für ein reduziertes Design benötigt werden.

Vor der Ausführung von ORTHOPLAN sollten noch die den Untersucher besonders interessierenden Eigenschaftskombinationen als Simulations-Karten spezifiziert werden. In unserem Beispiel wurden zwei Simulations-Karten gewählt, die durch die Befehle BEGIN DATA und END DATA eingeschlossen sind. Die Anzahl der gewünschten Holdout-Karten, die ebenfalls von ORTHOPLAN erzeugt werden, kann mit Hilfe des Unterbefehls HOLDOUT angegeben werden. Durch die Prozedur LIST wird in obigem SPSS-Job das Ergebnis abschließend angezeigt und mit Hilfe des Befehls SAVE OUTFILE in der Systemdatei DESIGN.SAV hinterlegt.[18]

Generierung von Produktkarten mit Hilfe der Prozedur PLANCARDS

Abbildung 9.30 zeigt den verwendeten SPSS-Job zur Generierung der Produktkarten.

[18] Bei den folgenden Prozeduren wird auf die Datei DESIGN.SAV zurückgegriffen, um die zufallsbedingten Änderungen im reduzierten Design auszuschließen. Die entsprechenden Daten- und Programmdateien können mit der Bestellkarte am Ende dieses Buches angefordert werden.

```
TITLE "Multivariate Analysemethoden (15. Auflage) - Conjoint".

SUBTITLE "Erstellung der Produktkarten aus dem reduzierten Design".

get file    = "a:\Conjoint - Orthogonales Design - DATEN.sav".

PLANCARDS
  /factor   = preis verwend geschmac kalorien
  /format both
  /paginate
  /title    = "Margarine PLANCARDS"
  /outfile  = "a:\Conjoint - Produktkarten - OUTPUT.txt".
```

Abbildung 9.30: SPSS-Job zur Erstellung der Produktkarten

SPSS-Prozedur „Plancards“

Die Prozedur PLANCARDS verwendet das Ergebnis der Prozedur ORTHOPLAN zur Erstellung von Produktkarten, die dann in der Befragung eingesetzt werden können. Da das Ergebnis der Prozedur ORTHOPLAN in der Systemdatei DESIGN.SAV abgespeichert wurde, wird dieses Ergebnis durch den Befehl GET FILE in den Job eingelesen. Mit Hilfe des Unterbefehls FORMAT kann festgelegt werden, ob die Produktkarten in einer Liste (LIST), als einzelne Karten (CARDS) oder als Liste und Karten (BOTH) ausgegeben werden sollen.

Darüber hinaus können die Produktkarten mit Hilfe des Unterbefehls TITLE mit Kopfzeilen und mit Hilfe des Unterbefehls FOOTER mit Fußzeilen versehen werden. Produktkarten können mit PLANCARDS aber auch unabhängig von der Prozedur ORTHOPLAN erstellt werden, indem der Benutzer selbst das reduzierte Design für die Produktkarten im Rahmen der SPSS-Datendefinitionen bestimmt. Mit dem Unterbefehl OUTFILE wird die Datei spezifiziert, in die das Ergebnis der Prozedur PLANCARDS geschrieben werden soll.

Datenauswertung mit Hilfe der Prozedur CONJOINT

Das für unser Beispiel verwendete Programm zur Durchführung der Conjoint-Analyse mit Hilfe der Prozedur CONJOINT ist in Abbildung 9.31 dargestellt.

```
* MVA: Fallbeispiel Conjointanalyse.

* Conjointanalyse für den Margarinemarkt (Erzeugung der Nutzenwerte).
CONJOINT
  /PLAN = "C:\Design.sav"
  /DATA = "C:\Conjoint.sav"
  /FACTORS = preis (LINEAR MORE) verwend (DISCRETE) geschmac
(DISCRETE) kalorien (LINEAR LESS)
  /SUBJEKT = person
  /RANK = stim1 TO stim9 hold1 hold2
  /PRINT = ALL
  /UTILITY = "C:\Utility.sav".
```

Abbildung 9.31: SPSS-Job zur Conjoint-Analyse

Die eigentliche Conjoint-Analyse wird durch die Prozedur CONJOINT durchgeführt. Mit dem Unterbefehl PLAN wird der Prozedur CONJOINT mitgeteilt, welche Datei die Daten für das *reduzierte Erhebungsdesign* enthält. In unserem Fall ist das die Datei DESIGN.SAV, die zuvor mit Hilfe der Prozedur ORTHOPLAN erzeugt wurde.

Jeder weitere Unterbefehl der Prozedur CONJOINT wird durch einen Schrägstrich (/) eingeleitet. In der vorliegenden Conjoint-Analyse wurden die folgenden Unterbefehle verwendet:

- Der Unterbefehl DATA:
  Die in diesem Beispiel verwendeten *Präferenzwerte der Befragten* wurden im SPSS-Datendokument CONJOINT.SAV hinterlegt. Auf dieses Datendokument kann nun im DATA-Unterbefehl der Prozedur CONJOINT zurückgegriffen werden. Alternativ können die Daten auch im Datendefinitionsteil eingegeben werden und sind damit bereits im ACTIVE-FILE (Spezifikation *) enthalten.
- Der Unterbefehl FACTORS:
  Der Unterbefehl FACTORS bestimmt den Zusammenhang zwischen den jeweiligen Ausprägungen der Stimuli-Eigenschaften (Faktoren) und damit auch die Beziehung zwischen den Eigenschaften und den erhobenen Präferenzurteilen. Bei den Spezifikationen ist streng auf die vorgenommene Kodierung der Stimuli-Eigenschaften zu achten. Vier Modelle stehen zur Verfügung, die bei den verschiedenen Eigenschaften verwendet werden können:
  - DISCRETE:
    Es liegen kategoriale Variable vor und es werden keinerlei Annahmen über die Beziehung zwischen Variablen und Rangwerten gemacht.
  - LINEAR:
    Die Rangwerte stehen in einer linearen Beziehung zu den Variablen. Dabei ist die Richtung des linearen Zusammenhangs durch die Angabe LINEAR MORE (steigende Ausprägungsnummer ist mit einer wachsenden Präferenz verbunden) oder LINEAR LESS (steigende Ausprägungsnummer ist mit einer fallenden Präferenz verbunden) zu konkretisieren.
  - IDEAL:
    Die Rangdaten stehen in einer *negativ* quadratischen Beziehung zu den Variablen. Dabei wird unterstellt, dass eine ideale Eigenschaftsausprägung einer Variablen existiert und zunehmende Abweichungen von diesem „Idealwert“ zu immer stärker werdenden Präferenzeinbußen führen.
  - ANTIIDEAL:
    Die Rangdaten stehen in einer *positiv* quadratischen Beziehung zu den Variablen. Dabei wird unterstellt, dass eine „schlechteste“ Eigenschaftsausprägung einer Variablen existiert und zunehmende Abweichungen von diesem „Antiideal“ zu immer stärker werdenden Präferenzen führen.
- Der Unterbefehl SUBJECT:
  Durch den Unterbefehl SUBJECT wird eine Identifikationsvariable für die Befragten bestimmt. In unserem Fall ist das die Variable „PERSON“, die die Personen-Nummer enthält. Wird keine Identifikationsvariable bestimmt, so gibt die Prozedur CONJOINT keine Einzelanalyse, sondern nur eine Gesamtanalyse aus.

- Die Unterbefehle RANK, SCORE und SEQUENCE:
  Zur Analyse der Präferenzdaten lässt CONJOINT alternativ drei Arten der Datenkodierung zu (vgl. Abschnitt 9.3.1):
  - RANK: (Methode der Rangverteilung)
    Dabei muss die Kodierung der Daten so erfolgen, dass die Reihenfolge der Variablen der Reihenfolge der Produktkarten entspricht. In unserem Fall entspricht die Variable „STIM1“ der Produktkarte Nr. 1, die Variable „STIM2“ der Produktkarte Nr. 2 usw. Beispielsweise hat der Datensatz für Auskunftsperson 33 folgende Form:

      33 5 10 11 6 8 9 2 1 3 4 7

    Nach der laufenden Nummer für die Personen folgen die Rangwerte für die elf Stimuli. Die Auskunftsperson hat dem Stimulus 1 (STIM1) den Rang 5, dem Stimulus 2 (STIM2) den Rang 10 vergeben usw. Die letzten beiden Ziffern 4 und 7 entsprechen den Rangwerten für die Holdout-Karten. Die Ziffern stehen für die Rangwerte der sortierten Stimuli.
  - SCORE: (Präferenzwertmethode)
    Dabei muss die Kodierung der Daten so erfolgen, dass die Reihenfolge der Variablen wiederum der Reihenfolge der Produktkarten entspricht.
  - SEQUENCE: (Methode des Rangordnens)
    Eine Beurteilung in Form von Rang- oder Präferenzwerten ist nicht erfolgt. Die Kodierung der Daten muss hier allerdings so erfolgen, dass die Produktkarte mit der höchsten Präferenz als erste Variable und diejenige mit der kleinsten Präferenz als letzte Variable kodiert wird. Für Auskunftsperson 33 hätte der Datensatz bei der Methode des Rangordnens wie folgt ausgesehen:

      33 8 7 9 10 1 4 11 5 6 2 3

    Stimulus Nr. 8 bekam die höchste Präferenz, Stimulus Nr. 7 die zweithöchste Präferenz usw. zugeordnet. Die Ziffern stehen für die Nummer des jeweiligen Stimulus.
- Der Unterbefehl PRINT:
  Der PRINT-Unterbefehl steuert die Druckausgabe der Prozedur CONJOINT.
- Der Unterbefehl UTILITY:
  Durch den Unterbefehl UTILITY wird ein Systemfile unter dem Namen UTILITY.SAV erzeugt, in dem für jede Person folgende Informationen abgespeichert sind.
  - Personenkennung (Variable „PERSON“)
  - Konstanter Term der Conjoint-Schätzung (Variable „CONSTANT“)
  - Teilnutzenwerte (Variable „VERWEND1“ bis „KALORIEN_L“)
  - Gesamtnutzenwerte des reduzierten Designs (Variable „SCORE1“ bis „SCORE9“)
  - Gesamtnutzenwerte der Holdout-Karten (Variable „SCORE10“ und „SCORE11“)
  - Gesamtnutzenwerte der Simulations-Karten (Variable „SIMUL01“ und „SIMUL02“)

Abbildung 9.32 zeigt den Inhalt der UTILITY-Datei für Person 33. Dabei ist jedoch zu beachten, dass bei den Eigenschaften „Preis“ und „Kaloriengehalt“ nur der Wert des Regressionskoeffizienten B angegeben wird, da mit seiner Hilfe, wie in Abschnitt 9.3.2.1 beschrieben, auf die Teilnutzenwerte geschlossen werden kann.

```
The variables are listed in the following order:

LINE   1: PERSON CONSTANT VERWEN1 VERWEN2 VERWEN3 GESCHM1 GESCHM2
LINE   2: PREIS_L KALORI_L SCORE1 SCORE2 SCORE3 SCORE4 SCORE5
LINE   3: SCORE6 SCORE7 SCORE8 SCORE9 SCORE10 SCORE11 SIMUL01
LINE   4: SIMUL02
.
. Ausdruck für Person 1 bis 32
.
PERSON:   33,00   3,92  -1,67   ,67   1,00   2,25  -2,25
PREIS_L:    ,50   -,50   6,67  2,33   ,00   5,50   4,50
SCORE6:    3,67   7,33   7,67  7,33   7,17   4,00   3,17
SIMUL02:   6,83
.
. Ausdruck für Person 34 bis 40
.
Number of cases read:  40    Number of cases listed:  40
```

Abbildung 9.32: Auszug aus dem Systemfile UTILITY.SAV für Person 33

Entscheidend ist dabei die Angabe der Teilnutzenwerte (VERWEND1 bis KALORIEN_L), da sich mit ihrer Hilfe die Gesamtnutzenwerte aller Stimuli errechnen lassen und sie durch andere Prozeduren (wie z. B. durch die Clusteranalyse) eingelesen werden können. Darüber hinaus lassen sich durch den File UTILITY.SAV unmittelbar die Gesamtnutzenwerte der Stimuli des reduzierten Designs und der Holdout-Karten ablesen, die im Ausdruck der Individualanalysen (vgl. Abbildung 9.21) nicht enthalten sind.

Abschließend sei noch darauf hingewiesen, dass die Prozedur ORTHOPLAN jeweils nach einem Zufallsprinzip reduzierte Designs erstellt, wodurch mit jedem ORTHOPLAN-Aufruf jeweils unterschiedliche reduzierte Designs erzeugt werden. Die Prozedur CONJOINT lässt aber auch die *Vorgabe eines reduzierten Designs durch den Anwender* zu.

Die Behandlung von Missing Values

Missing Values

Als fehlende Werte (MISSING VALUES) bezeichnet man Variablenwerte, die von den Befragten entweder außerhalb des zulässigen Beantwortungsintervalls vergeben wurden oder überhaupt nicht eingetragen wurden. Die Prozedur CONJOINT ist nicht in der Lage, solche fehlenden Werte zu handhaben. Sobald fehlende Werte bei den Rang- oder Präferenzwerten auftreten, wird der entsprechende Fall aus der Analyse ausgeschlossen.

9.4 Anwendungsempfehlungen

9.4.1 Durchführung einer klassischen Conjoint-Analyse

Zusammenfassend lassen sich für den Einstieg in eine Conjoint-Analyse folgende Empfehlungen geben:

1. Eigenschaften und Eigenschaftsausprägungen:
   Die Zahl der Eigenschaften und Eigenschaftsausprägungen ist möglichst gering zu halten. Weiterhin ist darauf zu achten, dass es sich um voneinander unabhängige Eigenschaften handelt, die für die Untersuchung relevant sein müssen. Ebenso müssen die Eigenschaftsausprägungen bei der Produktgestaltung konkret umsetzbar sein.

2. Erhebungsdesign:
   Nach Möglichkeit sollten im Erhebungsdesign nicht mehr als maximal 20 fiktive Produkte enthalten sein. Wird diese Zahl im vollständigen Design überschritten, so sollte ein reduziertes Design unter Verwendung der Profilmethode erstellt werden.

3. Bewertung der Stimuli:
   Die Befragungsmethode kann jeweils nur in Abhängigkeit von der konkreten Fragestellung festgelegt werden.

4. Schätzung der Nutzenwerte:
   Der Schätzung sollte ein additives Nutzenmodell zugrunde liegen.

5. Spezifikation der Nutzenmodelle:
   Wird bei der Verwendung von stark reduzierten Designs für keine der Eigenschaften ein Nutzenmodell festgelegt, d. h. wenn alle Eigenschaften vom Typ (DISCRETE) sind, so führt dies dazu, dass die Zahl der Freiheitsgrade für die individuelle Modellschätzung sehr gering oder gar Null ist. In einem solchen Fall können die empirischen Präferenzdaten zwar perfekt abgebildet werden, jedoch ist aufgrund der starken Orientierung an den Präferenzurteilen des reduzierten Designs nur von einer eingeschränkten Validität der Ergebnisse auszugehen. Eine entsprechende Prüfung anhand von Holdout-Karten ist hier stark empfohlen.

6. Aggregation der Nutzenwerte:
   Die gemeinsame Conjoint-Analyse kann zu einer größeren Differenzierung der Teilnutzenwerte einzelner Eigenschaften und damit zu besser interpretierbaren Werten führen. Wenn die Anzahl der Daten nicht zu groß ist, ist die gemeinsame Conjoint-Analyse der Aggregation der Einzelanalysen vorzuziehen.

7. Segmentierung:
   Eine Aggregation (oder gemeinsame Analyse) über alle Personen ist nur bei hinreichender Homogenität der individuellen Teilnutzenwerte gerechtfertigt. Dies sollte mit Hilfe einer Clusteranalyse (vgl. Kapitel 8) überprüft werden. Bei ausgeprägter Heterogenität sind segmentspezifische Analysen durchzuführen.

Obige Empfehlungen können nur Grundsatzaussagen darstellen, die in der konkreten Anwendungssituation einer geeigneten Relativierung bedürfen und kritisch zu hinterfragen sind.[19]

[19] Vgl. Weiber/Rosendahl (1997), S. 113 ff.

9.4.2 Anwendung alternativer conjointanalytischer Verfahren

Die Conjoint-Analyse hat in jüngster Zeit weite Verbreitung in der empirischen Forschung gefunden. Entsprechend breit sind auch die existierenden Verfahrensvarianten der Conjoint-Analyse. Die nachfolgend differenzierten Ansätze (vgl. Abb. 9.33) der Conjoint-Analyse unterscheiden sich vor allem im Hinblick auf die Erhebung der Präferenzurteile. Dabei ist jedoch zu beachten, dass innerhalb der jeweiligen Verfahren noch eine Vielzahl von Optionen zur Verfügung steht, wie z. B. Art der Erhebung, Wahl des Schätzalgorithmus, Art der verwendeten Skala, die entweder in einem oder aber auch in mehreren Verfahren Anwendung finden können.[20] Aufgrund ihrer in der Praxis und Literatur erlangten Bedeutung und der Verfügbarkeit entsprechender Softwareprodukte werden im Folgenden aber nur die in Abbildung 9.33 grau hinterlegten conjointanalytischen Verfahren einer kurzen Betrachtung unterzogen:

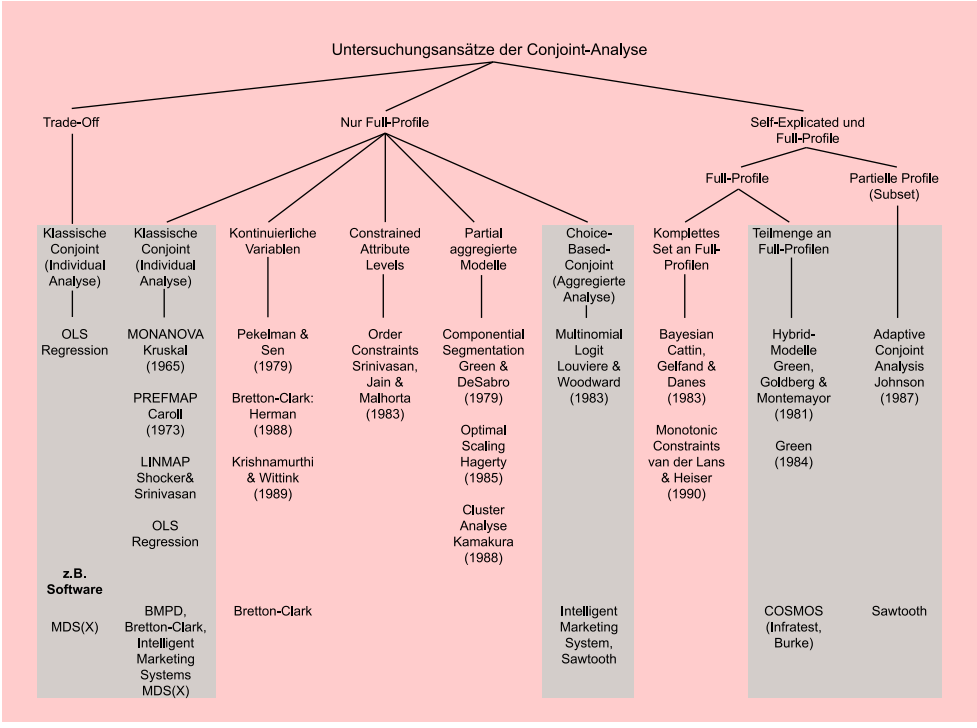


Abbildung 9.33: Alternative Untersuchungsansätze der Conjoint-Analyse (in Anlehnung an: Carroll/Green (1995), S. 386.)

Untersuchungsansätze der Conjoint-Analyse

Die ***klassischen Untersuchungsansätze*** der Trade-off- und der Profilmethode wurden bereits in Abschnitt 9.2.2 dieses Kapitels behandelt, sodass an dieser Stelle lediglich darauf hingewiesen werden soll, dass insbesondere die Profilmethode Gegenstand einer Reihe von Erweiterungen und Verbesserungen geworden ist. Zum einen wurde die traditionelle Teilnutzenwert-Modellierung um eine Mischung aus linearen und quadratischen Teilnutzen-Parametern erweitert. Zum anderen wurde eine Verbesserung von Validität und Reliabilität durch „Constrained Attribute

[20] Vgl. zu den nachfolgenden Ausführungen insb. Weiber/Rosendahl (1997), S. 107 ff. sowie Voeth (2000), S. 77 ff.

Levels“ erreicht, um die Monotonie innerhalb der Attribute sicherzustellen sowie durch die Verwendung unterschiedlicher partialer Aggregationsmethoden.[21]

Hybrid-Conjoint-Analyse
Self-Explicated Modell

Um umfangreich fraktionierte Conjoint-Designs auf mehrere Personen verteilen zu können, wird im Rahmen der ***Hybrid-Conjoint-Analyse*** die Verknüpfung eines Punktbewertungsmodells (Self-Explicated-Modell) mit einem Conjoint-Ansatz vorgenommen. Mit Hilfe des Punktbewertungsmodells werden zunächst die individuellen Wichtigkeiten aller relevanten Merkmale sowie die Erwünschtheit ihrer Merkmalsausprägungen individuell erfragt und die hier gewonnenen Beurteilungswerte zur Bildung von Personengruppen mit homogenen Beurteilungsstrukturen verwendet.[22] Darauf aufbauend wird das für eine Auskunftsperson zu große Master-Design in Teilblöcke zerlegt, und jedes der Gruppenmitglieder beurteilt nur noch *einen Teilblock*. Damit lassen sich zunächst Nutzenwerte auf Gruppenebene ermitteln. Zur Bestimmung der individuellen Nutzenwerte werden im Unterschied zur klassischen Conjoint-Analyse zusätzlich zu den empirischen Präferenzurteilen auch die Daten des Self-Explicated-Models herangezogen. Dadurch ergibt sich die für hybride Modelle typische Verknüpfung eines dekompositionellen mit einem kompositionellen Ansatz.[23] Allerdings ist zu beachten, dass im Gegensatz zur klassischen Conjoint-Analyse keine „rein“ individuellen Nutzenfunktionen berechnet werden können, da die aus den folgenden Schätzergebnissen „quasi-individuell“ hergeleiteten Teilnutzenbeträge immer noch von den Parametern des aggregierten Conjoint-Modells beeinflusst sind. Dies liegt darin begründet, dass die Schätzung der Funktionsparameter nur auf Gruppenniveau erfolgen kann, da die Bewertungen aller Stimuli des fraktionierten Designs in die Schätzung einzubeziehen sind und diese vollständig nur auf Gruppenebene vorliegen.

Adaptive-Conjoint-Analyse

Auch die ***Adaptive-Conjoint-Analyse*** stellt ein hybrides Modell dar, da die ganzheitlich zu beurteilenden Alternativkonzepte (dekompositioneller Teil) aufgrund der vorher *individuell erfragten Relevanz* und *Wichtigkeit* der Merkmale und Merkmalsausprägungen (kompositioneller Teil) erzeugt werden. Allgemein umfasst die Adaptive-Conjoint-Analyse folgende Ablaufschritte:

1. Bewertung der individuell relevanten Eigenschaftsausprägungen. In diesem Schritt können auch durch den Befragten völlig unakzeptable Ausprägungen eliminiert werden.

2. Bestimmung der Wichtigkeit jeder Eigenschaft anhand der zuvor festgelegten besten und schlechtesten Eigenschaftsausprägung.

3. Paarweise Präferenzbestimmung bei Teilprofilen mit maximal fünf Eigenschaften. Hierbei wird eine Annäherung von einfachen zu realistischeren Konzepten empfohlen, wobei sich Teilprofile mit drei Eigenschaften bewährt haben.

4. Präferenzbestimmung und Ermittlung von Teilnutzenwerten anhand kalibrierter und hinsichtlich des Gesamtnutzens ähnlicher Einzelkonzepte. Dies stellt eine möglichst genaue Schätzung der Teilnutzenwerte sicher.

Der gesamte Befragungsablauf erfolgt dabei computergestützt und orientiert sich am Beurteilungsverhalten jeder einzelnen Auskunftsperson. Da bei der adaptiven

[21] Vgl. Carroll/Green (1995), S. 387 ff.; Kamakura (1988), S. 157 ff.
[22] Vgl. Green (1984), S. 156 ff.
[23] Vgl. Hensel-Börner (2000), Kap. 2.5.2.

Conjoint-Analyse tatsächlich spezifische Erhebungsdesigns für jede Auskunftsperson erstellt werden, kann hier von einer *echten Individualanalyse* gesprochen werden. Die individuell angepassten Ehrhebungsdesigns bieten dabei den Vorteil auch Studien mit einer großen Anzahl von Eigenschaften (bis max. 30) und Eigenschaftsausprägungen (bis max. 9) durchführen zu können.[24]

Beurteilungskriterien	Klassiche Ansätze	Choice-Based-CA	Hybrid-CA	Adaptive CA
Erhebungsart:				
persönlich, schriftlich	++	+	++	−−
persönlich, computer-gestützt	Ø	++	Ø	++
postalisch, schriftlich	Ø	Ø	Ø	−−
postalisch, computer-gestützt	−	++	−	++
telefonisch	Ø	(+)	−	+
Erhebungssituation:				
Große Merkmalsanzahl	−−	−−	++	++
Individualanalyse	++	−−	+	++
individuelle Erhebungs-profile	−	−	−−	++
Anwendungssituation:				
Auswahlentscheidungen	Ø	++	Ø	Ø
Berücksichtigung der Simularität	−	++	−	−
Bestimmung von Marktreaktionen	Ø	++	Ø	Ø
Marktsegmentierung	++	(Ø)	−	(+)
Auswertungssituation:				
Inferenzstatistik	−	++	−	−

Legende:
Eignung: ++ = sehr gut + = gut Ø = durchschnittlich − = gering −− = ungeeignet

Abbildung 9.34: Vergleichende Bewertung alternativer conjointanalytischer Untersuchungsansätze

Auswahlbasierte Conjoint-Analyse (Choice-Based-Conjoint-; Discrete-Choice-Analyse)

Die ***Choice-Based-Conjoint-Analyse***, häufig auch als ***Discrete-Choice-Analyse*** bezeichnet, gehört zur Gruppe der „Auswahlbasierten Conjoint-Analyse“ und wird in diesem Buch den „*Fortgeschrittenen Verfahren der multivariaten Datenanalyse*“ zugeordnet. Aufgrund ihrer großen praktischen Bedeutung werden diese Verfahren als eigenständiges Kapitel in dem Buch „*Backhaus, K./Erichson, B./Weiber, R.: Fortgeschrittene Multivariate Analysemethoden, 3. Auflage, Berlin 2015.*“ behandelt. Hier sei nur darauf hingewiesen, dass sich diese Verfahren nicht nur bei der Bewertung der Stimuli von den vorher genannten Verfahren unterscheiden, sondern auch bezüglich ihrer theoretischen Grundlagen.[25] Im Gegensatz zu den zuvor erläuterten Untersuchungsansätzen werden im Rahmen der Choice-Based-Conjoint-Analyse von den Auskunftspersonen Präferenzurteile in Form von *Auswahlentscheidungen* verlangt. Die „Bewertung“ der Stimuli erfolgt dabei durch

[24] Vgl. Schubert (1995), Sp. 380.
[25] Vgl. Louviere/Woodworth (1983), S. 352 ff.

einmalige oder wiederholte Auswahl eines Stimulus aus einem Alternativen-Set. Im Gegensatz zu allen anderen Methoden kann damit auch eine *Nichtwahl-Möglichkeit* im Alternativen-Set berücksichtigt werden.[26]

Zusammenfassende Empfehlungen

Abschließend seien hier durch die nachfolgende Abbildung Empfehlungen für den Einsatz der oben skizzierten Verfahrensvarianten der Conjoint-Analyse gegeben, wobei die Beurteilungskategorien nach Erhebungsart sowie Erhebungs-, Anwendungs- und Auswertungssituation zu unterscheiden sind.

Bezüglich der Auswertungssituation ist hier noch anzumerken, dass lediglich bei der Choice-Based-Conjoint-Analyse Inferenzstatistiken berechnet werden können, während die nicht-metrischen Verfahren nur Fitmaße bereitstellen können.

9.5 Mathematischer Anhang

Berechnung der Teilnutzenwerte durch Regressionsanalyse

Dummy-Variable

Bei Durchführung einer Regression der p-Werte auf die Dummy-Variablen ist darauf zu achten, dass von den M_j Dummy-Variablen einer Eigenschaft j nur $(M_j - 1)$ Variablen linear unabhängig sind. Je Eigenschaft ist daher eine der Dummy-Variablen zu eliminieren, sodass insgesamt nur

$$Q = \sum_{j=1}^{J} M_j - K \tag{9.14}$$

Dummy-Variablen zu berücksichtigen sind. Im Beispiel ergibt sich $Q = 3$. Die der eliminierten Dummy-Variable zugehörige Merkmalsausprägung wird als Basisausprägung der betreffenden Eigenschaft betrachtet. Geschätzt werden sodann die Abweichungen von den jeweiligen Basisausprägungen. Wählt man jeweils die letzte Ausprägung einer Eigenschaft als Basisausprägung, so gelangt man zu folgender Datenmatrix:

Empirische Werte	Dummies			Geschätzte Werte
p_k	X_{A1}	X_{A2}	X_{B1}	y_k
2	1	0	1	1,6667
1	1	0	0	1,3333
3	0	1	1	3,6667
4	0	1	0	3,3333
6	0	0	1	5,6667
5	0	0	0	5,3333

Regression

Die zu schätzende Regressionsgleichung lautet allgemein:

$$y_k = a + \sum_{j=1}^{J} \sum_{m=1}^{M_j - 1} b_{jm} \cdot x_{jm} \tag{9.15}$$

Für das Beispiel ergibt sich:

$$y_k = 5,3333 - 4,0 \cdot x_{A1} - 2,0 \cdot x_{A2} - 0,3333 \cdot x_{B1} \qquad (R^2 = 0,924)$$

[26] Bei der TCA kann durch Legen einer Limit-Card eine Auswahlentscheidung „künstlich“ integriert werden. Vgl. Hahn/Voeth (1997).

Diese Gleichung liefert dieselben Gesamtnutzenwerte y_k, die man auch bei Anwendung der Varianzanalyse erhält. Die Teilnutzenwerte b_{jm} sind gegenüber den zuvor erhaltenen Werten β_{jm} andersartig skaliert. Die β_{jm} sind für jede Eigenschaft j um den Nullpunkt zentriert, und man erhält sie durch folgende Transformation:

$$\beta_{jm} = b_{jm} - \bar{b}_j \tag{9.16}$$

Die Differenzen zwischen den Teilnutzenwerte für die Eigenschaft j sind dagegen identisch, wie sich leicht nachprüfen lässt. Damit liefern beide Verfahren auch gleiche Wichtigkeiten der Eigenschaften.

Literaturhinweise

A. Basisliteratur zur Conjointanalyse

Dietz, W. (2007), Grundlagen der Conjoint-Analyse: Varianten, Vorgehensweise, Anwendungen, Bochum.

Hair, J./Black, W./Babin, B./Anderson, R. (2014), Multivariate Data Analysis, 7. Auflage, Englewood Cliffs (N.J.), Kapitel 3.

Kaltenborn, T./Fiedler, H./Lanwehr, R./Melles, T. (2013), Conjoint-Analyse, München.

Teichert, T. (2001), Nutzenschätzung in Conjoint-Analysen, Wiesbaden.

Teichert, T./Sattler, H./Völckner, F. (2008), Traditionelle Verfahren der Conjoint-Analyse, in: Herrmann A./Homburg C./Klarmann, M. (Hrsg.): Handbuch Marktforschung - Methoden, Anwendungen und Praxisbeispiele, 3. Auflage, Wiesbaden, S. 651–685.

B. Zitierte Literatur

Addelman, S. (1962), Orthogonal Main-Effect Plans for Asymmetrical Factorial Experiments, in: *Technometrics*, Vol. 4, Nr. 1, S. 21–46.

Carroll, D./Green, P. (1995), Psychometric Methods in Marketing Research: Part I, Conjoint Analysis, in: *Journal of Marketing Research*, Vol. 32, Nr. 4, S. 385–391.

Green, P. (1984), Hybrid Models for Conjoint Analysis: An Expository Review, in: *Journal of Marketing Research*, Vol. 21, Nr. 2, S. 155–169.

Hahn, C./Voeth, M. (1997), Limit-Cards in der Conjoint-Analyse – eine Modifikation der traditionellen Conjoint-Analyse, Arbeitspapier Nr. 21 des Betriebswirtschaftlichen Instituts für Anlagen und Systemtechnologien der Westfälischen Wilhelms-Universität Münster, Münster.

Hensel-Börner, S. (2000), Validität computergestützter hybrider Conjoint-Analysen, Wiesbaden.

Johnson, R. (1974), Trade-Off-Analysis of Consumer Values, in: *Journal of Marketing Research*, Vol. 11, Nr. 2, S. 121–127.

Kamakura, W. (1988), A Least Squares Procedure for Benefit Segmentation with Conjoint Experiments, in: *Journal of Marketing Research*, Vol. 25, Nr. 2, S. 157–167.

Kruskal, J. (1964a), Multidimensional Scaling by Optimizing Goodness of Fit to a Nonmetric Hypothesis, in: *Psychometrika*, Vol. 29, Nr. 1, S. 1–27.

Kruskal, J. (1964b), Nonmetric Multidimensional Scaling: A Numerical Method, in: *Psychometrika*, Vol. 29, Nr. 2, S. 115–129.

Kruskal, J. (1965), Analysis of factorial experiments by estimating a monotone transformation of data, in: *Journal of Royal Statistical Society*, Series B, Vol. 27, Nr. 2, S. 251–263.

Louviere, J./Woodworth, G. (1983), Design and Analysis of Simulated Consumer Choice or Allocation Experiments: An Approach Based on Aggregate Data, in: *Journal of Marketing Research*, Vol. 20, Nr. 4, S. 350–367.

Nitschke, T./Völckner, S. (2006), Präferenzmessung bei unsicheren Produkteigenschaften: Risikoberücksichtigung bei Ergebnissen aus Conjoint-Analysen, in: *Zeitschrift für betriebswirtschaftliche Forschung*, Vol. 58, Nr. 6, S. 743–770.

Schirm, K. (1995), Glaubwürdigkeit von Produktvorankündigungen, Wiesbaden.

Schubert, B. (1995), Conjoint-Analyse, in: Tietz B./Köhler, R./Zentes J. (Hrsg.): Handwörterbuch des Marketing, 2. Auflage, Stuttgart, S. 376–389.

Street, D. J./Burgess, L./Louviere, J. J. (2005), Quick and easy choice sets: Constructing optimal and nearly optimal stated choice experiments, in: *International Journal of Research in Marketing*, Vol. 22, Nr. 4, S. 459–470.

Szuppa, S. (2009), Marktforschung für das „Intelligente Haus“, in: Baier, D./Brusch, M. (Hrsg.): Conjoint-Analyse: Methoden, Anwendungen und Praxisbeispiele, Berlin/Heidelberg, S. 265–284.

Teichert, T./Shehu, E. (2009), Diskussion der Conjointanalyse in der Forschung, in: Baier, D./Brusch, M. (Hrsg.): Conjoint-Analyse, Berlin, S. 19–39.

Voeth, M. (2000), Nutzenmessung in der Kaufverhaltensforschung: Die Hierarchische Individualisierte Limit Conjoint-Analyse (HILCA), Wiesbaden.

Weiber, R./Billen, P. (2004), Das Markenspannen-Portfolio zur Bestimmung des Dehnungspotentials einer Dachmarke: Theoretische Analyse und empirische Belege, in: Boltz, D.-M./Leven, W. (Hrsg.): Effizienz in der Markenführung, Hamburg, S. 72–91.

Weiber, R./Mühlhaus, D. (2009), Auswahl von Eigenschaften und Ausprägungen bei der Conjoint Analyse, in: Baier, D./Brusch, M. (Hrsg.): Conjoint-Analyse, Berlin.

Weiber, R./Rosendahl, T. (1997), Anwendungsprobleme der Conjoint-Analyse: Die Eignung conjointanalytischer Untersuchungsansätze zur Abbildung realer Entscheidungsprozesse, in: *Marketing – Zeitschrift für Forschung und Praxis*, Vol. 19, Nr. 2, S. 107–118.

Wirtz, B./Olderog, T./Heithecker, S. (2003), Präferenzen und Zahlungsbereitschaften für Anwendungen und Dienste im Mobilen Internet und deren Implikationen für die Diffusion, in: *Zeitschrift für Betriebswirtschaft*, Vol. 73, Nr. 1, S. 73–98.

Teil III

Fortgeschrittene Verfahren der multivariaten Analyse

Mit den in **Teil II** dieses Buches vorgestellten neun Verfahren der multivariaten Datenanalyse wird aus Sicht der Autoren des vorliegenden Buches das in der Bachelorausbildung an Hochschulen primär relevante Spektrum an multivariaten Analysemethoden abgedeckt. Diese Verfahren wurden deshalb hier auch als „Grundlegende Verfahren“ bezeichnet.

Darüber hinaus existieren weitere multivariate Analysemethoden, die aus Sicht der Autoren eher in der Masterausbildung oder im Rahmen von Promotionsstudien von Bedeutung sind. Diese Verfahren werden von uns auf folgende sieben Analysemethoden eingegrenzt, die wir als „Fortgeschrittene Verfahren“ bezeichnen:

- Nichtlineare Regression,
- Strukturgleichungsanalyse,
- Konfirmatorische Faktorenanalyse,
- Auswahlbasierte Conjoint-Analyse,
- Neuronale Netze,
- Multidimensionale Skalierung,
- Korrespondenzanalyse.

Im nachfolgenden **Teil III** werden diese Verfahren jeweils auf ca. 5-7 Seiten kurz erläutert, wobei alle Kapitel folgende Dreiteilung aufweisen:

1. Problemstellung
2. Allgemeine Vorgehensweise
3. Umsetzung mit SPSS

Durch die kurzen Beschreibungen soll auch den Lesern des vorliegenden Buches ein erster Eindruck von den Einsatzmöglichkeiten der „fortgeschrittenen Verfahren“ gegeben werden. Eine ausführliche und einsteigergerechte Behandlung der Verfahren findet der Leser in unserem Buch:

> **Backhaus, Klaus/Erichson, Bernd/Weiber, Rolf:**
> Fortgeschrittene Multivariate Analysemethoden.
> Eine anwendungsorientierte Einführung, 3. Aufl., Berlin 2015

In obigem Buch werden alle Verfahren ebenfalls am Beispiel des Margarinekaufs erläutert und mit geringstmöglichen Anforderungen an mathematische Vorkenntnisse behandelt. Auch hier wird die Grundidee der Verfahren zunächst mit Hilfe eines „kleinen Handbeispiels“ verdeutlicht und anschließend die Durchführung mit Hilfe von SPSS an Hand eines umfangreicheren Fallbeispiels erläutert.

10 Nichtlineare Regression

Die Nichtlineare Regression wird in diesem Buch den „*Fortgeschrittenen Verfahren der multivariaten Analyse*“ zugeordnet und in diesem Kapitel nur in den Grundzügen behandelt. Eine ausführliche Darstellung zum Kapitel „Nichtlineare Regression“ ist verfügbar in dem Buch „*Backhaus, K./Erichson, B.Weiber, R.: Fortgeschrittene Multivariate Analysemethoden, 3. Aufl., Berlin Heidelberg 2015.*“

10.1 Problemstellung

Modellkomplexität

Wissenschaft und insbesondere Modellbildung sind immer eine Gratwanderung zwischen Simplifizierung und Verkomplizierung. Generell ist simpleren Modellen gegenüber komplexeren Modellen der Vorzug zu geben, solange keine relevanten Aspekte vernachlässigt werden. Die Komplexität eines Modells muss der Problemstellung angemessen sein, d. h. dem abzubildenden Phänomen wie auch dem Zweck des Modells. Für viele Fragestellungen im Rahmen der Regressionsanalyse ist es ausreichend, nichtlineare Zusammenhänge durch lineare Modelle zu approximieren.

Nichtlineare Modelle

Die Schätzung von nichtlinearen Modellen mittels nichtlinearer Regression ist mit einer Reihe von Schwierigkeiten verbunden. Sie ist sehr rechenaufwändig, da sich das dabei entstehende mathematische Problem nicht mehr analytisch lösen lässt. Vielmehr müssen die Schätzwerte numerisch mittels iterativer Algorithmen berechnet werden. Außerdem gibt es oft keine Gewähr, dass der dabei verwendete Optimierungsalgorithmus konvergiert oder ein globales Optimum findet. Iterative Verfahren erfordern, dass der Untersucher Startwerte für die zu schätzenden Parameter vorgibt. Von der Wahl dieser Startwerte hängt ab, ob und wie schnell der Algorithmus das Optimum findet. Es werden also bei Anwendung der nichtlinearen Regression auch erhöhte Anforderungen an den Untersucher gestellt. Ein weiterer Nachteil ist, dass die statistischen Tests, die bei der linearen Regressionsanalyse zur Prüfung der Güte des Modells oder der Signifikanz der Parameter verwendet werden, für die nichtlineare Regression nicht anwendbar sind. Allerdings erfordert nicht jedes nichtlineare Modell die Anwendung der nichtlinearen Regression.

Ein Modell $Y = f(X, u)$ ist linear in Bezug auf die unabhängige Variable X, wenn die Darstellung im X, Y-Diagramm eine Gerade ergibt und somit folgende Form hat (Abbildung 10.1a):

$$Y = \alpha + \beta \cdot X + u \tag{10.1}$$

K. Backhaus, B. Erichson, W. Plinke, R. Weiber, *Multivariate Analysemethoden*,
https://doi.org/10.1007/978-3-662-56655-8_11

Das Modell (10.1) ist aber auch linear in α und β, d. h. Y wächst linear mit α und β.

Das folgende Modell, das Potenz-Modell, ist zwar ebenfalls linear in α und β, aber nicht in dem Parameter γ:

$$Y = \alpha + \beta \cdot X^{\gamma} + u \tag{10.2}$$

Es ist außerdem nichtlinear in X, falls $\gamma \neq 1$ gilt. Ein Diagramm, das Y über X abbilden würde, hätte dann einen nichtlinearen Verlauf (Abbildung 10.1 b). Ebenso hätte ein Diagramm, das Y über γ abbilden würde, einen nichtlinearen Verlauf.

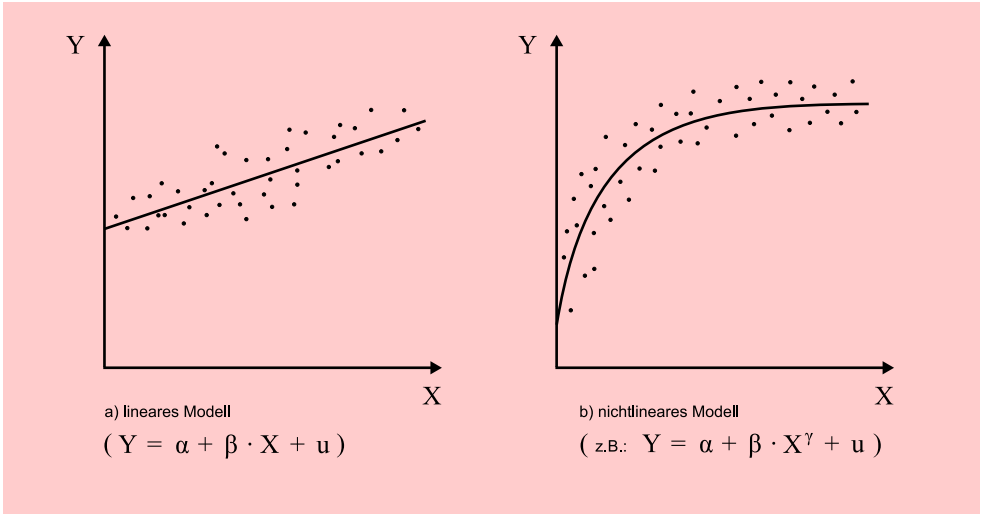


Abbildung 10.1: Lineare und nichtlineare Regressionsgleichung

Es ist generell zwischen zwei *Arten von nichtlinearen Modellen* zu unterscheiden:

Intrinsisch lineare Modelle

a) Linearisierbare Modelle (sog. *intrinsisch lineare Modelle*). Sie sind nichtlinear in den Variablen, aber linear in den zu schätzenden Parametern.

Intrinsisch nichtlineare Modelle

b) Nicht linearisierbare Modelle (sog. *intrinsisch nichtlineare Modelle*). Sie sind nichtlinear in den Variablen und in den zu schätzenden Parametern.

In Abbildung 10.2 werden einfache Beispiele von intrinsisch linearen und nichtlinearen Modellen gegenübergestellt, die kurz besprochen werden sollen. Alle dargestellten Modelle finden z. B. Anwendung im Rahmen der Werbewirkungsforschung (Abhängigkeit der Werbeerinnerung von der Zahl der Werbekontakte, Abhängigkeit der Absatzmenge von der Höhe des Werbebudgets). Das Potenzmodell (10.2) bzw. (a2) ist nichtlinear in X und γ und somit intrinsisch nichtlinear. Gibt man aber für den Parameter γ einen beliebigen Wert c vor, so lässt es sich damit linearisieren und mittels linearer Regression schätzen:

Potenzmodell

$$Y = \alpha + \beta \cdot X' + u \quad \text{mit} \quad X' = X^c \tag{10.3}$$

Quadratwurzel-Modell

Für $c = 0,5$ ergibt sich z. B. das Quadratwurzel-Modell (a1). Soll aber auch γ geschätzt werden, so muss die nichtlineare Regression angewendet werden.

intrinsisch lineare Modelle	intrinsisch nichtlineare Modelle
Quadratwurzel-Modell (a1) $Y = \alpha + \beta \cdot X^{0,5} + u$	Potenz-Modell (a2) $Y = \alpha + \beta \cdot X^{\gamma} + u$
Multiplikatives Modell (b1) $Y = \alpha \cdot X^{\beta} \cdot u$ $ln(Y) = ln(\alpha) + \beta \cdot ln(X) + ln(u)$	Multiplikatives Modell mit additiver Störgröße (b2) $Y = \alpha + X^{\beta} + u$
Logistisches Modell (c1) $Y = \frac{c}{1+e^{\alpha+\beta \cdot X+u}}$ $ln(\frac{c}{Y} - 1) = \alpha + \beta \cdot X + u$	Logistisches Modell (c2) $Y = \frac{c}{1+e^{\alpha+\beta \cdot X}} + u$ (c3) $Y = \frac{\gamma}{1+e^{\alpha+\beta \cdot X+u}}$

Abbildung 10.2: Beispiele für nichtlineare Modelle

Intrinsisch lineare Modelle mit J unabhängigen Variablen enthalten maximal $J+1$ zu schätzende Parameter (bzw. J Parameter, wenn man auf ein konstantes Glied verzichtet), während intrinsisch nichtlineare Modelle auch mehr Parameter enthalten können. So enthält das Modell (a2) hier mit J = 1 drei zu schätzende Parameter.

Parameterzahl

Die Modelle (b1) und (b2) unterscheiden sich nur in Bezug auf die Spezifikation des stochastischen Terms u. In das multiplikative Modell (b1) geht auch die Störgröße u multiplikativ ein. Modell (b2) ist dagegen nicht linearisierbar, da die Störgröße hier additiv eingeht und der rechte Teil der Gleichung so nicht mehr logarithmiert werden kann.

Multiplikatives Modell

Das logistische Modell ist in der Form (c1) linearisierbar, wenn c gegeben ist. (c2) unterscheidet sich davon nur in der Spezifikation des stochastischen Terms u. In der Form (c3) enthält das logistische Modell drei zu schätzende Parameter (α, β und γ) und ist deshalb nicht linearisierbar.

Logistisches Modell

10.2 Allgemeine Vorgehensweise

Nichtlineare Phänomene kommen in vielen Bereichen und Formen vor, und entsprechend existieren vielfältige nichtlineare Modelle. Eine Gruppe nichtlinearer Modelle sind Wachstumsmodelle, die zur Beschreibung und Prognose von Wachstumsprozessen in der Biologie, Medizin, Soziologie und Wirtschaft Anwendung finden.

Wachstumsmodelle

Als ein Beispiel für einen Wachstumsprozess wollen wir die Ausbreitung des Mobiltelefons in Deutschland betrachten (vgl. Abbildung 10.3).

Mobiltelefon

Die Entwicklung der Mobilfunkteilnehmer in Abbildung 10.3 lässt keinen idealtypischen Verlauf erkennen. Um das Jahr 2000 zeigt sich im Wachstum ein Wendepunkt, d. h. das Wachstum schwächt sich ab. Es scheint dann aber noch einmal progressiv anzusteigen, wenngleich nicht so stark wie in der zweiten Hälfte der 90er Jahre. Diese Zunahme ist vermutlich durch die zahlreichen Innovationen und Preissenkungen in diesem Markt bedingt.

Zur Beschreibung von Wachstumsprozessen kann auf ein umfangreiches Arsenal von Modellen zurückgegriffen werden. In Abbildung 10.4 sind elementare Wachstumsmodelle zusammengestellt. Die unabhängige Variable bildet die Zeit t. Da sich in nichtlinearen Modellen die stochastische Komponente unterschiedlich spezifizieren lässt, ist hier anstelle eines zugrundeliegenden stochastischen Modells nur dessen geschätzte Funktion dargestellt.

t	Jahr	Mobilfunkteil-nehmer in Mio.
1	1990	0,273
2	1991	0,532
3	1992	0,953
4	1993	1,768
5	1994	2,482
6	1995	3,764
7	1996	5,554
8	1997	8,286
9	1998	13,913
10	1999	23,470
11	2000	48,247
12	2001	56,126
13	2002	59,128
14	2003	64,839
15	2004	71,322
16	2005	79,271
17	2006	85,652
18	2007	97,151

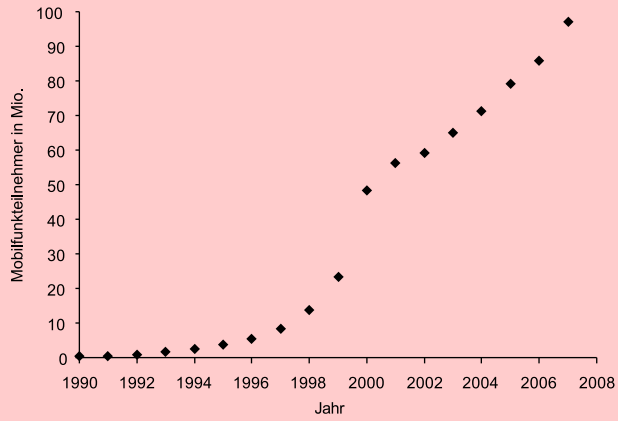


Abbildung 10.3: Entwicklung der Mobilfunkteilnehmer in deutschen Mobilfunknetzen

Wachstumsmodelle im engeren Sinn

Wachstumsmodelle im engeren Sinn sind solche, die gegen eine obere Grenze konvergieren (man spricht auch von Sättigungsmodellen). Beispiele sind das *Exponentielle Wachstumsmodell*, das einen konkaven Verlauf hat, sowie das *Logistische Wachstumsmodell* und das *Gompertz-Modell*, die beide einen S-förmigen Verlauf haben. Sie konvergieren alle gegen die Wachstumsgrenze (Sättigungsgrenze) M, die hier wegen ihrer besonderen Bedeutung durch einen großen Buchstaben symbolisiert sei. Da in Modellen mit Sättigungsgrenze ein zusätzlicher Parameter zu schätzen ist, lassen sie sich nicht linearisieren und sind somit intrinsisch nichtlinear.

Sättigungsgrenze

Elementare Wachstumsmodelle

Für unser Mobilfunkbeispiel kommt aufgrund der visuellen Betrachtung (Abbildung 10.3) wie auch aufgrund diffusionstheoretischer Überlegungen insbesondere ein S-förmiges Modell in Frage, also entweder ein Logistisches Modell oder ein Gompertz-Modell.

Logistisches Modell

Gompertz-Modell

Das Logistische Modell und das Gompertz-Modell unterscheiden sich primär dadurch, dass das Logistische Modell symmetrisch um den Wendepunkt verläuft, der somit bei $y_w = M/2$ liegt. Der Wendepunkt des Gompertz-Modells dagegen liegt etwas tiefer bei $y_w = M/e \approx 0,37M$. In Abbildung 10.5 erfolgt ein Vergleich dieser beiden Modelle.

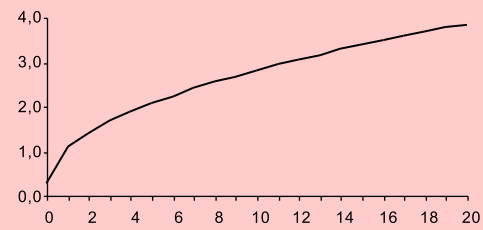
KQ-Methode

Auch intrinsisch nichtlineare Modelle lassen sich, wie lineare Modelle, mit Hilfe der Methode der kleinsten Quadrate (vgl. Kapitel 1) schätzen. Die zu schätzenden Modellparameter werden so bestimmt, dass die *Summe der quadrierten Abweichungen* (Sum of Squares, SS) zwischen den beobachteten Werten y_k und den geschätzten Werten $\hat{y}_k$ minimal wird:

$$SS = \sum_k (y_k - \hat{y}_k)^2 \rightarrow \min! \tag{10.4}$$

Iterative Berechnung

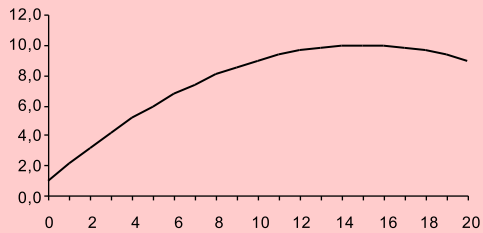
Das dabei entstehende mathematische Problem aber lässt sich nicht mehr, wie bei der linearen Regression, analytisch lösen, d. h. für die Berechnung der Schätzwerte lassen sich keine Formeln bilden. Vielmehr müssen die Schätzwerte numerisch mittels iterativer Algorithmen berechnet werden. In einem intelligenten Trial-and-Error-Verfahren werden die unbekannten Parameterwerte solange variiert, bis SS ein Minimum annimmt (was gleichbedeutend mit der Maximierung des Bestimmtheitsmaßes ist). Dies macht die Berechnung nicht nur sehr aufwändig, sondern wirft auch eine Reihe von Problemen auf.



Konkav

Potenz-Modell

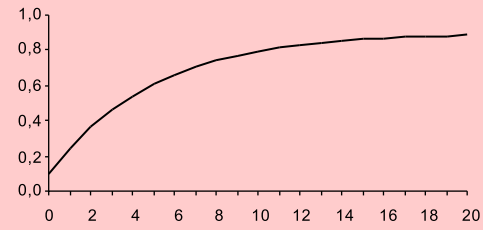
$$\hat{Y} = a + b \cdot t^c \qquad (0 < c < 1)$$



Konkav mit Abschwung

Quadratisches Modell

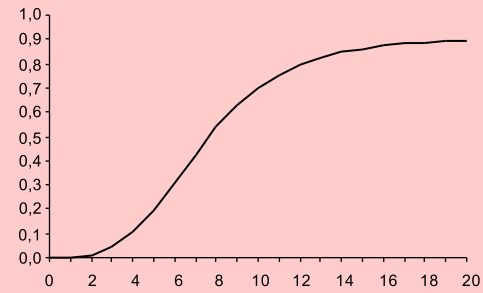
$$\hat{Y} = a + b \cdot t - c \cdot t^2 \qquad (c > 0)$$



Konkav mit Sättigungsgrenze

Exponentielles Modell

$$\hat{Y} = M - a \cdot e^{-b \cdot t} \qquad (0 < b < 1)$$



S-förmig mit Sättigungsgrenze

Logistisches Modell

$$\hat{Y} = \frac{M}{1 + e^{a - b \cdot t}} \qquad (b > 0)$$

Compertz-Modell

$$\hat{Y} = M \cdot e^{-a \cdot b^t} \qquad (0 < b < 1)$$

Abbildung 10.4: Elementare Wachstumsmodelle

	Logistisches Modell	Gompertz-Modell
Modellbeispiel	$\hat{Y} = \frac{M}{1+e^{a-b \cdot t}}$	$\hat{Y} = M \cdot e^{-a \cdot b^t}$
Ursprung	$y_0 = M/(1+e^a)$	$y_0 = M/e^a$
Wendepunkt	$y_w = M/2$ $t_w = a/b$	$y_w = M/e = M/2,7183 \approx 0,37M$ $t_w = ln(a)/-ln(b)$
Linearisierung für M gegeben	$ln(\frac{M}{Y} - 1) = a + b \cdot t$	$ln(ln(M) - ln(Y)) = ln(a) + ln(b) \cdot t$

Abbildung 10.5: Vergleich von Logistischem Modell und Gompertz-Modell

Konvergenzkriterium

Da der minimale Wert von SS nicht bekannt ist, weiß man auch nicht, ob man ihn gefunden hat. Deshalb definiert man ein Konvergenzkriterium für SS, mittels dessen entschieden wird, ob man ein Minimum gefunden hat und der Iterationsprozess beendet werden kann. Es besteht damit aber keine Gewähr, dass das so gefundene Minimum auch ein globales Minimum bildet oder ob es sich möglicherweise nur um ein lokales Minimum handelt.

Levenberg/Marquardt-Algorithmus

In SPSS wird standardmäßig der Levenberg/Marquardt-Algorithmus verwendet und das Konvergenzkriterium auf den Wert $SSCON = 1,0E-8$ gesetzt. d. h., der Iterationsprozess soll enden, wenn die Änderung von SS den Wert $SSCON$ unterschreitet.

Um einen iterativen Algorithmus zu starten, müssen Startwerte für die zu schätzenden Parameter des Modells vorgegeben werden. Ob und wie schnell eine optimale Lösung gefunden wird, hängt in entscheidendem Maße von diesen Startwerten ab, die der Untersucher festlegen muss.

Startwerte

Als Startwert für die Schätzung der Sättigungsgrenze M wählen wir einen Wert, der über dem größten beobachteten Wert liegt, z. B. $M = 100$. Damit lassen sich das Logistische Modell und das Gompertz-Modell linearisieren (vgl. Abbildung 10.5). Für das Logistische Modell ergibt sich:

$$ln(\frac{100}{Y} - 1) = a + b \cdot t \tag{10.5}$$

Mittels linearer Regression erhält man die Schätzwerte $a = 6,2$ und $b = 0,5$, die sich als Startwerte für die nichtlineare Schätzung des Logistischen Modells verwenden lassen.

Analog erhält man für das Gompertz-Modell mit $M = 100$ die Gleichung:

$$ln(ln(100) - ln(Y)) = ln(a) + ln(b) \cdot t \tag{10.6}$$

Die Regressionsanalyse liefert hierfür die Schätzwerte $a = 13,1$ und $b = 0,77$.

Die Ergebnisse der Modellschätzung, die man mit den obigen Startwerten erhält, sind in Abbildung 10.6 zusammengefasst.

Modell	Iterationen	a	b	M	R^2
Logistisch	8	5,960	0,509	92,0	0,982
Gompertz	8	23,023	0,750	103,4	0,987

Abbildung 10.6: Ergebnisse der Modellschätzung

Das Gompertz-Modell erzielt hier ein geringfügig größeres Bestimmtheitsmaß als das Logistische Modell. Mit den gefundenen Schätzwerten lautet es:

$$\hat{Y} = 103,4 \cdot e^{-23,023 \cdot 0,75^t} \tag{10.7}$$

Abbildung 10.7 zeigt den geschätzten und prognostizierten Verlauf des Gompertz-Modells im Streudiagramm der Daten.

Beim Logistischen Modell liegt der geschätzte Wert für die Sättigungsgrenze M unter dem für 2007 realisierten Wert von 97,151 Mio. Die Prognose mit diesem Modell würde also definitiv zu niedrige Werte liefern. Aber auch die durch das Gompertz-Modell geschätzte Sättigungsgrenze von M = 103,4 Mio. erscheint angesichts des empirischen Verlaufs recht niedrig. Eine für prognostische Zwecke bessere Modellschätzung lässt sich mittels gewichteter nichtlinearer Regression erhalten, indem man die quadratischen Abweichungen für jüngere Perioden stärker gewichtet als die für weiter zurückliegende Perioden.

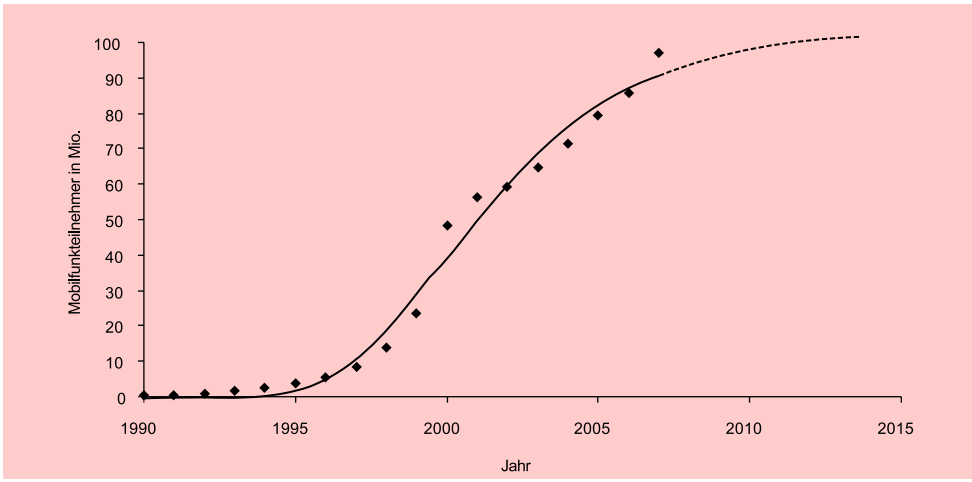


Abbildung 10.7: Gompertz-Modell: Geschätzter und prognostizierter Verlauf

10.3 Umsetzung mit SPSS

Die Durchführung von nichtlinearen Regressionsanalysen erfolgt in SPSS mit der Prozedur NLR, die über den Menüpunkt „Analysieren/ Regression/ Nichtlineare Regression“ aufgerufen werden kann. Es öffnet sich das in Abbildung 10.8 wiedergegebene Dialogfeld. Hier ist zunächst die abhängige Variable aus der Liste der Variablen in der Arbeitsdatei auszuwählen. Dies ist die Variable „Mobiles“, die die Zeitreihe der Mobilfunkteilnehmer enthält. In dem Feld „Modellformel“ ist das Modell zu formulieren. Zur Definition der Modellparameter und Eingabe der Startwerte ist das Feld „Parameter...“ anzuklicken.

Prozedur NLR

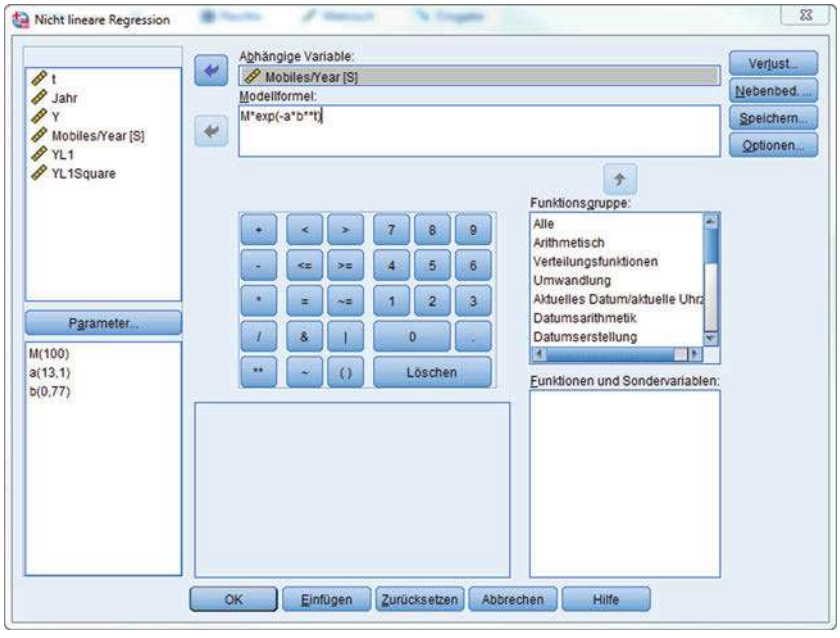


Abbildung 10.8: Dialogfeld „Nichtlineare Regression“

11 Strukturgleichungsanalyse

Strukturgleichungsmodelle werden in diesem Buch den „*Fortgeschrittenen Verfahren der multivariaten Analyse*“ zugeordnet und in diesem Kapitel nur in den Grundzügen behandelt. Eine ausführliche Darstellung zum Kapitel „Strukturgleichungsmodelle“ ist verfügbar in dem Buch *„Backhaus, K./Erichson, B.Weiber, R.: Fortgeschrittene Multivariate Analysemethoden, 3. Aufl., Berlin Heidelberg 2015.“*

11.1 Problemstellung

Theorie und Sachlogik als zentrale Ausgangspunkte der Modellbildung

Insbesondere im wissenschaftlichen Bereich besitzt die Bildung von Modellen einen zentralen Stellenwert zur Erklärung und Prognose unterschiedlicher Betrachtungsgegenstände. Aber auch in der Praxis werden häufig Modelle entwickelt, um komplexe reale Sachverhalte beschreiben (Erklärungsmodelle) und auch zukünftige Entwicklungen abschätzen (Prognosemodelle) zu können. Voraussetzung ist dabei jeweils, dass der Anwender über eine klare und in einer Theorie oder in der Sachlogik (z. B. durch Plausibilitätsüberlegungen) begründete Vorstellung über die Zusammenhänge des jeweils betrachteten Sachverhalts verfügt. Sollen diese Überlegungen einer empirischen Prüfung unterzogen werden, so müssen die betrachteten Größen auch empirisch messbar sein und die postulierten Zusammenhänge über geeignete mathematische Formulierungen in eine formale Struktur überführt werden. Liegen einfache Dependenzzusammenhänge bei direkt messbaren Variablen vor, so stellt die Regressionsanalyse die klassische Prüfmethodik dar (vgl. Kapitel 1 dieses Buches). Eine Einschränkung ergibt sich dabei allerdings dadurch, dass z. B. im Rahmen einer multiplen Regression nur eine abhängige Variable bei mehreren unabhängigen Variablen betrachtet wird. Es können somit nur einfache Dependenzstrukturen überprüft werden.

Strukturgleichungsmodelle sind Mehrgleichungssysteme

Bei komplexen Modellen existieren aber oftmals mehrere abhängige, d. h. zu erklärende Variablen, bei denen untereinander ebenfalls häufig kausale Zusammenhänge vermutet werden. In diesem Fall mündet die formale Abbildung in ein Modell mit mehreren Regressionsgleichungen *(Mehrgleichungssystem)*, und eine Modellprüfung erfordert die *simultane* Prüfung aller Beziehungen. In diesen Fällen sprechen wir von der *Strukturgleichungsanalyse*. Folgendes einfache Beispiel möge der Verdeutlichung dienen: Ein Hersteller von Margarine geht aufgrund von sachlogischen Überlegungen von folgenden Vermutungen aus:

1. Die „Attraktivität“ einer Margarine hat einen positiven Einfluss auf deren „Kaufabsicht“.

K. Backhaus, B. Erichson, W. Plinke, R. Weiber, *Multivariate Analysemethoden*,
https://doi.org/10.1007/978-3-662-56655-8_12

2. Die „Kaufabsicht“ von Margarine wird insbesondere durch den „Gesundheitsgrad“, die „Verwendungsbreite“ und das „Preisniveau“ einer Margarine bestimmt, wobei das „Preisniveau“ die „Kaufabsicht“ negativ beeinflusst.

3. Gleichzeitig haben die Erfahrungen des Margarineherstellers gezeigt, dass der wahrgenommene „Gesundheitsgrad“ und die „Verwendungsbreite“ einer Margarine auch die zentralen Bestimmungsgrößen für die „Attraktivität“ einer Margarine darstellen.

Pfaddiagramm

Obige Überlegungen lassen sich mit Hilfe eines *Pfaddiagramms* auch graphisch verdeutlichen (vgl. Abbildung 11.1), wobei aus den Vermutungen des Margarineherstellers folgende Zusammenhänge zwischen den Variablen deutlich werden:

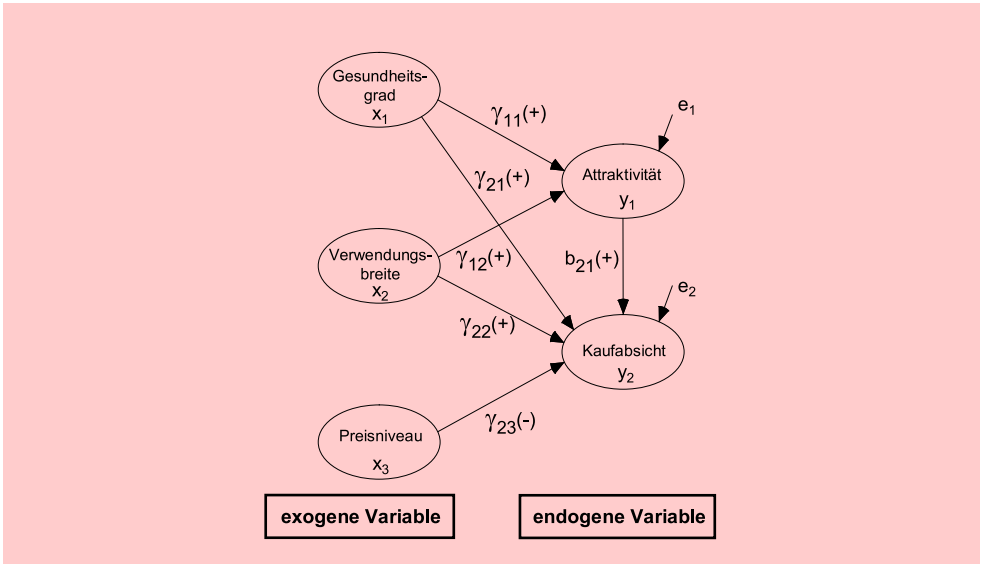


Abbildung 11.1: Strukturgleichungsmodell mit manifesten Variablen

Endogene und exogene Modellvariable

Die Variablen „Kaufabsicht“ und „Attraktivität“ stellen abhängige Größen dar, die durch die unabhängigen Größen „Gesundheitsgrad“, „Verwendungsbreite“ und „Preisniveau“ erklärt werden. Abhängige Variablen werden in Strukturgleichungsmodellen als *endogene Variablen* bezeichnet, um deutlich zu machen, dass sie durch die Größen im Modell erklärt werden. Gleichzeitig weisen endogene Variablen die Besonderheit auf, dass zwischen ihnen wiederum kausale Beziehungen bestehen können. So beeinflusst in unserem Beispiel die endogene Variable „Attraktivität“ auch die endogene Variable „Kaufabsicht“. Demgegenüber werden die erklärenden Größen als *exogene Variablen* bezeichnet, da sie quasi „von außen“ im Modell vorgegeben sind und durch das Modell selbst *nicht* erklärt werden. In Abhängigkeit der Messbarkeit der Variablen eines Strukturgleichungsmodells lassen sich folgende zwei Arten von Strukturgleichungsmodellen unterscheiden:

Strukturgleichungsmodelle mit manifesten Variablen

1. *Strukturgleichungsmodelle mit manifesten Variablen*:
   Hier wird unterstellt, dass alle Variablen des Modells *direkt* auf metrischem Skalenniveau messbar sind. Damit ergibt sich für das obige Beispiel folgendes

Gleichungssystem (ohne konstante Terme):

$$y_1 = \gamma_{11}x_1 + \gamma_{12}x_2 + e_1 \tag{11.1}$$
$$y_2 = b_{21}y_1 + \gamma_{21}x_1 + \gamma_{22}x_2 + \gamma_{23}x_3 + e_2 \tag{11.2}$$

Gleichungssystem

oder in *Matrixschreibweise*:

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ b_{21} & 0 \end{pmatrix} \cdot \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} + \begin{pmatrix} \gamma_{11} & \gamma_{12} & 0 \\ \gamma_{21} & \gamma_{22} & \gamma_{23} \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \end{pmatrix}$$

bzw. $y = \mathbf{B}y + \mathbf{\Gamma}x + e$

Dabei beinhalten die Matrizen $\mathbf{B}$ (Beta) und $\mathbf{\Gamma}$ (Gamma) die Regressionskoeffizienten des Gleichungssystems, während der Vektor e die Fehlervariablen der beiden Regressionsgleichungen umfasst.

2. *Strukturgleichungsmodelle mit latenten Variablen (Kausalanalyse)*:
   In diesem Fall stellen die Modellvariablen *latente Variable* bzw. hypothetische Konstrukte dar, die sich einer direkten Messbarkeit entziehen und deshalb über geeignete Indikatorvariablen gemessen werden müssen, die sich auf der empirischen Ebene direkt auf metrischem Skalenniveau erheben lassen. Damit erweitert sich das Strukturmodell aus Abbildung 11.1 um jeweils ein Messmodell für die latenten endogenen und die latenten exogenen Variablen, und es ergibt sich das in Abbildung 11.2 dargestellte Bild.

Strukturgleichungssystem mit latenten Variablen

Vollständiges Strukturgleichungsmodell

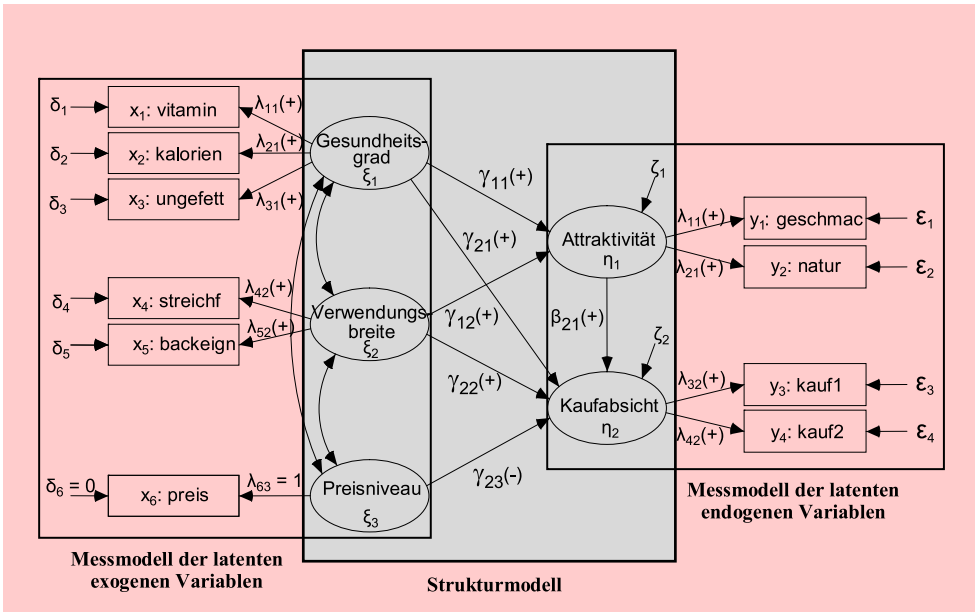


Abbildung 11.2: Strukturgleichungsmodell mit latenten Variablen

Die beiden Messmodelle folgen dabei jeweils dem Ansatz der Faktorenanalyse (vgl. Kapitel 7 in diesem Buch). Die Faktorenanalyse unterstellt, dass sich die empirischen Indikatorvariablen einer latenten Variablen durch hohe Korrelationen auszeichnen, wobei diese Korrelationen durch die jeweils betrachtete latente Variable verursacht werden. Um latente Variablen in einem solchen Strukturmodell auch formal zu kennzeichnen, werden latente exogene Variablen mit Ksi (ξ) und latente endogene Variablen mit Eta (η) bezeichnet. Demgegenüber wird für die empirischen Indikatorvariablen der latenten exogenen Größen die Bezeichnung x und für die der latenten endogenen Größen die Bezeichnung y verwendet.

Gleichungssystem eines vollständigen Strukturgleichungsmodells

Strukturmodelle mit latenten Variablen können ebenfalls über ein Gleichungssystem formal gefasst werden, wobei sich für das in Abbildung 11.2 dargestellte Pfaddiagramm folgendes Gleichungssystem ergibt (ohne konstante Terme):

$$\eta_1 = \gamma_{11}\xi_1 + \gamma_{12}\xi_2 + \zeta_1 \tag{11.3}$$
$$\eta_2 = \beta_{21}\eta_1 + \gamma_{21}\xi_1 + \gamma_{22}\xi_2 + \gamma_{23}\xi_3 + \zeta_2 \tag{11.4}$$

$$y_1 = \lambda_{11}\eta_1 + \epsilon_1 \tag{11.5}$$
$$y_2 = \lambda_{21}\eta_1 + \epsilon_2 \tag{11.6}$$
$$y_3 = \lambda_{32}\eta_2 + \epsilon_3 \tag{11.7}$$
$$y_4 = \lambda_{42}\eta_2 + \epsilon_4 \tag{11.8}$$

$$x_1 = \lambda_{11}\xi_1 + \delta_1 \tag{11.9}$$
$$x_2 = \lambda_{21}\xi_1 + \delta_2 \tag{11.10}$$
$$x_3 = \lambda_{31}\xi_2 + \delta_3 \tag{11.11}$$
$$x_4 = \lambda_{42}\xi_2 + \delta_4 \tag{11.12}$$
$$x_5 = \lambda_{52}\xi_2 + \delta_5 \tag{11.13}$$
$$x_6 = 1\xi_3 + 0 \tag{11.14}$$

oder in *Matrixschreibweise*:

$$\eta = B\eta + \Gamma\xi + \zeta \quad \text{Strukturgleichungsmodell} \tag{11.15}$$
$$y = \Lambda_y\eta + \epsilon \quad \text{Messmodell der latenten endogenen Variablen} \tag{11.16}$$
$$x = \Lambda_x\xi + \delta \quad \text{Messmodell der latenten exogenen Variablen} \tag{11.17}$$

Auch hier beinhalten die Matrizen $\mathbf{B}$ (Beta) und $\mathbf{\Gamma}$ (Gamma) die Regressions- bzw. Pfadkoeffizienten des Strukturgleichungsmodells und der Vektor ζ umfasst die Fehlervariablen der Strukturgleichungen. Die Matrizen $\mathbf{\Lambda_y}$ (Lambda-y) und $\mathbf{\Lambda_x}$ (Lambda-x) stellen Faktorladungsmatrizen dar, die die Beziehungsstärke (Faktorladungen) zwischen den Indikatorvariablen y bzw. x und den zugehörigen latenten Variablen η bzw. ξ zum Ausdruck bringen. Die Vektoren ϵ und δ beinhalten wiederum die Fehlerterme der jeweiligen Messgleichungen.

Ein Vergleich der beiden Arten von Strukturgleichungsmodellen macht deutlich, dass Strukturgleichungsmodelle mit manifesten Variablen als Spezialfall in dem Ansatz der Strukturgleichungsmodelle mit latenten Variablen enthalten sind. Soll ein Strukturgleichungsmodell nur mit manifesten Variablen gerechnet werden, so werden die manifesten Variablen zunächst als latente Variable interpretiert, denen jeweils *genau eine* Indikatorvariable zugeordnet wird. Anschließend werden a priori die jeweils zugehörige Faktorladung zwischen latenter Variable und Indikatorvariable mit 1 und der Fehlerterm der Indikatorvariablen mit 0 spezifiziert. Durch diese Festlegungen sind die Matrizengleichungen (11.16) und (11.17) vollständig bestimmt und die „latenten" Variablen als manifeste Variablen definiert. In Abbildung 11.2 wurde auf diese Weise die Variable „Preisniveau" als manifeste Variable definiert (vgl. auch Gleichung (11.14)). Damit stellen Strukturgleichungsmodelle mit latenten Variablen den allgemeinen Fall von Strukturgleichungsmodellen dar und bilden deshalb den Fokus dieses Kapitels.

Spezifizierung von latenten Variablen als manifeste Variable

11.2 Allgemeine Vorgehensweise

Die Vorgehensweise bei Strukturgleichungsmodellen mit latenten Variablen lässt sich durch folgende vereinfachte Schrittfolge beschreiben:

1. Modellformulierung
2. Parameterschätzungen
3. Beurteilung der Schätzergebnisse

Die *Modellformulierung* hat auf einer theoretischen Basis oder aufgrund fundierter sachlogischer Überlegungen zu erfolgen, die sich im Ergebnis graphisch mit Hilfe eines Pfaddiagramms abbilden lassen. Besondere Beachtung ist dabei der Formulierung der beiden Messmodelle zu schenken. Klassischerweise werden im Rahmen der Messmodelle solche Indikatorvariablen gewählt, die die betrachtete latente Variable jeweils in ihrer Gesamtheit reflektieren (sog. *reflektive Indikatoren*). Erfüllen die einzelnen Indikatorvariablen diese Annahme, so müssen sich zwischen den Indikatorvariablen, die eine bestimmte latente Variable messen sollen, hohe empirische Korrelationen ergeben. Weiterhin wird *unterstellt*, dass diese Korrelationen durch das Wirken der latenten Variablen verursacht werden, womit die Konstruktion der Messmodelle dem *allgemeinen Ansatz der Faktorenanalyse* (vgl. Kapitel 7 in diesem Buch) folgt. Die latenten Variablen stellen danach Faktoren dar, die als hypothetische Größen die Korrelationen zwischen den Indikatorvariablen begründen. Da in den beiden Messmodellen nur diejenigen Faktorladungen (= Korrelationen zwischen Indikatorvariablen und latenter Größe) geschätzt werden, die gemäß der theoretischen Überlegungen begründbar sind und alle anderen Faktorladungen a priori auf Null gesetzt werden, folgen die beiden Messmodelle jeweils dem Ansatz der *konfirmatorischen Faktorenanalyse* (vgl. Kapitel 12 in diesem Buch).

1. Schritt: Modellformulierung

Reflektive Messmodelle und reflektive Indikatoren

Konfirmatorische Faktorenanalyse

Bei der *Schätzung der Parameter* des Modells sind verschiedene Wege denkbar: Im einfachsten Fall kann ein Strukturgleichungsmodell in zwei Schritten sukzessive geschätzt werden. Dabei werden im ersten Schritt mit Hilfe von zwei Faktorenanalysen die Faktorladungen (Lambda-Koeffizienten) des exogenen sowie des endogenen Messmodells geschätzt und jeweils die Faktorwerte berechnet. Die Faktorwerte bilden Messwerte für die Faktoren und liefern damit „geschätzte Beobachtungswerte" für

2. Schritt: Parameterschätzungen

Zusammenspiel von Faktoren- und Regressionsanalyse

alle latenten Variablen bei allen befragten Personen. Mit Hilfe der Faktorwerte kann dann im zweiten Schritt eine Regressionsanalyse (vgl. Kapitel 1 in diesem Buch) mit den endogenen latenten Variablen als abhängige Größen und den exogenen latenten Variablen als unabhängige Größen gerechnet werden. Die Regressionsschätzung liefert über die Regressionskoeffizienten die Schätzung der Beziehungen im Strukturmodell (Gamma-Koeffizienten).

Varianzanalytischer Ansatz von PLS

Dieser Vorgehensweise folgt z. B. das *Partial Least Squares (PLS)-Verfahren*, das damit einen *varianzanalytischen Ansatz* verfolgt.

Kovarianzanalytischer Ansatz von AMOS

Demgegenüber nimmt die von SPSS angebotene Software AMOS (Analysis of Moment Structures) eine *simultane* Schätzung *aller Modellparameter* vor. Bei der Parameterschätzung folgt AMOS einem faktoranalytischen Ansatz, wobei mehrere Schätzalgorithmen zur Verfügung stehen. Die verschiedenen Schätzalgorithmen unterscheiden sich vor allem im Hinblick auf bestimmte Verteilungsannahmen über die Variablen und der Möglichkeit zur Berechnung von Inferenzstatistiken. Bei allen Schätzverfahren wird zunächst die empirische Korrelationsmatrix zwischen *allen* Indikatorvariablen berechnet (vgl. Abbildung 11.3).

Allgemeiner Aufbau der empirischen Korrelationsmatrix

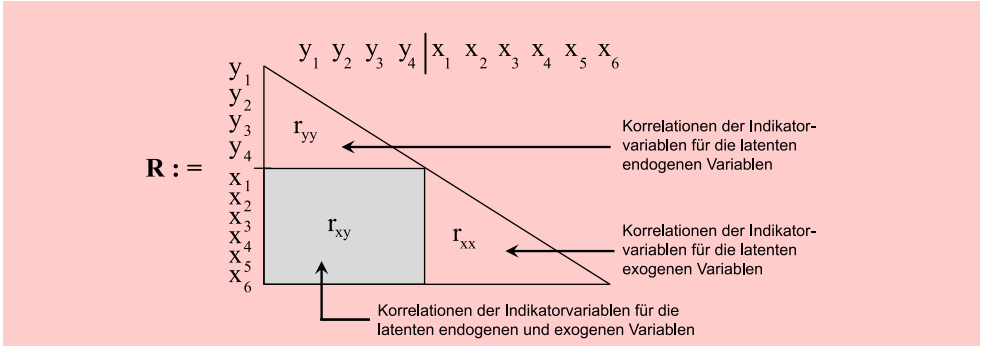


Abbildung 11.3: Aufbau der empirischen Korrelationsmatrix

Dabei bilden die Korrelationen zwischen den y-Indikatoren die Basis zur Schätzung der Parameter des Messmodells der latenten endogenen Variablen, während die Korrelationen zwischen den x-Indikatoren die Grundlage zur Schätzung der Parameter des Messmodells der latenten exogenen Variablen darstellen. Die Korrelationen zwischen den x- und y-Indikatorvariablen bilden die maßgebliche Grundlage zur Schätzung der Parameter des Strukturmodells. Die (simultane) Parameterschätzung wird von AMOS so vorgenommen, dass die mit Hilfe der geschätzten Parameter erzeugte modelltheoretische Korrelationsmatrix ($\sum$) eine möglichst gute Reproduktion der empirischen Korrelationsmatrix erbringt. Damit lautet das allgemeine Zielkriterium: $f(\mathbf{R} - \sum) \Rightarrow$ Min! Die verschiedenen Schätzfunktionen unterscheiden sich dabei in der unterschiedlichen Spezifizierung der Funktionsvorschrift f. Da der Ansatz von AMOS an Stelle der Korrelationsmatrix auch die Varianz-Kovarianzmatrix der Indikatorvariablen verwenden kann (auch Momente-Matrix genannt), wird diese Vorgehensweise häufig auch als *Kovarianzstrukturanalyse* bezeichnet.

Simulatane Schätzung aller Modellparameter

Kovarianzstrukturanalyse

3. Schritt: Beurteilung der Schätzergebnisse

Zur *Beurteilung der Schätzergebnisse* werden von AMOS unterschiedliche Gütekriterien angeboten, die sich einerseits auf die Beurteilung der Modellstruktur als Ganzes und andererseits auf die Beurteilung von Teilstrukturen des Modells beziehen.

11.3 Umsetzung mit SPSS

Die Analyse von Strukturgleichungsmodellen erfolgt im Rahmen von SPSS mit Hilfe des eigenständigen Programms AMOS. Dem Anwender wird durch das Untermodul „AMOS Graphics“ eine Grafikoberfläche mit diversen Zeichenwerkzeugen zur Verfügung gestellt, mit deren Hilfe sich das Pfaddiagramm des zu prüfenden Hypothesensystems inkl. der Indikatorvariablen leicht erstellen lässt (vgl. Abbildung 11.4).

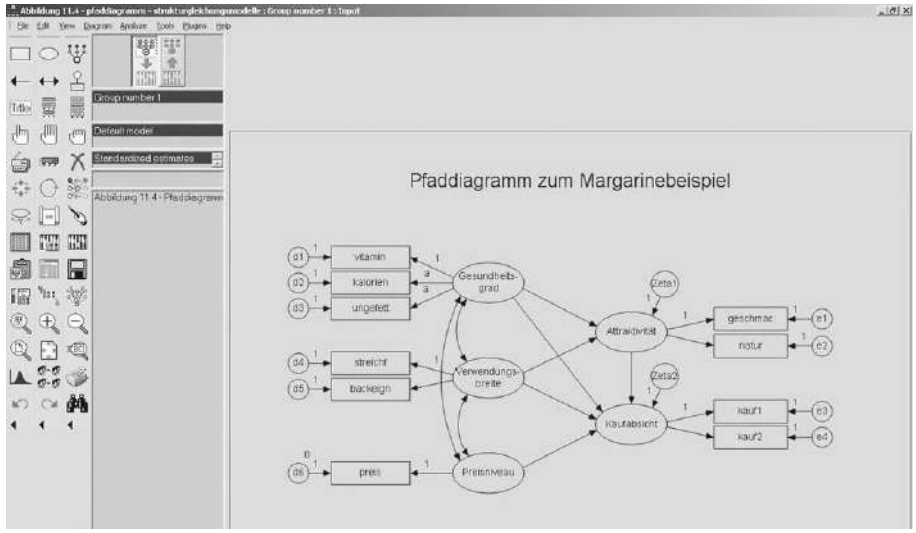


Abbildung 11.4: Grafikoberfläche und Toolbox von AMOS (Modul Graphics)

Erstellung eines Pfaddiagramms mit AMOS

Aus dem vom Anwender erstellten Pfaddiagramm ermittelt AMOS automatisch das Strukturgleichungssystem. Durch die Wahl des gewünschten Schätzalgorithmus können dann die Parameterschätzungen vorgenommen werden.

12 Konfirmatorische Faktorenanalyse

Die Konfirmatorische Faktorenanalyse wird in diesem Buch den „*Fortgeschrittenen Verfahren der multivariaten Analyse*“ zugeordnet und in diesem Kapitel nur in den Grundzügen behandelt. Eine ausführliche Darstellung zum Kapitel „Konfirmatorische Faktorenanalyse“ ist verfügbar in dem Buch *„Backhaus, K./Erichson, B.Weiber, R.: Fortgeschrittene Multivariate Analysemethoden, 3. Aufl., Berlin Heidelberg 2015.“*

12.1 Problemstellung

Operationalisierung hypothetischer Konstrukte (latenter Variable) als Kernanliegen

Reflektive Messmodelle als Basis

Bei vielen Problemstellungen sowohl in der Wissenschaft als auch in der Praxis sind Phänomene von Interesse, die sich einer direkten Messbarkeit auf der empirischen Ebene entziehen, weshalb sie auch als *hypothetische Konstrukte* oder *latente Variable* bezeichnet werden. Beispiele für solche hypothetischen Konstrukte sind z. B. Autorität, Angst, Emotion, Einstellung, Intelligenz, Involvement, Kaufabsicht, Loyalität, Macht, Motivation, Reputation, Stress, Vertrauen oder Zufriedenheit. Um hypothetische Konstrukte auch auf einer empirischen Ebene erfassen zu können, müssen diese über geeignete *Messmodelle* operationalisiert werden. Die konfirmatorische Faktorenanalyse greift dabei auf sog. *reflektive Messmodelle* mit empirisch direkt messbaren Variablen (sog. *Indikatorvariable*) zurück. Die Indikatorvariablen müssen dabei so definiert werden, dass ihre Messwerte jeweils beispielhafte Manifestierungen des betrachteten hypothetischen Konstruktes darstellen. Das Konstrukt muss sich durch die Indikatorvariablen jeweils in seiner Gesamtheit möglichst gut abbilden bzw. reflektieren lassen (sog. *reflektive Indikatoren*), d. h. die Indikatorvariablen müssen beliebig austauschbar sein. Es werden deshalb i. d. R. auch mehrere Indikatorvariablen für ein hypothetisches Konstrukt definiert, um auf diese Weise mögliche Verzerrungen einzelner Indikatorvariablen bei der Abbildung eines Konstruktes auszugleichen (sog. *Konzept multipler Items*).

gemeinsame Basis von explorativer und konfirmatorischer Faktorenanalyse

Ebenso wie die klassische bzw. explorative Faktorenanalyse basiert auch die konfirmatorische Faktorenanalyse auf dem *Fundamentaltheorem der Faktorenanalyse* (vgl. Kapitel 7 in diesem Buch). Dabei wird zunächst unterstellt, dass sich jeder Messwert einer Indikatorvariablen x_j als eine Linearkombination mehrerer (hypothetischer) Faktoren beschreiben lässt. Werden die Indikatorvariablen zuvor standardisiert, so lässt sich der Messwert einer standardisierten Indikatorvariablen (z_j) bei einem Objekt k

K. Backhaus, B. Erichson, W. Plinke, R. Weiber, *Multivariate Analysemethoden*,
https://doi.org/10.1007/978-3-662-56655-8_13

bei Existenz von Q Faktoren wie folgt berechnen:

$$z_{kj} = a_{j1}p_{k1} + a_{j2}p_{k2} + \ldots + a_{jQ}p_{kQ} \tag{12.1}$$

Die Gewichtungsgrößen a_{jq} werden als *Faktorladungen* bezeichnet, die den Zusammenhang zwischen einem Faktor q und einer Indikatorvariablen j zum Ausdruck bringen. Aus statistischer Sicht spiegeln die Faktorladungen die *Korrelation* zwischen einem Faktor und einer Indikatorvariablen wider. Um die Notation zu verkürzen, wird der Ausdruck (12.4) häufig auch in folgender Matrixschreibweise ausgedrückt, die die *Grundgleichung der Faktorenanalyse* darstellt:

Grundgleichung der Faktorenanalyse

$$\mathbf{Z} = \mathbf{AP} \tag{12.2}$$

Dabei ist $\mathbf{Z}$ die JxK-Matrix der standardisierten Ausgangsdaten, mit den J Indikatorvariablen in den Zeilen und den befragten K Personen in den Spalten. Die Faktorladungsmatrix $\mathbf{A}$ ist eine JxF-Matrix mit den J Indikatorvariablen in den Zeilen und den F Faktoren in den Spalten. Die Matrix $\mathbf{P}$ ist die FxK-Matrix der sog. *Faktorwerte*, die die (noch unbekannten) Messwerte der F Faktoren für alle K Personen beinhaltet.

Die Korrelationsmatrix ($\mathbf{R}$) der Indikatorvariablen lässt sich dann durch Multiplikation der standardisierten Ausgangsdatenmatrix ($\mathbf{Z}$) mit ihrer Transponierten ($\mathbf{Z}'$) berechnen, und es gilt:

Empirische Korrelationsmatrix bei standardisierten Variablen

$$\mathbf{R} = \frac{1}{K-1}\mathbf{ZZ}' \tag{12.3}$$

Wird für $\mathbf{Z}$ die Beziehung aus (12.2) verwendet, so lässt sich die empirische Korrelationsmatrix der Indikatorvariablen auch mit Hilfe der Faktorladungsmatrix wie folgt bestimmen:

$$\mathbf{R} = \frac{1}{K-1}(\mathbf{AP})(\mathbf{AP})' = \frac{1}{K-1}\mathbf{APP}'\mathbf{A}' = \mathbf{A}\left[\frac{1}{K-1}\mathbf{PP}'\right]\mathbf{A}' \tag{12.4}$$

Da der Ausdruck $[1/K-1\mathbf{PP}']$ die Korrelationsmatrix der (standardisierten) Faktorwerte darstellt, entspricht diese der Einheitsmatrix, wenn unterstellt wird, dass *keine* Korrelationen zwischen den Faktoren bestehen. In diesem Fall vereinfacht sich (12.4) zu:

$$\mathbf{R} = \mathbf{AA}' \tag{12.5}$$

Die Beziehungen (12.4) und (12.5) werden auch als das *Fundamentaltheorem der Faktorenanalyse* bezeichnet und beschreiben den Zusammenhang zwischen der empirischen Korrelationsmatrix ($\mathbf{R}$) und der Faktorladungsmatrix ($\mathbf{A}$).

Fundamentaltheorem der Faktorenanalyse

Ziel der konfirmatorischen Faktorenanalyse ist es, anhand einer vorgegebenen Beziehungsstruktur zwischen Indikatorvariablen und Faktoren die Faktorladungsmatrix ($\mathbf{A}$) so zu schätzen, dass sich mit ihrer Hilfe entsprechend Gleichung (12.5) die empirische Korrelationsmatrix der Indikatorvariablen möglichst gut reproduzieren lässt. Mit Hilfe der Parameterschätzungen können dann die postulierten Beziehungen überprüft werden.

	explorative Faktorenanalyse	konfirmatorische Faktorenanalyse
Zielsetzung	Entdeckung von Faktoren als ursächliche Größen für hoch korrelierende Variable	Prüfung der Beziehungen zwischen Indikatorvariablen und hypothetischen Größen
Zuordnung der Indikatorvariablen zu Faktoren	wird aufgrund statistischer Kriterien vom Verfahren vorgenommen	vom Anwender a priori vorgegeben
Anzahl der Faktoren	wird aufgrund statistischer Kriterien ermittelt	vom Anwender a priori vorgegeben
Schätzung der Faktorladungsmatrix	es wird eine vollständige Faktorladungsmatrix geschätzt	i.d.R. wird eine Einfachstruktur der Faktorladungsmatrix unterstellt
Interpretation der Faktoren	erfolgt a posteriori mit Hilfe der Faktorladungsmatrix	wird vom Anwender a priori vorgegeben

Abbildung 12.1: Explorative versus konfirmatorische Faktorenanalyse

Unterschiede zwischen explorativer und konfirmatorischer Faktorenanalyse

Die zentralen Unterschiede zwischen einer explorativen und einer konfirmatorischen Faktorenanalyse sind damit darin zu sehen, dass bei der explorativen Faktorenanalyse die Zuordnung von Ausgangsvariablen zu Faktoren sowie die Anzahl der zu extrahierenden Faktoren das *Ergebnis* der Faktorenanalyse ist, womit die Faktorenanalyse klassischerweise den *struktur-entdeckenden* Verfahren zugeordnet wird. Demgegenüber erfolgt bei einer konfirmatorischen Faktorenanalyse sowohl die Zuordnung der Indikatorvariablen zu Faktoren als auch die Festlegung der Anzahl der Faktoren sowie ihrer inhaltlichen Bedeutung a priori durch den Anwender aufgrund von theoretischen oder sachlogischen Überlegungen. Dementsprechend dient die konfirmatorische Faktorenanalyse allein der Prüfung vorab festgelegter Zusammenhänge und ist damit den *struktur-prüfenden* Verfahren der multivariaten Datenanalyse zuzuordnen.

Konfirmatorische Faktorenanalyse als Bestandteil von Strukturgleichungsmodellen

Gleichzeitig stellt die konfirmatorische Faktorenanalyse damit eine zentrale Basis zur Formulierung von Messmodellen bei der Analyse des Zusammenhangs zwischen latenten Variablen dar, der im Rahmen von *Strukturgleichungsmodellen* (vgl. Kapitel 11 in diesem Buch) analysiert wird.

12.2 Allgemeine Vorgehensweise

Die Vorgehensweise der konfirmatorischen Faktorenanalyse lässt sich durch vier zentrale Ablaufschritte beschreiben:

1. Modellformulierung
2. Modellspezifizierung
3. Parameterschätzungen
4. Beurteilung der Schätzergebnisse

1. Schritt: Modellformulierung

Im Rahmen der *Modellformulierung* hat der Anwender das Messmodell für das oder die betrachteten hypothetischen Konstrukte aufgrund von theoretischen oder fundierten sachlogischen Überlegungen festzulegen. Dabei ist insbesondere der Formulierung der Indikatorvariablen besondere Beachtung zu schenken, über die ein hypothetisches Konstrukt gemessen werden soll. Zur Konstruktion von Messmodellen stehen zwei grundsätzliche Wege zur Verfügung:

1. *Reflektive Messmodelle*:
   Reflektive Messmodelle
   Bei einem reflektiven Messmodell wird unterstellt, dass Veränderungen in den Messwerten der Indikatorvariablen (x_j) durch die latente Variable (ξ) kausal verursacht werden. Veränderungen des hypothetischen Konstruktes führen damit gleichermaßen auch zu Veränderungen bei den Indikatorvariablen. Damit können die Indikatorvariablen als „austauschbare Messungen“ der latenten Variablen interpretiert werden, weshalb sie auch als *reflektive Indikatoren* der latenten Variablen bezeichnet werden. Formal lässt sich diese Beziehung wie folgt fassen:

$$x_j = \lambda_j \xi + \delta_j \quad (j = 1,\ldots, n) \tag{12.6}$$

2. *Formative Messmodelle*:
   Formative Messmodelle
   Bei einem formativen Messmodell wird unterstellt, dass die Indikatorvariablen (x_j) Bestimmungsgrößen der betrachteten latenten Variablen (ξ) darstellen. Veränderungen bei einer Indikatorvariablen verursachen auch Veränderungen in der Ausprägung des hypothetischen Konstruktes. Damit stellen die Indikatorvariablen *keine* austauschbaren Messungen der latenten Variablen dar sondern sind „Bausteine“ oder unterschiedliche „Dimensionen“ der latenten Variablen, die diese formieren. Wir sprechen auch von *formativen Indikatoren*. Formal lässt sich diese Beziehung wie folgt fassen:

$$\xi = \lambda_1 x_1 + \lambda_2 x_2 + \ldots + \lambda_n x_n + \zeta \tag{12.7}$$

Der zentrale Unterschied zwischen reflektiven und formativen Messmodellen liegt damit in der Umkehrung der Beziehungsrichtung bzw. der unterstellten Kausalität zwischen den Indikatorvariablen und einer latenten Variablen. Formative Messmodelle folgen einem *regressionsanalytischen Ansatz* mit der Besonderheit, dass für die latente Variable als abhängige Größe der Regressionsbeziehung *keine* empirischen Messwerte verfügbar sind und diese in Relation zu anderen latenten Variablen geschätzt werden. So wird in dem in Abbildung 12.2 dargestellten Beispiel etwa davon ausgegangen, dass die „Verwendungsbreite“, der „Gesundheitsgrad“ und die „Geschmacksintensität“ einer Margarine jeweils einzelne Facetten bzw. zentrale Einflussgrößen des hypothetischen Konstruktes „Produktattraktivität“ bilden.

Demgegenüber folgen reflektive Messmodelle einem *faktoranalytischen Ansatz* und unterstellen, dass hohe Korrelationen zwischen den Indikatorvariablen bestehen, deren verursachende Größe die betrachtete latente Variable darstellt. In diesem Fall sind die Indikatorvariablen so zu definieren, dass sie jeweils für sich betrachtet ein Konstrukt in seiner Gesamtheit möglichst gut widerspiegeln. In Abbildung 12.2 wurde unterstellt, dass die Indikatorvariablen „Produktinteresse“, „Preisbereitschaft“ und „Produktkenntnis“ gute Reflektoren des Konstruktes „Produktattraktivität“ darstellen bzw. Veränderungen der Produktattraktivität auch zu Veränderungen bei diesen Indikatorvariablen führen.

Konfirmatorische Faktorenanalysen unterstellen reflektive Messmodelle

Von entscheidender Bedeutung ist es, dass die konfirmatorische Faktorenanalyse *reflektive Messmodelle* voraussetzt, was jedoch bei vielen praktischen und wissenschaftlichen Anwendungen nicht immer hinreichend beachtet wird.

Im Folgenden gehen wir als Beispiel für eine Modellformulierung davon aus, dass ein Margarinehersteller aufgrund seiner Erfahrungen vermutet, dass der „Gesundheitsgrad“ und die „Verwendungsbreite“ einer Margarine für den Umsatzerfolg seiner Margarine sehr wichtig sind. Er möchte deshalb wissen, wie diese beiden hypothetischen

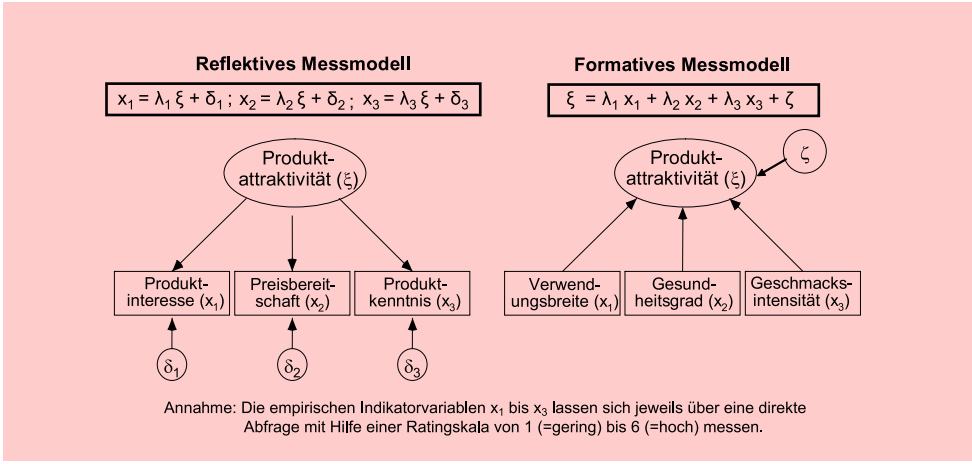


Abbildung 12.2: Reflektives versus formatives Messmodell

Größen von seinen Käufern eingeschätzt werden. Um Verzerrungen sowie den Einfluss zufälliger Effekte bei den Käufereinschätzungen zu reduzieren, erfragt er die Beurteilungen zu seiner Margarine bezüglich dieser beiden latenten Größen *nicht* direkt mit Hilfe einer Ratingskala. Stattdessen erhebt er sie anhand unterschiedlicher Indikatoren, von denen er glaubt, dass sie die beiden latenten Variablen jeweils gut widerspiegeln (reflektieren): Für den „Gesundheitsgrad“ sieht er die Größen „Vitamingehalt“, „Kaloriengehalt“ und „Anteil ungesättigter Fettsäuren“ jeweils als gute Indikatorvariablen an. Bei der „Verwendungsbreite“ glaubt er, dass die Indikatorvariablen „Streichfähigkeit“ sowie „Brat- und Backeignung“ diese latente Variable in geeigneter Form reflektieren können. Alle Indikatorvariablen werden im Rahmen einer Kundenbefragung auf einer Ratingskala von 1 (= gering) bis 6 (= hoch) erhoben. Der damit von dem Margarinehersteller sachlogisch formulierte Zusammenhang lässt sich graphisch mit Hilfe eines *Pfaddiagramms*, wie in Abbildung 12.3 dargestellt, verdeutlichen.

Umkehrung der Kausalbeziehungen bei reflektiven und formativen Messmodellen

Pfaddiagramm eines Messmodells für exogene latente Variable

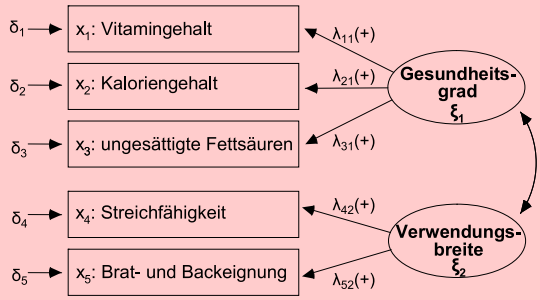


Abbildung 12.3: Pfaddiagramm der konfirmatorischen Faktorenanalyse für das Margarinebeispiel

Bei der Darstellung im Pfaddiagramm wird unterstellt, dass Wahrnehmungsveränderungen bei einem hypothetischen Konstrukt auch eine Ausstrahlung auf die jeweils definierten Indikatorvariablen besitzen, weshalb sich empirisch zwischen den Indika-

torvariablen Korrelationen ergeben. Außerdem wird eine eindeutige Zuordnung von Indikatorvariablen zu den Faktoren vorgenommen, deren Stärke im Rahmen der konfirmatorischen Faktorenanalyse über die Faktorladungen geschätzt wird.

2. Schritt: Modellspezifizierung

Die formale *Modellspezifizierung* lässt sich auf Basis des Pfaddiagramms vornehmen und mündet für das obige Margarinebeispiel in folgendes Gleichungssystem:

$$x_1 = \lambda_{11}\xi_1 + \delta_1 \tag{12.8}$$
$$x_2 = \lambda_{21}\xi_1 + \delta_2 \tag{12.9}$$
$$x_3 = \lambda_{31}\xi_2 + \delta_3 \tag{12.10}$$
$$x_4 = \lambda_{42}\xi_2 + \delta_4 \tag{12.11}$$
$$x_5 = \lambda_{52}\xi_2 + \delta_5 \tag{12.12}$$

Gleichungssystem

oder in *Matrixschreibweise*:

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} = \begin{pmatrix} \lambda_{11} & 0 \\ \lambda_{21} & 0 \\ \lambda_{31} & 0 \\ 0 & \lambda_{42} \\ 0 & \lambda_{52} \end{pmatrix} \cdot \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} + \begin{pmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \\ \delta_4 \\ \delta_5 \end{pmatrix}$$

bzw. $x = \mathbf{\Lambda_x}\xi + \delta$

Das Gleichungssystem macht deutlich, dass – entsprechend der Hypothesen des Margarineherstellers – diejenigen Faktorladungen a priori auf Null gesetzt wurden, bei denen der Hersteller *keinen* Zusammenhang zwischen einer Indikatorvariablen und einer latenten Variablen postulierte. Solche, bereits a priori festgelegten Parameterwerte, werden auch als „feste Parameter“ bezeichnet und gehen mit dem vorgegebenen Wert in die Lösung des Modells ein. Alle übrigen Parameter sind „freie Parameter“, deren Werte als unbekannt gelten und erst aus den empirischen Daten geschätzt werden.

feste versus freie Parameter

3. Schritt: Parameterschätzungen

Bei der *Schätzung der Parameter* des Modells nimmt die von SPSS angebotene Software AMOS (Analysis of Moment Structures) eine *simultane* Schätzung *aller* Modellparameter vor, wobei mehrere Schätzalgorithmen zur Verfügung stehen. Die verschiedenen Schätzalgorithmen unterscheiden sich vor allem im Hinblick auf bestimmte Verteilungsannahmen über die Variablen und der Möglichkeit zur Berechnung von Inferenzstatistiken. Die (simultane) Parameterschätzung wird von AMOS so vorgenommen, dass die mit Hilfe der geschätzten Parameter erzeugte modelltheoretische Korrelationsmatrix ($\sum$) eine möglichst gute Reproduktion der empirischen Korrelationsmatrix erbringt. Damit lautet das allgemeine Zielkriterium: $f(\mathbf{R} - \sum) \Rightarrow$ Min! Dabei errechnet sich die modelltheoretische Korrelationsmatrix nach dem Fundamentaltheorem der Faktorenanalyse gem. Gleichung (12.5). Die verschiedenen Schätzfunktionen unterscheiden sich in der unterschiedlichen Spezifizierung der Funktionsvorschrift f.

4. Schritt: Beurteilung der Schätzergebnisse

Zur *Beurteilung der Schätzergebnisse* werden von AMOS unterschiedliche Gütekriterien angeboten, mit deren Hilfe insbesondere die Validität und die Relibilität der Konstruktmessungen beurteilt werden können.

12.3 Umsetzung mit SPSS

Konfirmatorische Faktorenanalysen werden im Rahmen von SPSS mit Hilfe des eigenständigen Programms AMOS durchgeführt. Dem Anwender wird durch das Untermodul „AMOS Graphics“ eine Grafikoberfläche mit diversen Zeichenwerkzeugen zur Verfügung gestellt, mit deren Hilfe sich das Pfaddiagramm einer konfirmatorischen Faktorenanalyse leicht erstellen lässt (vgl. Abbildung 12.4).

Erstellung eines Pfaddiagramms mit AMOS

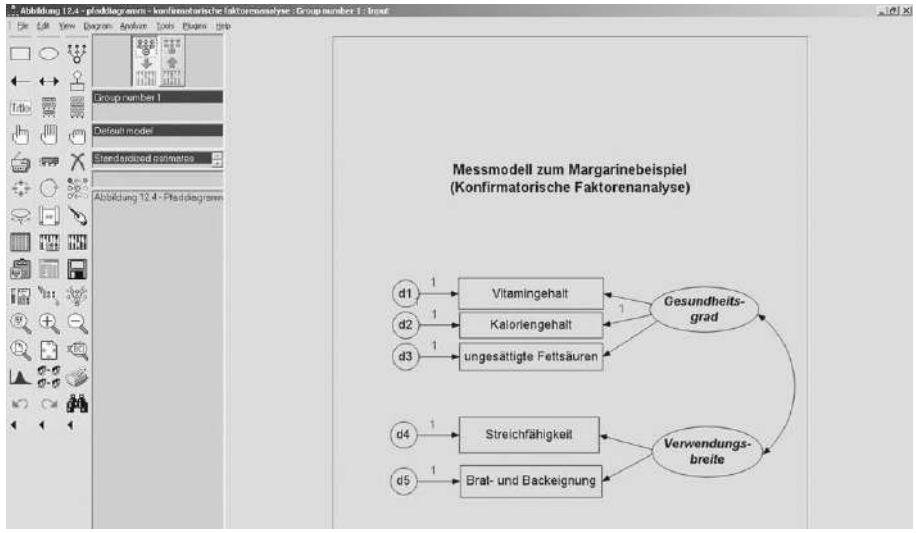


Abbildung 12.4: Grafikoberfläche und Toolbox von AMOS (Modul Graphics)

Aus dem vom Anwender erstellten Pfaddiagramm ermittelt AMOS automatisch das Gleichungssystem der konfirmatorischen Faktorenanalyse. Durch die Wahl des gewünschten Schätzalgorithmus können dann die Parameterschätzungen vorgenommen werden.

13 Auswahlbasierte Conjoint-Analyse

Die auswahlbasierte Conjoint-Analyse wird in diesem Buch den „*Fortgeschrittenen Verfahren der multivariaten Analyse*“ zugeordnet und in diesem Kapitel nur in den Grundzügen behandelt. Eine ausführliche Darstellung zum Kapitel „Auswahlbasierte Conjoint-Analyse“ ist verfügbar in dem Buch *„Backhaus, K./Erichson, B.Weiber, R.: Fortgeschrittene Multivariate Analysemethoden, 3. Aufl., Berlin Heidelberg 2015.“*

13.1 Problemstellung

Kaufprognose

Bei der in diesem Buch detailliert beschriebenen traditionellen Conjoint-Analyse wird die Vorziehenswürdigkeit eines Produkts im Vergleich zu einem oder mehreren anderen Produkten auf individueller Ebene erhoben. So kann beispielsweise prognostiziert werden, welche Margarine ein Befragungsteilnehmer favorisiert und deshalb wahrscheinlich in Zukunft kaufen wird. Dabei wird generell davon ausgegangen, dass für die Probanden alle untersuchten Alternativen potenziell kaufenswert erscheinen. Nur unter dieser Annahme kann die Präferenz für ein bestimmtes Produkt als Indikator für die Prognose von Auswahlentscheidungen bei einem realen Produktkauf herangezogen werden. Um künftige Kaufentscheidungen vorherzusagen, wird der geschätzte Gesamtnutzen für verschiedene hypothetische Alternativen durch Anwendung von Entscheidungsmodellen (probability-of-choice models) in Auswahlwahrscheinlichkeiten transformiert. Dabei kann beispielsweise das Max-Utility-Modell angewendet werden. Wie im Text zur traditionellen Conjoint-Analyse beschrieben, wird bei dem Max-Utility-Modell (dieses wird auch als First-Choice-Modell bezeichnet) das Produkt mit einer Wahlwahrscheinlichkeit von 1 ausgewählt, das den höchsten Nutzen aufweist. Durch Aggregation dieser Werte über alle Befragungsteilnehmer können schließlich Marktanteile geschätzt werden.

Max-Utility-Modell

First-Choice-Modell

Die Nutzung solcher Entscheidungsmodelle ist jedoch dann problematisch, wenn keine Informationen zu realen Kaufentscheidungsprozessen vorliegen. In diesem Fall muss der Anwender sich für eines der zur Verfügung stehenden Entscheidungsmodelle entscheiden. In Abhängigkeit von dieser Wahl ist jedoch mit unterschiedlichen Ergebnissen zu rechnen. Die Selektion eines spezifischen Modells hat demnach einen Einfluss auf den prognostizierten Marktanteil und damit auf die Vorteilhaftigkeit eines Produkts aus Unternehmenssicht.

Entscheidungsmodelle

K. Backhaus, B. Erichson, W. Plinke, R. Weiber, *Multivariate Analysemethoden*,
https://doi.org/10.1007/978-3-662-56655-8_14

Diesen Nachteil greifen die auswahlbasierten Verfahren (diese werden auch als Choice-Based Conjoint-Analysen bezeichnet) zur Präferenzmessung auf. Der Befragungsteilnehmer wird bei der auswahlbasierten Conjoint-Analyse gebeten, aus verschiedenen Choice-Sets jeweils die für ihn attraktivste Alternative auszuwählen. Ein Choice-Set besteht dabei aus verschiedenen hypothetischen Produkten, wobei oftmals drei bis vier Alternativen simultan präsentiert werden. Zudem ist es auch möglich, in jedes Choice-Set eine „Nicht-Auswahl-Option“ aufzunehmen. So kann sichergestellt werden, dass in die Nutzenschätzung lediglich als kaufenswert erachtete Produkte eingehen. Bei dieser Methode wird zudem vermutet, dass Auswahlentscheidungen das „natürliche“ Kaufverhalten der Konsumenten realistischer widerspiegeln als Bewertungen mit Hilfe einer Rangfolge oder durch Nutzung von Rating-Skalen.

Choice-Set

13.2 Allgemeine Vorgehensweise

Ziel der auswahlbasierten Conjoint-Analyse

Das Ziel beider Varianten der Conjoint-Analyse ist identisch. So kann ein Margarinehersteller nach Anwendung von traditionellen, aber auch bei den auswahlbasierten Varianten der Conjoint-Analyse Aussagen darüber treffen, auf welche Eigenschaften die Kunden einer Zielgruppe besonders achten und wie eine neue Margarinesorte konkret gestaltet werden sollte, um einen möglichst hohen Marktanteil zu erzielen. Auch die Abfolge der Ablaufschritte bei der Durchführung einer auswahlbasierten Conjoint-Analyse entspricht der traditionellen Vorgehensweise. Allerdings unterscheidet sich die konkrete inhaltliche Ausgestaltung zwischen den traditionellen und auswahlbasierten Conjoint-Analysen. Keinerlei Besonderheiten gibt es im ersten Schritt. Hier werden die zu untersuchenden Eigenschaften und Ausprägungen festgelegt. Unterschiede ergeben sich jedoch bei den anschließenden Schritten, die für die auswahlbasierten Conjoint-Analysen im Folgenden bis zu Schritt 4 (Teilnutzenschätzung) kurz skizziert werden.

Erhebungsdesign

Im zweiten Schritt wird ein Erhebungsdesign erstellt. Dabei wird generell die Profilmethode genutzt, d. h. es werden verschiedene hypothetische Produkte konstruiert, die durch sämtliche Eigenschaften beschrieben werden.[1] Im Gegensatz zur traditionellen Conjoint-Analyse, die meist auf einem orthogonalen Erhebungsdesign aufsetzt, werden bei der auswahlbasierten Variante die Stimuli meist zufällig konstruiert. Im dritten Schritt werden diese Alternativen den Befragungsteilnehmern im Rahmen verschiedener Choice-Sets zur Bewertung vorgelegt. Auch die Zuordnung der Stimuli zu einem Choice-Set erfolgt meist durch eine Zufallsauswahl.

Hypothetische Kaufsituation

In Abbildung 13.1 ist ein solches Choice-Set für eine hypothetische Kaufsituation, in der sich Margarinekäufer befinden, dargestellt. In diesem Fall entscheidet sich der Befragungsteilnehmer für Margarine I.

Dichotome Urteile

Bei der in Abbildung 13.1 dargestellten Auswahlentscheidung handelt es sich um ein dichotomes Urteil der Befragungsteilnehmer, d. h. eine Alternative wird aus einem Set verschiedener Alternativen ausgewählt oder nicht. Der Vorteil einer solchen Vorgehensweise ist ein vergleichsweise geringer kognitiver Aufwand bei der Bewertung der Alternativen, der durch die verstärkte Anwendung von Urteilsheuristiken zustande kommt. Es wird demnach vermutet, dass dem Probanden die Selektion der „besten“ Alternative aus einem Set möglicher Produkte relativ leicht fällt. Allerdings ist die Informationseffizienz dichotomer Bewertungen, auf denen alle auswahlbasierten Conjoint-Analysen beruhen, weitaus niedriger als bei den von den traditionellen

[1] Bei einigen auswahlbasierten Verfahren können auch sog. „partial profiles“ untersucht werden, d. h. die präsentierten Produktalternativen bestehen aus einer Teilmenge der zu untersuchenden Eigenschaften.

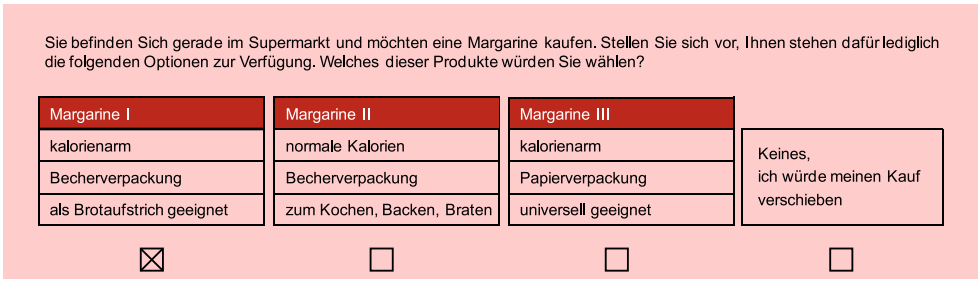


Abbildung 13.1: Darstellung eines Choice-Sets

Conjoint-Analysen genutzten Rangreihungen aller verfügbarer Alternativen bzw. bei der Bewertung der Stimuli auf einer Rating-Skala. Deshalb sind bei auswahlbasierten Conjoint-Analysen tendenziell mehr Einschätzungen notwendig. Um die Befragungsteilnehmer durch eine zu große Anzahl an Choice-Sets nicht zu überfordern, wird jedem Probanden lediglich eine Teilmenge an möglichen Produkten und Choice-Sets zur Einschätzung vorgelegt. Bei der praktischen Anwendung variiert die Anzahl der Choice-Sets, die ein Befragungsteilnehmer bewertet, meist zwischen acht und 20.

Anzahl der Einschätzungen

Eine solche Aufteilung der möglichen Choice-Sets über alle Befragungsteilnehmer hat zur Folge, dass keine (im engen Sinne) individuellen Nutzenfunktionen geschätzt werden können. Um stabile Schätzergebnisse zu erzielen, muss aufgrund dieser Aufteilung zudem die Zahl der Befragungsteilnehmer ausreichend groß sein. Schließlich können im Vergleich zur traditionellen Conjoint-Analyse tendenziell weniger Eigenschaften simultan untersucht werden. Dies gilt insbesondere, weil aufgrund der wiederholten Präsentation von Choice-Sets häufiger mit der Anwendung von Entscheidungsheuristiken bei der Auswahlentscheidung zu rechnen ist. Um den kognitiven Aufwand der wiederholten Auswahlentscheidung zu mindern, könnte ein Befragungsteilnehmer beispielsweise lediglich eine Teilmenge der zur Verfügung gestellten Informationen (z. B. nur die aus Sicht des Kunden wichtigsten Merkmale) zur Bewertung der Alternativen nutzen. Bei einem Vergleich der Ergebnisse zwischen traditionellen und auswahlbasierten Varianten der Conjoint-Analyse ist deshalb damit zu rechnen, dass bei Anwendung auswahlbasierter Präferenzmessmethoden den „wichtigen“ Eigenschaften ein größeres, den anderen Merkmalen dagegen ein kleineres Gewicht zugewiesen wird. Bei Eigenschaften wie „Preis“ und visuell wahrnehmbaren Merkmalen wie „Marke“ oder „Verpackungsgröße“ ist deshalb mit vergleichsweise hohen Bedeutungsgewichten zu rechnen.

In Abbildung 13.2 werden die grundlegenden Merkmale von auswahlbasierten Conjoint-Analysen im Vergleich zu traditionellen Conjoint-Analysen dargestellt.

grundlegende Merkmale der auswahlbasierten Conjoint-Analyse

Die Auswahl einer spezifischen Variante der Conjoint-Analyse hat immer auch einen Einfluss auf die zu erwartenden Schätzergebnisse. In der Praxis wird deshalb empfohlen, die Variante der Conjoint-Analyse ausgehend von den vermuteten Bewertungsstrategien in realen Kaufsituationen auszuwählen. Traditionelle Varianten der Conjoint-Analyse erscheinen dann vorteilhaft, wenn Produkte in high-involvement Kategorien untersucht werden sollen, bei denen auch innerhalb des realen Kaufentscheidungsprozesses mit einem sorgfältigen Abwägen der verschiedenen Produktmerkmale zu rechnen ist. Ein Beispiel dafür sind Fernseher oder Laptops. Andererseits ist bei low-involvement Produktkategorien auch in der Realität eher mit vereinfachten Kaufentscheidungsprozessen zu rechnen. So könnte ein Konsument sich mehr oder

Auswahlbasierte Conjoint-Analysen
vermutete höhere Realitätsnähe
Integration einer Nicht-Kauf-Alternative
„direkte“ Prognose von Marktanteilen
Analyse einer geringeren Anzahl von Eigenschaften
geringe Informationseffizienz dichotomer Urteile
tendenziell stärkere Anwendung vereinfachter Bewertungsstrategien
Traditionelle Conjoint-Analysen
Bewertung verschiedener Produktalternativen
echte Individualanalysen möglich, deshalb auch bei kleinen Stichprobengrößen geeignet
tendenziell stärkere Anwendung bei high-involvement Gütern

Abbildung 13.2: Ausgewählte Vor- und Nachteile auswahlbasierter und traditioneller Conjoint-Analysen

weniger „spontan“ für eine Margarine der Marke „Flora“ entscheiden, d. h. allein aufgrund vergleichsweise weniger Schlüsselmerkmale wie „Marke“, „Verpackungsgröße“ und „Preis“ wird eine Kaufentscheidung getroffen. Auswahlbasierte Verfahren sind demnach insbesondere für die Prognose von Kaufentscheidungen geeignet, bei denen die Auswahl eines Produkts mehr oder weniger „unmittelbar“ erfolgt und diese keine langfristigen Auswirkungen auf den Konsumenten hat.

Direkte Wahlentscheidungen

Die auswahlbasierten Varianten der Conjoint-Analyse haben in den vergangenen Jahren eine weite Verbreitung gefunden, was vor allem dem vermuteten höheren Realitätsgehalt von Auswahlentscheidungen und dem Vorhandensein einfach zu bedienender Software zur Erstellung und Durchführung entsprechender Studien zugeschrieben wird. Teilweise wird davon ausgegangen, dass inzwischen mehr auswahlbasierte als traditionelle Conjoint-Analysen durchgeführt werden.

Teilnutzenschätzung

Zur Schätzung der Teilnutzen (Schritt 4) auf Basis der erhobenen Auswahlentscheidungen ist die Anwendung der folgenden drei Ansätze üblich:

- aggregierte Choice Analyse,
- Latent Class Analyse und
- Hierarchical Bayes Schätzung.

multinomiales Logit-Modell

Da, wie beschrieben, die Informationseffizienz von Auswahlentscheidungen aufgrund der Nutzung dichotomer Urteile gering ist, sind vergleichsweise viele Einschätzungen notwendig, bevor die Teilnutzenwerte geschätzt werden können. Deshalb wird bei der aggregierten auswahlbasierten Conjoint-Analyse die Vielzahl der zu bewertenden Alternativen über alle Befragungsteilnehmer aufgeteilt. Durch Nutzung eines multinomialen Logit-Modells werden danach die Schätzwerte auf aggregierter Ebene bestimmt. Ergebnis ist eine Nutzenfunktion, die die Auswahlentscheidungen aller Befragungsteilnehmer repräsentiert. Grundannahme dieser Vorgehensweise ist, dass die Befragungsteilnehmer alle über vergleichsweise ähnliche Präferenzen verfügen. Entgegen dieser Anwendungsvoraussetzung ist jedoch oftmals damit zu rechnen, dass verschiedene Zielgruppen am Markt existieren, die durch unterschiedliche Bedürfnisse gekennzeichnet sind. Um das Problem aggregierter Nutzenschätzungen abzumildern,

kann eine Latent Class Analyse bzw. ein Hierarchical Bayes Ansatz zur Schätzung der Teilnutzenwerte angewendet werden.

Latent-Class Analyse

Mit Hilfe der Latent-Class Analyse ist es möglich, auf Basis der Auswahlentscheidungen simultan zur Nutzenschätzung verschiedene Zielgruppen zu bestimmen und die jeweiligen segmentspezifische Nutzenfunktionen zu ermitteln. Ergebnis sind demnach weiterhin nicht individuelle Nutzenwerte, sondern die Präferenzen für eine spezifische Zielgruppe. Anwendungsvoraussetzung sind also möglichst homogene Zielgruppen innerhalb der Stichprobe. Eine Näherung an individuelle Nutzenfunktionen kann erzielt werden, indem die Wahrscheinlichkeiten der Zugehörigkeit zu verschiedenen Zielgruppen berücksichtigt wird.

Hierarchical Bayes

Mit Hilfe der Hierarchical Bayes Schätzung ist es schließlich möglich, quasi „individuelle“ Nutzenfunktionen zu ermitteln. Dabei gehen allerdings Informationen über die Verteilung der Teilnutzen aller Befragungsteilnehmer als a priori Informationen in die befragungsteilnehmerspezifische Schätzung ein, sodass insgesamt weniger Informationen auf individueller Ebene benötigt werden und trotzdem für jeden Befragungsteilnehmer Nutzenwerte geschätzt werden können. Dabei können „Punktschätzungen“ für einzelne Teilnutzen, aber auch Schätzungen über die Wahrscheinlichkeit bestimmter Schätzparameter ermittelt werden. Auswahlbasierte Conjoint-Analysen beruhen in der praktischen Anwendung inzwischen meist auf dem Hierarchical Bayes Ansatz.

13.3 Umsetzung mit SPSS, Excel und Spezialsoftware

Auswahlbasierte Conjoint-Analysen (Choice-Based-Conjoint-Analysis; CBCA) können nicht ohne weiteres mit dem Programmpaket SPSS durchgeführt werden, weshalb in der Anwendungspraxis häufig auf Spezialsoftware zurückgegriffen wird. Eine weite Verbreitung hat dabei das Programmpaket der Sawtooth Software, Inc. (SSI) gefunden, das speziell für die Erstellung von conjointanalytischen Designs, deren Implementierung in Versuchsplänen (Programm-System SSI Web) und die Auswertung der erhobenen Daten (Programm-System SSI SMRT) konzipiert wurde. Abhängig von der erworbenen Ausstattung des Programmpakets hat der Nutzer die Wahl zwischen Latent-Class und Hierarchical Bayes basierten Choice-Based Conjoint-Analysen.

SSI Web wurde vor allem zur Erstellung von Fragebögen für Internet-Befragungen entwickelt und ermöglicht die Konstruktion des Erhebungsdesigns für eine CBCA incl. der Definition der Choice Sets. Die eigentliche Datenauswertung mit Schätzung der Teilnutzenwerte erfolgt dann über SSI SMRT.

Alternativ zur Verwendung einer Spezialsoftware für die CBCA (wie z. B. SSI) können die Teilnutzenwerte einer CBCA aber auch mit Hilfe von Excel oder der Prozedur/„Cox-Regression“ in SPSS geschätzt werden. In MS Excel ist die Schätzung mit Hilfe des sog. Solver möglich. In SPSS ist Prozedur COXREG unter dem Menüpunkt „*Analysieren / Überleben / Cox-Regression*“ verfügbar. Die mittels Cox-Regression geschätzten Regressionskoeffizienten entsprechen den Teilnutzenwerten.

In der dritten Auflage unseres Lehrbuchs „Fortgeschrittene Multivariate Analysemethoden“ wird ausführlich die Durchführung einer CBCA mit Hilfe von MS Excel, IBM SPSS und der SSI-Software gezeigt. Für das in diesem Buch verwendete Fallbeispiel führen alle drei Vorgehensweisen zu den gleichen Schätzergebnissen der Teilnutzenwerte. Zusätzlich zeigen wir im Anhang zu diesem Kapitel auch alternative Wege zur *Konstruktion von Erhebungsdesigns* für die CBCA auf, sodass auch hier nicht zwingend auf eine Spezialsoftware zurückgegriffen werden muss.

14 Neuronale Netze

Neuronale Netze werden in diesem Buch den „*Fortgeschrittenen Verfahren der multivariaten Analyse*“ zugeordnet und in diesem Kapitel nur in den Grundzügen behandelt. Eine ausführliche Darstellung zum Kapitel „Neuronale Netze“ ist verfügbar in dem Buch *„Backhaus, K./Erichson, B. Weiber, R.: Fortgeschrittene Multivariate Analysemethoden, 3. Aufl., Berlin Heidelberg 2015.“*

14.1 Problemstellung

In der Realität sind die Wirkungsbeziehungen zwischen Variablen häufig sehr komplex, wobei sich die Komplexität einerseits in einer großen Anzahl von miteinander verknüpften Einflussfaktoren äußert, andererseits darin, dass die Beziehungen zwischen den Variablen häufig nicht-linear sind. Auch kann der Anwender in vielen Fällen *keine* begründeten Hypothesen über die Art der Zusammenhänge aufstellen. In solchen Fällen sind sog. Künstliche Neuronale Netze (KNN) von großem Nutzen, da der Anwender bei dieser Gruppe von Analyseverfahren nicht zwingenderweise eine Vermutung über den Zusammenhang zwischen Variablen treffen muss. Das bedeutet, dass weder eine kausale Verknüpfung zwischen Variablen postuliert noch die Verknüpfung zwingend als linear unterstellt werden muss. Außerdem können Neuronale Netze auch Variablen mit unterschiedlichem Skalenf verarbeiten. Durch KNN werden die Zusammenhänge zwischen Variablen selbständig durch einen Lernprozess ermittelt und sie können dabei eine Vielzahl von Variablen berücksichtigen.

Nicht-Linearität

Grundsätzlich können mit Neuronalen Netzen klassische multivariate Analysemethoden substituiert werden, soweit großzahlige Untersuchungen vorliegen. Es existieren zahlreiche Typen von Neuronalen Netzen, die ein sehr breites Einsatzspektrum, z. B. Prognosen (vgl. Regressionsanalyse) oder Zuordnungen zu bestehenden Gruppen (vgl. Diskriminanzanalyse), abdecken. Der Einsatz von Neuronalen Netzen bietet sich immer dann an, wenn die Wirkungszusammenhänge zwischen den einzelnen Einflussgrößen nicht unbedingt aufgedeckt werden müssen, sondern durch „Trainieren“ des Netzes Lernprozesse erzeugt werden, die Lösungsansätze für die jeweilige Fragestellung bieten. Ausgewählte Fragestellungen, die durch den Einsatz von Neuronalen Netzen beantwortet werden können, zeigt Abbildung 14.1.

Substitution traditioneller Verfahren

Anwendungsbeispiele

K. Backhaus, B. Erichson, W. Plinke, R. Weiber, *Multivariate Analysemethoden*,
https://doi.org/10.1007/978-3-662-56655-8_15

Fragestellung	Vorgehensweise	Problemtyp
Wie verhält sich der Aktienkurs bei Variation verschiedener Einflussfaktoren?	Es werden die Einflussfaktoren auf den Aktienkurs während einer bestimmten Periode und der korrespondierende Aktienkurs erhoben. Anschließend wird das KNN auf neue Situationen, also zur kurzfristigen Prognose von Aktienkursen, angewendet.	Prognose
Wie hoch ist der Umsatz eines Unternehmens bei verschiedenen Szenarien?	Es wird der Umsatz eines Unternehmens in vergangenen Umweltsituationen untersucht. Die Umweltsituationen werden durch eine Reihe von Merkmalen beschrieben. Das KNN berechnet den Umsatz für neue Umweltsituationen.	Prognose
Soll ein Bankkredit gewährt werden?	Ausgangsbasis ist ein Datensatz, der kreditwürdige und nicht-kreditwürdige Kunden sowie deren soziodemographischen und ökonomischen Angaben umfasst. Das KNN ordnet Kunden bei der Beantragung eines Kredites einer der beiden Gruppen zu.	Klassifizierung (Zuordnung)
Wie ist die Bonität anhand von Jahresabschlüssen zu beurteilen?	Mit Hilfe von Kennzahlen aus Jahresabschlüssen vergangener Perioden werden die betrachteten Unternehmen in verschiedene Insolvenzklassen eingeteilt.	Klassifizierung (Zuordnung)
Wie lassen sich die Käufer in verschiedene Gruppen einteilen?	Käufer werden über soziodemographische und ökonomische Merkmale definiert. Das KNN generiert eine Ausgabe, die über die Ähnlichkeit zwischen den verschiedenen Käufern Aufschluss gibt und als Grundlage für die Bildung von verschiedenen Käufergruppen dient.	Klassifizierung (Gruppenbildung)

Abbildung 14.1: Anwendungsbeispiele

14.2 Allgemeine Vorgehensweise

Biologische Lernprozesse

Das Konstruktionsprinzip von KNN beruht auf dem Vorbild der Abläufe im Nervensystem von Menschen und Tieren und versucht, „biologisches Lernen“ über geeignete mathematische Operationen nachzuvollziehen, wodurch sehr gute Ergebnisse erzeugt werden können. Die Grundstruktur von KNN lässt sich anhand von Abbildung 14.2 verdeutlichen.

Jedes KNN besteht aus einer Eingabeschicht (Input-Layer), einer oder mehreren verdeckten Schichten (Hidden-Layer) und einer Ausgabeschicht (Output-Layer).

In Abbildung 14.2 ist beispielhaft ein Netz mit zwei verdeckten Schicht aufgezeigt. Dabei besteht die Eingabeschicht aus vier Neuronen ($1 - 4$), die erste verdeckte Schicht aus drei Neuronen ($5 - 7$), die zweite verdeckte Schicht aus zwei Neuronen ($8, 9$) und die Ausgabeschicht aus drei Neuronen ($10 - 12$). Auf der Eingabeschicht werden alle (empirisch erhobenen) Variablen als Eingabeneuronen abgebildet, die dann eine Aktivierung der Neuronen auf den beiden verdeckten Schichten bewirken. Auf der Ausgabeschicht dienen diejenigen (empirisch erhobenen) Variablen als Referenzgrößen für die Ausgabeneuronen, die den Output (Zielvariable) des Neuronalen Netzes widerspiegeln. Die Besonderheit von Neuronalen Netzen ist nun darin zu sehen, dass für die Verarbeitung der Eingangsvariablen auf den verdeckten Schichten *keine Vorgaben* (z. B. durch festgelegte Beziehungen oder Kausalhypothesen) gemacht werden, sondern das Netz durch einen Lernprozess den Aktivierungsgrad der einzelnen Neuronen auf den verdeckten Schichten jeweils so bestimmt, dass die Ausgabeneuronen die empirisch erhobenen Ergebnisgrößen möglichst genau abbilden können. Das Grundprinzip der Informationsverarbeitung in den Neuronen der verdeckten Schicht lässt sich wie folgt verdeutlichen: Zunächst werden die auf ein Neuron j treffenden

Eingabeschicht

Ausgabeschicht

Verdeckte Schichten

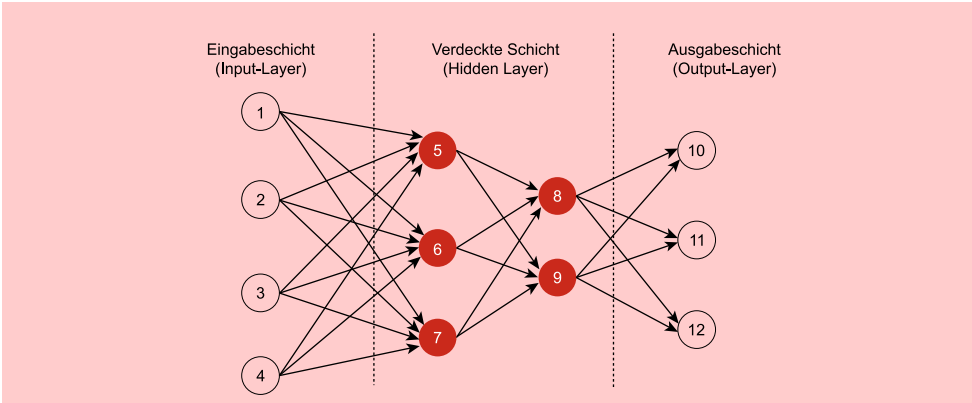


Abbildung 14.2: Grundstruktur eines (dreischichtigen) neuronalen Netzes

Signale zu einem Nettoeingabewert (net_j) für das Neuron verdichtet. Innerhalb des Neurons wird dann dieser Nettoeingabewert nach Maßgabe einer Aktivierungsfunktion verarbeitet, die im Ergebnis den Aktivierungsgrad des Neurons bestimmt. Die Verdichtung der auf ein Neuron treffenden Signale zu einem Nettoeingabewert erfolgt nach Maßgabe der sog. Propagierungsfunktion, die im einfachsten Fall als Summenfunktion definiert ist und den Nettoeingabewert aus der Summe der gewichteten Eingabesignale berechnet. Für die Summenfunktion gilt:

Propagierungsfunktion

$$net_j = \sum w_{ij}\, o_i \tag{14.1}$$

Durch die sog. Aktivierungsfunktion wird aus dem Nettoeingabewert dann der Aktivierungszustand des Neurons bestimmt, wobei im einfachsten Fall der Aktivierungszustand des Neurons ($a(net)$) zweiwertig – im Sinne von „aktiviert (1)“ / „nicht aktiviert (1)“ – ist. Für eine solche „einfache Schwellenwertfunktion“ gilt:

Aktivierungsfunktion

$$a\,(net_j) = \begin{cases} 1 & \text{falls } net_j > \Theta_j \\ 0 & \text{sonst} \end{cases} \tag{14.2}$$

Sowohl die Gewichte (w_{ij}) der Propagierungsfunktion als auch der Parameter Θ der Aktivierungsfunktion werden durch den Lernprozess des Netzes solange verändert und angepasst, bis die Ausgabeneuronen die empirisch gemessenen Ergebnisse (Zielvariable) möglichst gut abbilden können. Bereits hier wird deutlich, dass Neuronale Netze hohe Fallzahlen erfordern, mit denen das Netz trainiert werden muss.

Die obigen Zusammenhänge lassen sich an folgendem vereinfachten Beispiel verdeutlichen: Ein Margarinehersteller geht davon aus, dass die Kaufentscheidung für Margarine maßgeblich durch das Geschlecht, den Preis und das Gesundheitsbewusstsein des Käufers (gemessen auf einer Skala von 0 = kein Gesundheitsbewusstsein bis 10 = sehr hohes Gesundheitsbewusstsein) bestimmt wird. Für eine große Zahl von Probanden werden deshalb diese Kriterien erhoben und gleichzeitig wird festgehalten, ob die jeweiligen Personen einen Kauf getätigt haben (1) oder nicht (0). Die erhobenen Daten für die ersten drei Personen sind in Abbildung 14.3 dargestellt:

Der Margarinehersteller möchte nun mit Hilfe eines KNN prüfen, ob sich mit Hilfe der drei Kaufkriterien (= Eingabeneuronen) das Kaufverhalten (= Ausgabeneuron) abbilden lässt. Dabei wird im Folgenden aus didaktischen Gründen *keine* verdeckte Schicht betrachtet, sondern direkt eine Beziehung zwischen den drei Eingabeneuronen

	Eingangsvariable			Zielvariable
	Geschlecht (x_1)	Preis (x_2)	Gesundheits-bewusstsein (x_3)	Kaufverhalten (x_4)
Person 1	m (1)	1,80 €	8	Kauf (1)
Person 2	w (0)	2,00 €	8	Nichtkauf (0)
Person 3	m (1)	1,50 €	9	Nichtkauf (0)

Abbildung 14.3: Empirisch erhobene Daten der ersten drei Personen im Beispiel

und dem Ausgabeneuron unterstellt (sog. einschichtiges KNN). Kann das Ausgangsneuron durch die Informationsverarbeitung im Netz aktiviert werden, so erhält es den Wert 1, was bei der Zielvariablen einem Kauf entspricht. Weiterhin definiert der Margarinehersteller die Propagierungsfunktion als Summenfunktion gemäß Gleichung 14.1 und die Aktivierungsfunktion als Schwellenwertfunktion gem. Gleichung 14.2, wobei er $\Theta = 0,5$ setzt. Der Lernprozess, den das KNN nun durchläuft, lässt sich mit Hilfe der erhobenen Daten der ersten drei Probanden wie folgt verdeutlichen:

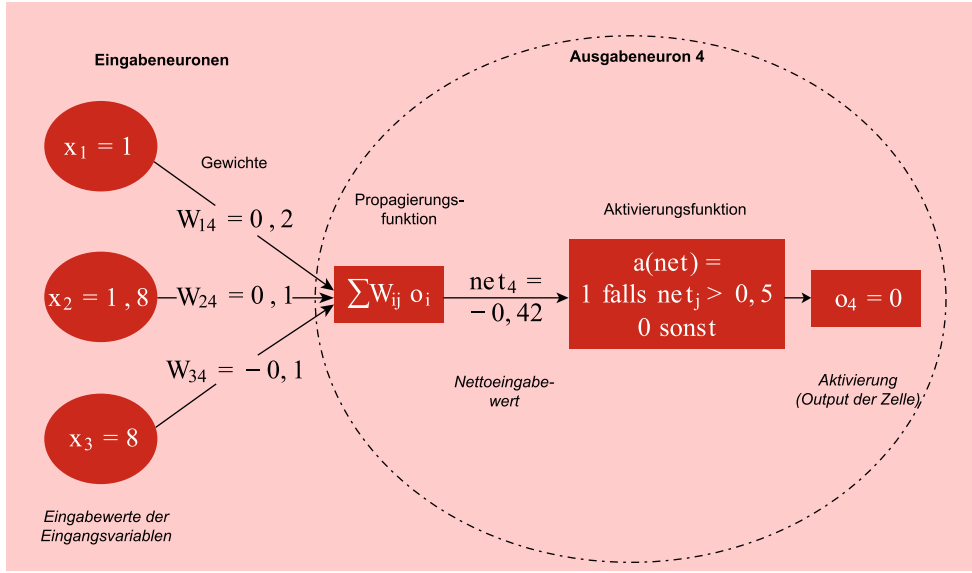


Abbildung 14.4: Informationsverarbeitungsprozess eines aktiven Neurons am Beispiel der Daten von Person 1

Im Ausgangspunkt müssen zunächst von Null verschiedene Startwerte für die Gewichte w_{ij} vorgegeben werden, damit aus den Eingangsvariablen der Nettoeingabewert für das Ausgabeneuron berechnet werden kann. Diese Startwerte können vom Anwender vorgegeben oder aber durch den Algorithmus erzeugt werden. Anschließend werden die Daten von Person 1 eingelesen und wie in Abbildung 14.4 dargestellt verarbeitet. Dabei ergibt sich für Person 1 als Nettoeingabewert:

$$0,2 \cdot 1 + 0,1 \cdot 1,8 + (-0,1) \cdot 8 = -0,42$$

Da dieser Wert $< 0,5$ ist, wird das Ausgabeneuron für Person 1 nicht aktiviert ($o_4 = 0$), womit das tatsächliche Kaufverhalten der ersten Person (Kauf) jedoch *nicht*

rekonstruiert werden kann. Um nun einen Lernprozess in Gang zu setzen, bedarf es der Definition einer Lernregel, die der Margarinehersteller wie folgt formuliert: Lernregel

„Entspricht der Aktivierungszustand des Ausgabeneurons nicht dem tatsächlich beobachteten Kaufverhalten einer Person, so sind die Gewichte der Eingabeneuronen um 5 % der Eingabewerte dieser Person zu erhöhen bzw. zu vermindern (sog. Lernrate) und dann die Daten der folgenden Person mit den neuen Gewichten einzulesen."

Nach dieser Lernregel sind die Gewichte wie folgt zu verändern:

$$w_{14} = (0,2 + 1 \cdot 0,05) = 0,25$$
$$w_{24} = (0,1 + 1,8 \cdot 0,05) = 0,19$$
$$w_{34} = (-0,1 + 8 \cdot 0,05) = 0,30$$

Die Daten von Person 2 werden dann mit den neuen Gewichten eingelesen und es ergibt sich als Nettoeingabewert für Neuron 4 bei Person 2 der Wert:

$$0,25 \cdot 0 + 0,19 \cdot 2 + 0,30 \cdot 8 = 2,78$$

Da $2,78 > 0,5$ ist, wird das Ausgabeneuron nach diesem Schritt aktiviert, womit allerdings auch in diesem Fall das tatsächliche gezeigte Kaufverhalten dieser Person (Nichtkauf) nicht abgebildet werden kann. Folglich werden die Gewichte für den nächsten Schritt mit den Daten von Person 2 um 5 % verringert und es folgt:

$$w_{14} = (0,25 - 0 \cdot 0,05) = 0,25$$
$$w_{24} = (0,19 - 2 \cdot 0,05) = 0,09$$
$$w_{34} = (0,30 - 8 \cdot 0,05) = -0,1$$

Die Daten der dritten Person werden dann mit diesen modifizierten Gewichten eingelesen und es ergibt sich als Nettoeingabewert für Neuron 4 bei Person 3 der Wert:

$$0,25 \cdot 1 + 0,09 \cdot 1,5 - 0,1 \cdot 9 = -0,515$$

Da $-0,515 < 0,5$ ist, wird das Ausgabeneuron jetzt nicht aktiviert, was in diesem Fall auch dem tatsächlichen Kaufverhalten der dritten Person (Nichtkauf) entspricht. Dieser Prozess wird auch für die weiteren befragten Personen fortgesetzt, wobei die Anpassung der Gewichte entsprechend der Lernregel sich nicht nur an dem Kaufverhalten der jeweils „eingelesenen" Person orientiert, sondern gleichzeitig auch „rückschauend" unter Berücksichtigung aller bereits eingelesenen Personen erfolgt, bis der Fehler (im Sinne der Abweichung zwischen Aktivierungszustand des Ausgabeneurons und den tatsächlichen Kaufverhaltensweisen der Personen) über alle Personen des Trainingsdatensatzes minimal wird.

Der in obigem Beispiel aufgezeigte Lernprozess des KNN weist bestimmte Eigenschaften auf, die sich zusammenfassend wie folgt charakterisieren lassen:

- In unserem Beispiel fand ein „überwachtes Lernen" statt, da für alle Personen auch das tatsächliche Kaufverhalten (Kauf/Nichtkauf) erhoben wurde. Anhand dieser „wahren Werte" konnte der korrekte Aktivierungszustand des Ausgabeneurons für jede Person kontrolliert werden. Überwachtes Lernen
- Die Informationsverarbeitung erfolgt in unserem Beispiel streng von der Eingabe- hin zur Ausgabeschicht des Netzes und es bestehen weder rückwärtsgerichtete Verbindungen noch Verbindungen zwischen den einzelnen Neuronen.

Dadurch ist eine *parallele Informationsverarbeitung* in den einzelnen Neuronen des Netzes möglich, da für die Rechenoperationen in den Neuronen lediglich die Ausgabewerte der vorgelagerten Schicht benötigt werden. Neuronale Netze, die in dieser Weise vorgehen, werden auch als „vorwärtsgerichtet“ (feedforward) bezeichnet.

Vorwärtsgerichtet

- Der Lernprozess beruhte in unserem Beispiel auf einer Anpassung der Gewichte der Eingabeneuronen, was für KNN als typisch anzusehen ist. Weiterhin erfolgte die Modifikation der Gewichte rückwärtsgerichtet, d. h. von der Ausgabe- hin zur Eingabeschicht. Da die Fehlerbestimmung auf der Ausgabeschicht ansetzt (Kauf/Nichtkauf), kann auch nur der Beitrag der Verbindungen, die direkt zur Ausgabeschicht führen, unmittelbar berechnet werden. Für die anderen Gewichte wird eine Fortpflanzung (*propagation*) des Fehlers von der Ausgabeschicht hin zur Eingabeschicht unterstellt. Dementsprechend erfolgt auch die Veränderung der Gewichte rückwärtsgerichtet (*back*). Diesem Grundsatzprinzip folgt der sog. Backpropagation-Algorithmus, der es erlaubt, die Veränderung der Gewichte effizient und strukturiert auch bei Existenz mehrerer verdeckter Schichten vorzunehmen.

Lernprozess

Backpropagation-Algorithmus

In Abhängigkeit davon, ob ein Lernen überwacht oder nicht überwacht erfolgt und ob die Informationsverarbeitung vorwärts- oder rückwärtsgerichtet ist, existieren unterschiedliche Typen von KNN. Bei praktischen Anwendungen besitzt das sog. Multi-Layer-Perceptron (MLP) eine große Bedeutung, bei dem ein überwachtes Lernen stattfindet und die Informationsverarbeitung vorwärtsgerichtet (feedforward) erfolgt. MLP-Netze werden meist mit Hilfe des sog. Backpropagation-Algorithmus trainiert, durch den die Lernmethodik in einem KNN abgebildet wird.

Multi-Layer-Perceptron

Abschließend sei an dieser Stelle noch herausgestellt, dass die in unserem einfachen Beispiel aufgezeigte Vorgehensweise zum Trainieren eines KNN nur zur Verdeutlichung der grundsätzlichen Vorgehensweise von KNN geeignet ist. Für den Einsatz in der Praxis ist diese Vorgehensweise jedoch insbesondere aus folgenden Gründen nicht anzutreffen:

- Die im Beispiel verwendete Lernregel zur Modifizierung der Gewichte ist nicht effizient, da nicht sicher ist, dass der Algorithmus auch tatsächlich ein Minimum der Fehlerrate über alle Probanden findet.
- Der Einsatz von einschichtigen Neuronalen Netzen ist nicht sinnvoll, da gerade erst durch den Einsatz *mehrerer verdeckter Schichten* mit jeweils *mehreren Neuronen* eine Approximation der Ausgabeneuronen an die empirisch erhobenen Outputdaten über alle Probanden deutlich besser erreicht werden kann.
- Insbesondere die im Beispiel verwendete Aktivierungsfunktion ist sehr einfach formuliert, da der Schwellenwertparameter Θ konstant gehalten und nicht im Verlauf des Lernprozesses modifiziert wurde. Nicht-lineare Zusammenhänge in den Daten können jedoch erst durch die Verwendung stetiger und/oder nicht-linearer Aktivierungsfunktionen erkannt werden.

Das in der praktischen Anwendung häufig verwendete Multi-Layer-Perceptron nimmt die von uns hier zur Veranschaulichung vorgenommene Simplifizierung natürlich nicht vor und findet durch Anwendung des Backpropagation-Algorithmus auch effiziente Schätzer für die Gewichte.

14.3 Umsetzung mit SPSS

Ab der Version 22 verfügt auch IBM SPSS über eine Analysefunktion zur Durchführung von Neuronalen Netzen. Die zugehörige Prozedur „*Multi-Layer-Perzeptron (MLP)*“ kann in SPSS über den Menüpunkt „*Analysieren*“ und dort den Unterpunkt „*Neuronale Netze / Mehrschichtiges Perzeptron*“ aufgerufen werden (vgl. Abbildung 14.5).

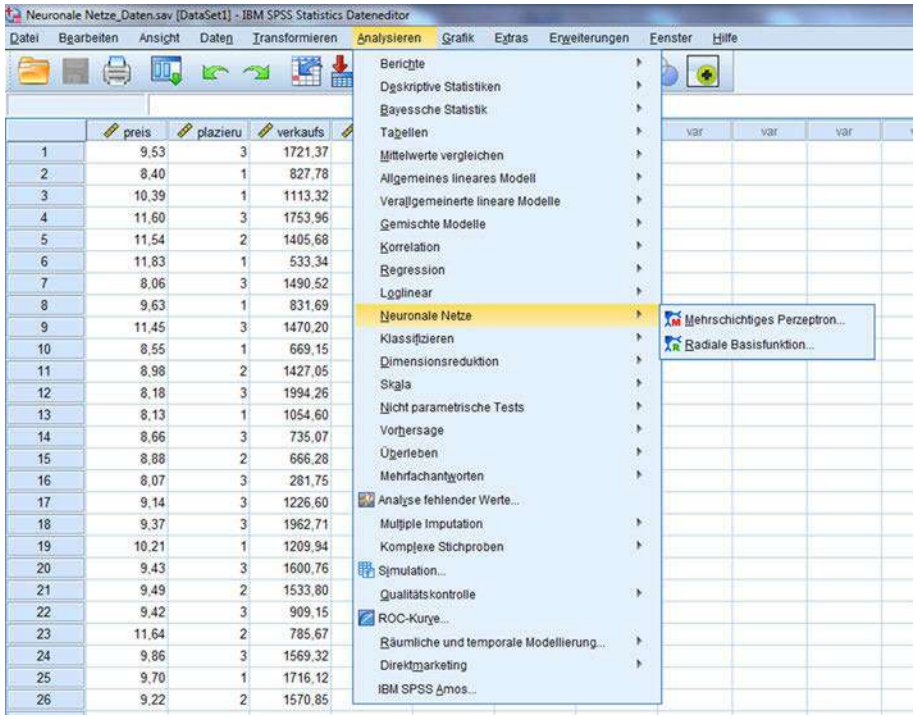


Abbildung 14.5: Daten-Editor mit Menüpunkt „Neuronale Netze“

Nach Aufruf der Prozedur „MLP“ sind die abhängigen und die unabhängigen Variablen festzulegen, wobei letztere sowohl metrisch (sog. Kovariate) als auch nominal skaliert (sog. Faktoren) sein können. In weiteren Dialogfeldern ist dann eine Deklaration der Trainings- bzw. Testdaten vorzunehmen und die Netztopologie (Anzahl verborgener Schichten, Art der Aktivierungsfunktion usw.) festzulegen.

In früheren Versionen von SPSS konnten Neuronale Netze nur mit Hilfe des Zusatzpaketes „*SPSS Clementine*“ durchgeführt werden. Diese Option besteht auch heute noch, wobei „*Clementine*“ aber in das von IBM angebotene Erweiterungsmodul „*IBM SPSS Modeler*“ integriert wurde. In der dritten Auflage unseres Lehrbuchs „Fortgeschrittene Multivariate Analysemethoden“ wird das Fallbeispiel mit der SPSS-Prozedur „MLP“ gerechnet und abschließend mit den Berechnungen unter Verwendung des *Modelers* verglichen. Beide Prozeduren führen in den zentralen Ergebnissen und den Gütekriterien zu den gleichen Ergebnissen. Für den interessierten Leser wurde eine Dokumentation der Vorgehensweise mit dem Modeler auf der Internetseite zum Buch „www.multivariate.de“ unter dem Register Service → Download kostenfrei hinterlegt.

15 Multidimensionale Skalierung

Die Multidimensionale Skalierung wird in diesem Buch den „*Fortgeschrittenen Verfahren der multivariaten Analyse*“ zugeordnet und in diesem Kapitel nur in den Grundzügen behandelt. Eine ausführliche Darstellung zum Kapitel „Multidimensionale Skalierung“ ist verfügbar in dem Buch *„Backhaus, K./Erichson, B.Weiber, R.: Fortgeschrittene Multivariate Analysemethoden, 3. Aufl., Berlin Heidelberg 2015.“*

15.1 Problemstellung

Die Multidimensionale Skalierung (MDS) umfasst Verfahren, mittels derer sich Objekte auf Basis ihrer Ähnlichkeiten oder Unähnlichkeiten gemeinsam in einem zwei- oder mehrdimensionalen Raum darstellen lassen. Ein primäres Anwendungsgebiet bilden Positionierungsanalysen, mittels derer sich die subjektive Wahrnehmung von Objekten durch Personen (z. B. der Wahrnehmung von Produkten durch Konsumenten, von Politikern durch Wähler, von Universitäten durch Studenten) visualisieren lässt. Man geht davon aus, dass die Objekte im *Wahrnehmungsraum* der Personen feste Positionen einnehmen und sich somit wie Städte auf einer Landkarte darstellen lassen. Je dichter zwei Objekte im *Wahrnehmungsraum* beieinander liegen, als desto ähnlicher werden sie empfunden, und je weiter sie voneinander entfernt liegen, als desto unähnlicher werden sie empfunden.

Wahrnehmungsraum

Die Menge der Objekte und ihrer relativen Positionen zueinander bezeichnet man als *Konfiguration*. Abbildung 15.1 zeigt beispielhaft eine solche Konfiguration für elf Streichfette (Margarine- und Buttermarken), wie man sie mit SPSS erhält. Die Abbildung basiert auf der Befragung von 32 Personen. Eine analoge Darstellung lässt sich aber auch für jede einzelne Person ermitteln.

Konfiguration

Abbildung 15.1 lässt sich auch als „Mapping“ oder „Perceptual Mapping“ bezeichnen und ähnelt den Mappings in Abbildung 7.7 oder 7.63 von Kapitel 7, die mittels Faktorenanalyse gewonnen wurden. Das Ziel ist in beiden Fällen das gleiche, die verwendeten Daten und die Methodik aber sind verschieden.

Perceptual Mapping

- Die *faktorielle Positionierung* basiert auf Beurteilungen der *Eigenschaften* von Objekten. Die Eigenschaften muss der Untersucher auswählen und verbalisieren.
- Die *Positionierung mittels MDS* basiert auf Beurteilungen der *Un-/ Ähnlichkeiten* zwischen den Objekten.

K. Backhaus, B. Erichson, W. Plinke, R. Weiber, *Multivariate Analysemethoden*,
https://doi.org/10.1007/978-3-662-56655-8_16

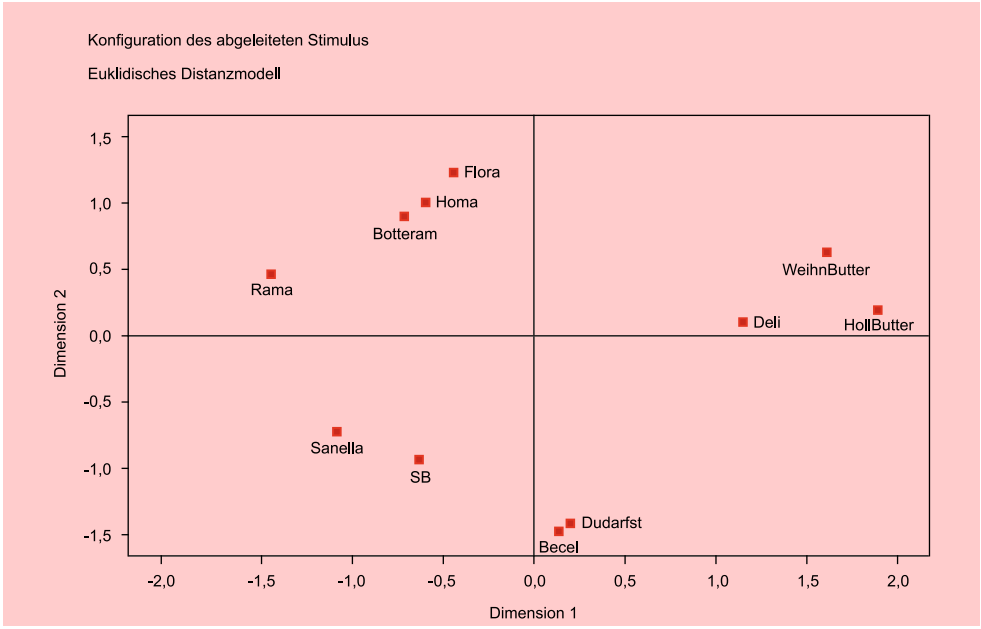


Abbildung 15.1: MDS für den Margarinemarkt

Vorteile

Vorteile der MDS gegenüber der *faktoriellen Positionierung* sind darin zu sehen, dass die relevanten Eigenschaften unbekannt sein können und dass keine Beeinflussung des Ergebnisses durch die Auswahl der Eigenschaften und deren Verbalisierung durch den Untersucher erfolgt.

Vorteile der faktoriellen Positionierung gegenüber der MDS sind darin zu sehen, dass die Datenerhebung leichter durchzuführen ist und dass sich die Ergebnisse leichter interpretieren lassen, da ein Bezug zwischen den abgeleiteten Dimensionen und den vorgegebenen Eigenschaften besteht.

Nichtmetrische MDS

Für die gebräuchlichen Verfahren der MDS ist es ausreichend, wenn die Input-Daten ordinales Skalenniveau haben. Man spricht daher auch von *nichtmetrischer Multidimensionaler Skalierung*. Nur diese Verfahren seien hier betrachtet. Die Ergebnisse dieser Verfahren haben aber immer metrisches Skalenniveau. Durch Verdichtung von Daten wird eine Erhöhung des Skalenniveaus ermöglicht.

15.2 Allgemeine Vorgehensweise

Die Durchführung einer MDS beginnt in der Regel mit der Datengewinnung.

Datengewinnung

Un-/Ähnlichkeitsurteile

Im Rahmen der Datengewinnung muss der Untersucher Ähnlichkeits- oder Unähnlichkeitsurteile von Personen (z. B. von potenziellen Käufern einer Produktklasse) erheben. Un-/Ähnlichkeitsurteile beziehen sich nicht isoliert auf einzelne Objekte, sondern immer auf Paare von Objekten. Zur Erhebung von Un-/Ähnlichkeitsurteilen werden unterschiedliche Methoden angewendet.

a) Methode der Rangreihung

Rangreihung

Die Methode der Rangreihung bildet das klassische Verfahren zur Erhebung von Un-/Ähnlichkeitsurteilen. Dabei wird eine Auskunftsperson veranlasst, die Objektpaare nach ihrer empfundenen Ähnlichkeit zu ordnen, d. h. sie nach aufsteigender oder abfallender Ähnlichkeit in eine Rangfolge zu bringen. Hierzu werden ihr Kärtchen vorgelegt, auf denen jeweils ein Objektpaar angegeben ist.

Bei K Objekten ergeben sich $K(K-1)/2$ Paare (Kärtchen), die zu ordnen sind. Bei $K = 11$ Objekten, wie in obigem Beispiel, ergeben sich 55 Paare. Die Zahl der Paare nimmt quadratisch mit der Zahl der Objekte zu. Hierin liegt ein Nachteil der MDS gegenüber der faktoriellen Positionierung, bei der der Erhebungsaufwand nur linear mit der Zahl der Objekte zunimmt.

Um bei größerer Anzahl von Objekten die Aufgabe zu erleichtern, lässt man die Auskunftsperson zunächst zwei Gruppen bilden: „ähnliche Paare“ und „unähnliche Paare“, welche im zweiten Schritt jeweils wieder in zwei Untergruppen wie „ähnlichere Paare“ und „weniger ähnliche Paare“ geteilt werden usw., bis letztlich eine vollständige Rangordnung vorliegt.

MDS-Algorithmen

Für die Anwendung von MDS-Algorithmen sind die Objektpaare entsprechend ihrer Reihenfolge mit Zahlen (Rangwerten) zu versehen, d. h. bei z. B. 55 Paaren sind diesen die Ränge 1 bis 55 zuzuordnen. Dies kann alternativ so erfolgen, dass man Ähnlichkeits- oder Unähnlichkeitsdaten (similarities and dissimilarities) erhält:

$$\text{Ähnlichkeitsdaten:} 1 = \text{unähnlichstes Paar}, 55 = \text{ ähnlichstes Paar}$$
$$\text{Unähnlichkeitsdaten:} 1 = \text{ähnlichstes Paar}, 55 = \text{ unähnlichstes Paar}$$

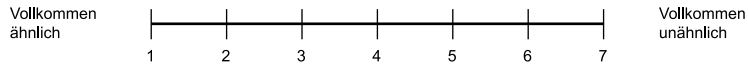
Üblich ist die zweite Alternative, d. h. mit Rangdaten sind üblicherweise Unähnlichkeitsdaten gemeint. Ein kleines Beispiel zeigt Abbildung 15.2. „Rama“ und „Holl.Butter“ bilden das unähnlichste Paar und „Becel“ und „Du darfst“ das ähnlichste. Bei der Auswertung mit Computer-Programmen sind prinzipiell beide Alternativen zulässig; es muss nur dem Programm korrekt mitgeteilt werden, wie die Daten kodiert wurden, da man andernfalls unsinnige Ergebnisse erhält. Da eine Un-/Ähnlichkeitsmatrix symmetrisch ist, braucht nur die untere oder obere Dreiecksmatrix angegeben zu werden.

	Becel	Du darfst	Rama	Delicado	Holl.Butter	Weih.Butter
Becel	0					
Du darfst	1	0				
Rama	7	8	0			
Delicado	6	5	9	0		
Holl.Butter	11	10	15	4	0	
Weih.Butter	13	12	14	3	2	0

Abbildung 15.2: Beispiel einer Unähnlichkeitsmatrix

b) Methode der Ratings

Bei dieser Methode werden die Objektpaare jeweils einzeln auf einer Ähnlichkeits- oder Unähnlichkeitsskala eingestuft, z. B.:
Die Marken „Becel“ und „Du darfst“ sind



Die Auskunftsperson soll jeweils den ihrer Meinung nach zutreffenden Punkt auf der Skala ankreuzen. Üblich sind 7− oder 9−stufige Skalen. Da hier größere Werte geringere Ähnlichkeit bedeuten, handelt es sich um eine Unähnlichkeitsskala.

Ratingverfahren

Das *Ratingverfahren* lässt sich von den Auskunftspersonen einfacher und schneller durchführen, da jedes Objektpaar isoliert beurteilt wird und nicht mit den anderen Paaren verglichen werden muss. Bei großer Anzahl von Objekten bzw. geringer Belastbarkeit der Auskunftspersonen ist es daher vorzuziehen. Es liefert aber auch ungenauere Daten, da zwangsläufig, wenn z. B. 55 Paare auf einer 7−stufigen Ratingskala

Ties

beurteilt werden, verschiedene Paare gleiche Ähnlichkeitswerte, sog. *Ties*, erhalten. Je größer die Zahl der Objekte und je geringer die Stufigkeit der Ratingskala, desto mehr derartige Ties treten auf.

Wahl eines Distanzmodells

Die Umsetzung der gewonnenen Un-/Ähnlichkeitsdaten in eine grafische Darstellung (Konfiguration) erfolgt mittels MDS so, dass Objekte, die als ähnlich beurteilt werden, dicht beieinander liegen sollen, und Objekte, die als unähnlich beurteilt werden, sollen weit auseinander liegen. Hierzu bedarf es der Wahl eines Distanzmaßes (Me-

Distanzmaß

trik). Die Mathematik kennt unterschiedliche Räume, die sich hinsichtlich ihrer Metrik unterscheiden.

Euklid-Metrik

Am gebräuchlichsten ist die Euklidische Metrik, die unserer räumlichen Vorstellung entspricht. Die Euklidische Distanz zweier Punkte entspricht der kürzesten Entfernung zueinander („Luftlinie“) (vgl. Abbildung 15.3):

$$d_{kl} = \left[\sum_{r=1}^{R}(x_{kr} - x_{lr})^2\right]^{\frac{1}{2}} \tag{15.1}$$

mit

d_{kl} : Distanz der Punkte k, l
x_{kr}, x_{lr}: Koordinaten der Punkte k, l auf der r-ten Dimension $(r = 1, 2, ..., R)$

Eine alternatives Distanzmodell bildet die City-Block-Metrik, bei welcher sich die Distanz zweier Punkte als Summe der absoluten Abstände auf den einzelnen Dimen-

City-Block-Metrik

sionen ergibt.

$$d_{kl} = \sum_{r=1}^{R}|x_{kr} - x_{lr}| \tag{15.2}$$

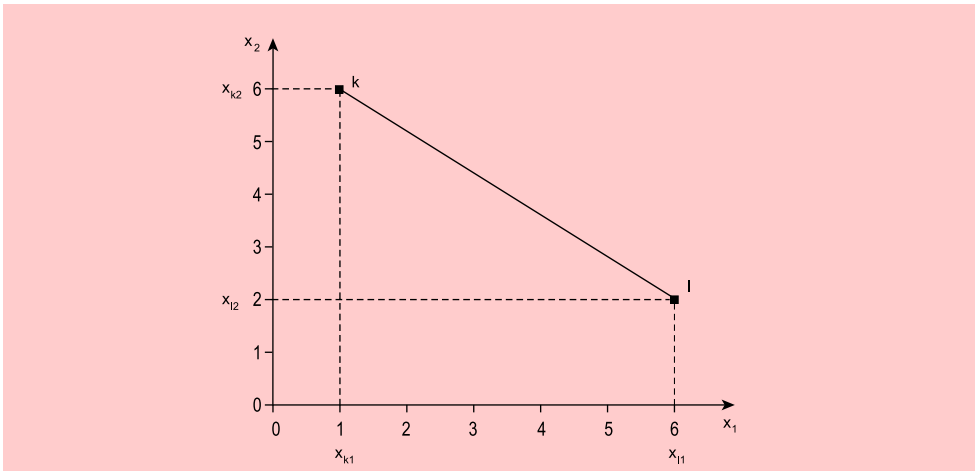


Abbildung 15.3: Euklidische Distanz

Ermittlung der Konfiguration

Verfahren der MDS

Das Verfahren der MDS lässt sich formal wie folgt umreißen: Es sei u_{kl} die Unähnlichkeit zwischen den Objekten k und l. Gegeben sind die Unähnlichkeiten zwischen K Objekten:

$$\{u_{kl}\}_{k,l=1,...,K}$$

In einem Raum mit möglichst geringer Dimensionalität R wird eine Konfiguration von Punkten $\mathbf{x} = (x_{k1}, x_{k2}, ..., x_{kR})$ gesucht, wobei durch x_{kr} die Koordinaten der Punkte bezeichnet seien.

Die Distanzen d_{kl} zwischen den Punkten k und l oder i und j sollten dabei möglichst gut die folgende Monotoniebedingung erfüllen:

$$\text{Wenn } u_{kl} > u_{ij}, \text{ dann } d_{kl} > d_{ij} \tag{15.3}$$

In der gesuchten Konfiguration sollte also die Rangfolge der Distanzen zwischen den Objekten *möglichst gut* die Rangfolge der vorgegebenen Unähnlichkeiten wiedergeben. Eine perfekte Erfüllung der Monotoniebedingung ist i. d. R. nicht möglich (und sollte auch, wie unten noch erläutert wird, nicht möglich sein).

Iteratives Vorgehen

Um die Konfiguration zu finden, geht man iterativ vor. Man startet mit einer Ausgangskonfiguration und versucht, diese schrittweise zu verbessern. Dazu bedarf es eines Gütemaßes, das als Zielkriterium dienen kann. Von J.B. Kruskal wurde hierfür das sog. *Stress-Maß* eingeführt:

STRESS

$$\text{STRESS} = \sqrt{\frac{\sum_k \sum_l (d_{kl} - \hat{d}_{kl})^2}{\text{Faktor}}} \tag{15.4}$$

mit

d_{kl} : Distanz zwischen Objekten k und l
$\hat{d}_{kl}$: Disparitäten für Objekte k und l

Für die Disparitäten $\hat{d}_{kl}$ gilt:

$$\text{Wenn } u_{kl} > u_{ij}, \text{ dann } \hat{d}_{kl} \geq \hat{d}_{ij} \tag{15.5}$$

Die Disparitäten bilden also schwach monotone Transformationen der Unähnlichkeiten. Damit wird deutlich, dass die Input-Daten der MDS, die Unähnlichkeiten, nur ordinales Skalenniveau haben müssen. Eine beliebige monotone Transformation der Rangwerte (z. B. der Quadrierung oder Logarithmierung) würde am Ergebnis nichts ändern. Entscheidend ist lediglich, dass die Reihenfolge der Distanzen erhalten bleibt.

Die Größe des Stress-Maßes wird bestimmt durch die Differenzen $(d_{kl} - \hat{d}_{kl})$ zwischen Distanzen und Disparitäten. Das obige Stress-Maß bildet ein Kleinstquadratekriterium. Eine Lösung ist um so besser, je kleiner das Stress-Maß ist. Im Fall einer exakten monotonen Anpassung entsprechen alle Distanzen den Disparitäten und der Stress nimmt den Wert 0 an.

Der Faktor im Nenner von (15.4) dient lediglich zur Normierung des Stress-Maßes auf Werte zwischen 0 und 1. Hier existieren unterschiedliche Varianten. Gebräuchlich sind die *Stress-Formeln 1 und 2* von Kruskal:

$$\text{STRESS-1} = \sqrt{\frac{\sum_k \sum_l (d_{kl} - \hat{d}_{kl})^2}{\sum_k \sum_l {d_{kl}}^2}} \tag{15.6}$$

$$\text{STRESS-2} = \sqrt{\frac{\sum_k \sum_l (d_{kl} - \hat{d}_{kl})^2}{\sum_k \sum_l (d_{kl} - \overline{d})^2}} \tag{15.7}$$

mit

$\overline{d}$: Mittelwert der Distanzen

Die obigen Stress-Formeln 1 und 2 führen zum selben Ergebnis und finden in den meisten Computer-Programmen für die MDS Verwendung (z. B. MDSCAL, KYST, POLYCON, PROXSCAL). Beim Vergleich der beiden Stress-Maße ist zu beachten, dass STRESS-2 etwa doppelt so große Werte wie STRESS-1 liefert.

S-STRESS

Ein weiteres Stress-Maß ist S-STRESS von Takane/Young/de Leeuw, das in dem Programm ALSCAL als Zielkriterium verwendet wird.

$$\text{S-STRESS} = \sqrt{\frac{\sum_k \sum_l (d_{kl}^2 - \hat{d}_{kl}^2)^2}{\sum_k \sum_l d_{kl}^2}} \tag{15.8}$$

Zahl und Interpretation der Dimensionen

Ein Wahrnehmungsraum wird neben der *Metrik* auch durch die *Zahl der Dimensionen* bestimmt. Beides muss vom Anwender einer MDS festgelegt werden. Aus praktischen Erwägungen wird man sich meist auf zwei oder drei Dimensionen beschränken, um eine grafische Darstellung der Ergebnisse zu ermöglichen und so die inhaltliche Interpretation zu erleichtern (aber auch eine einzige Dimension kann ausreichend sein). Da sich unsere räumliche Erfahrung und Vorstellung auf maximal drei Dimensionen beschränkt, erscheint dies auch für unseren Wahrnehmungsraum angemessen.

Als formales Kriterium zur Bestimmung der Zahl der Dimensionen kann das Stress-Maß herangezogen werden. Der Stress einer Lösung sollte möglichst niedrig sein. Dabei ist aber zu beachten, dass generell der Stress abnimmt, wenn die Zahl der Dimensionen erhöht wird. Eine stressfreie Lösung lässt sich immer erzielen, wenn die Zahl

der Dimensionen hinreichend groß gemacht wird (für K Objekte gilt dies immer bei $K-1$ Dimensionen). Bei nur geringfügiger Änderung des Stresses sollte daher die Lösung mit geringerer Anzahl von Dimensionen vorgezogen werden. Zur Unterstützung der Entscheidung kann das Elbow-Kriterium herangezogen werden (vgl. Kapitel 8: Cluster-Analyse).

Vorsicht ist geboten, wenn der Stress null oder sehr klein wird (z. B. $< 0,01$), da dies ein Indiz für eine *degenerierte Lösung* sein kann. Die Objekte klumpen sich dann meist im Mittelpunkt des Koordinatensystems. Ein gewisses Mindestmaß an Stress ist deshalb bei der MDS immer notwendig, um eine eindeutige Lösung zu erhalten.

Degenerierte Lösung

Ein inhaltliches Kriterium zur Bestimmung der Zahl der Dimensionen bildet die Interpretierbarkeit der Konfiguration wie auch der Dimensionen des Wahrnehmungsraumes. Wenngleich eine Interpretation der Dimensionen nicht immer möglich oder notwendig ist, so erhöht sich dadurch doch die Anschaulichkeit, und die Validität der gefundenen Lösung wird bestärkt. Zwecks besserer Interpretierbarkeit ist es oft hilfreich, die Konfiguration geeignet zu rotieren oder zu spiegeln.

Interpretierbarkeit

Aggregation von Personen

Eine MDS kann auf Basis der Un-/Ähnlichkeitsurteile einer einzelnen Person durchgeführt werden. Bei vielen Anwendungen interessieren jedoch nicht so sehr die individuellen Wahrnehmungen, sondern diejenigen von Gruppen, z. B. bei der Analyse der Markenwahrnehmung durch Käufergruppen oder Marktsegmente.

Grundsätzlich bieten sich drei Möglichkeiten zur Lösung des Aggregationsproblems an:

Aggregation

a) Es werden vor der Durchführung der MDS die Ähnlichkeitsdaten durch Bildung von Mittelwerten oder Medianen aggregiert. Auf die so aggregierten Daten wird dann die MDS angewendet.

b) Es wird eine MDS für jede Person durchgeführt und anschließend werden die Ergebnisse aggregiert, indem man für jedes Objekt und jede Dimension die Mittelwerte der Koordinaten berechnet. Da die Ergebnisse immer metrisch sind im Gegensatz zu den Input-Daten, erscheint diese Vorgehensweise adäquater. Diese Vorgehensweise ist allerdings sehr aufwändig und infolge von Ties und fehlenden Werten nicht immer möglich. Sie erfordert außerdem, dass die individuellen Konfigurationen durch Rotation vergleichbar gemacht wurden.

c) Die heute gebräuchlichen Computer-Programme erlauben auch eine gemeinsame Analyse der Ähnlichkeitsdaten einer Mehrzahl von Personen, für die dann eine gemeinsame Konfiguration ermittelt wird. Man bezeichnet diese Art der MDS auch als RMDS (replicated MDS). Für jede Person ist eine Matrix der Un-/Ähnlichkeitsurteile einzugeben (vgl. Abbildung 15.2).

Beim Vergleich einer MDS auf Basis von aggregierten Ähnlichkeitsdaten und einer RMDS ist zu berücksichtigen, dass letztere zwangsläufig höhere Stress-Werte liefert. Daraus darf nicht der Fehlschluss gezogen werden, dass die extern aggregierten Daten eine bessere Abbildung der Objekte im Wahrnehmungsraum liefern.

RMDS

Grundsätzlich ist bei der Aggregation über Personen zu prüfen, ob hinreichende Homogenität der Personen vorliegt. Andernfalls ist z. B. mit Hilfe der Cluster-Analyse (vgl. Kapitel 8) zuvor eine Segmentierung auf Basis der Ähnlichkeitsurteile vorzunehmen, d. h. es sind möglichst homogene Cluster zu bilden, innerhalb derer eine Aggregation zulässig ist.

15.3 Umsetzung mit SPSS

PROXSCAL

ALSCAL

Zur Durchführung einer MDS bietet SPSS alternativ die Prozeduren PROXSCAL und ALSCAL, die über den Menüpunkt „Analysieren/ Skalierung“ (vgl. Abbildung 15.4) aufgerufen werden können. In PROXSCAL wird als Optimierungskriterium das Stress-Maß von Kruskal verwendet. Es werden STRESS-1, STRESS-2 und S-STRESS ausgegeben. In ALSCAL wird S-STRESS als Optimierungskriterium verwendet. Abbildung 15.1 zeigt eine Lösung, die mit ALSCAL erzielt wurde. Da PROXSCAL ein anderes Optimierungskriterium verwendet, liefert es eine zwar ähnliche, aber doch unterschiedliche Lösung.

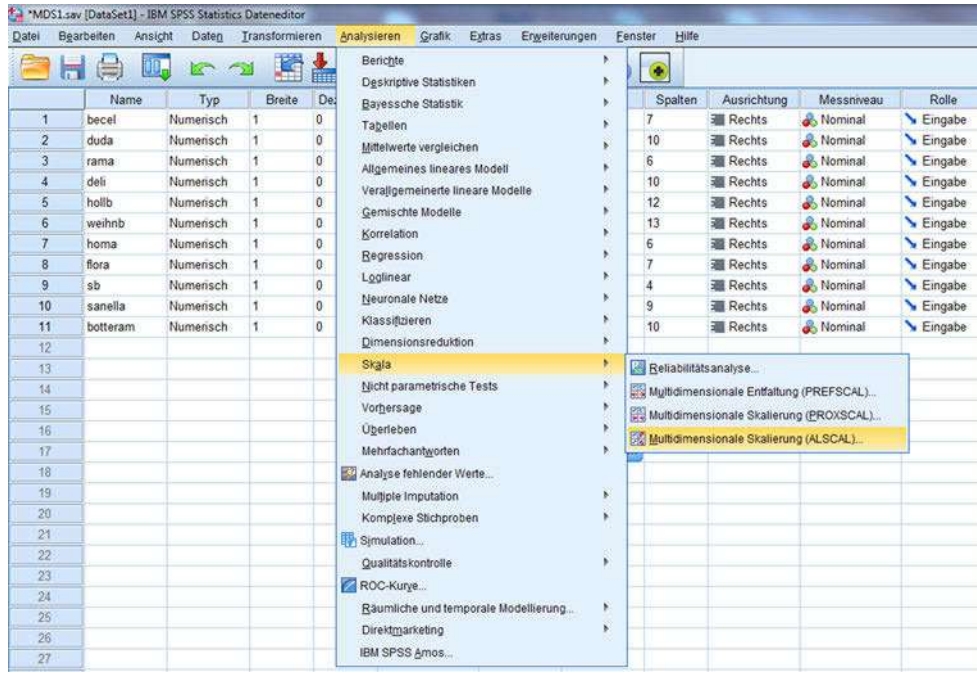


Abbildung 15.4: Daten-Editor mit Menüpunkt „Skala / Multidimensionale Skalierung“

16 Korrespondenzanalyse

Die Korrespondenzanalyse wird in diesem Buch den „*Fortgeschrittenen Verfahren der multivariaten Analyse*“ zugeordnet und in diesem Kapitel nur in den Grundzügen behandelt. Eine ausführliche Darstellung zum Kapitel „Korrespondenzanalyse“ ist verfügbar in dem Buch *„Backhaus, K./Erichson, B.Weiber, R.: Fortgeschrittene Multivariate Analysemethoden, 3. Aufl., Berlin Heidelberg 2015.“*

16.1 Problemstellung

Visualisierung von Kreuztabellen

Die *Korrespondenzanalyse* ist ein Verfahren zur Visualisierung von Kreuztabellen. In einer Kreuztabelle werden die *gemeinsamen Häufigkeiten* der Merkmalsausprägungen von zwei kategorialen Variablen zusammengestellt. Ein Beispiel zeigt Abbildung 16.1. Bei 45 Personen wurde ermittelt, welches die von ihnen bevorzugte Margarinemarke ist und welches Merkmal (Benefit) für sie bei einer Margarine besonders wichtig ist. In den Zellen der Kreuztabelle sind für jede Kombination von Marke und Merkmal die betreffenden Häufigkeiten zusammengestellt. Die Häufigkeiten addieren sich über die Zellen zur Fallzahl der Kreuztabelle, d. h. der Anzahl der Personen. Die gemeinsamen Häufigkeiten in den Zellen einer Kreuztabelle sind ein Ausdruck des Zusammenhangs (Korrespondenz) der betreffenden Kombination von Marke und Merkmal (Benefit).

Biometrik

Ähnliche Beispiele wären im Bereich der Biometrik die gemeinsamen Häufigkeiten von Haarfarbe und Augenfarbe von Personen, im Bereich der Linguistik das gemeinsame Vorkommen von bestimmten Vokalen und Konsonanten oder in der Wahlforschung die Häufigkeiten, mit denen die politischen Parteien von Wählern unterschiedlicher Berufsgruppen gewählt werden.

	Margarinemerkmale		
Margarinemarken	Geschmack	Energie	Gesundheit
Delicado	9	3	6
Rama	3	6	3
Becel	1	1	2
Homa	2	5	4

Abbildung 16.1: Kreuztabelle (Margarinebeispiel)

Mit derartigen Kreuztabellen befasst sich auch die Kontingenzanalyse (Kapitel 6). Während aber die Kontingenzanalyse die Daten statistisch analysiert, um die Si-

K. Backhaus, B. Erichson, W. Plinke, R. Weiber, *Multivariate Analysemethoden*,
https://doi.org/10.1007/978-3-662-56655-8_17

gnifikanz von Zusammenhängen zu prüfen, bezweckt die Korrespondenzanalyse die grafische Darstellung der Daten.

Verwandschaft mit der Faktorenanalyse

Die Korrespondenzanalyse dient damit der Veranschaulichung komplexer Sachverhalte und ist als strukturen-entdeckendes Verfahren einzuordnen. Sie ist damit verwandt mit der Faktorenanalyse (Kapitel 7) und der Multidimensionalen Skalierung (Kapitel 15). Aufgrund ihrer Ähnlichkeit mit der Faktorenanalyse bzw. deren spezieller Form der Hauptkomponentenanalyse wird die Korrespondenzanalyse auch als „Hauptkomponentenanalyse mit kategorialen Daten“ bezeichnet. Die Methodik wie auch die Ergebnisse sind aber doch recht verschieden.

Durch Anwendung der Korrespondenzanalyse lassen sich die Zeilen- und Spaltenelemente einer Kreuztabelle als Punkte in einem gemeinsamen Raum (Korrespondenzraum, joint space) darstellen. Für die Beispieldaten erhält man die Abbildung 16.2,

Biplot

die auch als *Biplot* bezeichnet wird. Aus der Lage der Punkte lässt sich erkennen, dass Rama und Homa von den befragten Personen als ähnlich wahrgenommen werden, da die betreffenden Punkte relativ dicht beieinander liegen, während Delicado und Becel als unähnlich zueinander wie auch zu Rama und Homa empfunden werden, da sie relativ isolierte Positionen einnehmen. Da auch die drei Merkmale durch Punkte in demselben Raum repräsentiert werden, erkennt man weiterhin, dass die Position von Delicado sehr nahe bei dem Merkmal „Geschmack“ liegt, während Rama und Homa in der Nähe des Merkmals „Energie“ positioniert sind.

Bei der Korrespondenzanalyse werden Zeilen und Spalten in gleicher Weise behandelt (anders als bei der Faktorenanalyse). Das Ergebnis einer Korrespondenzanalyse ändert sich nicht, wenn man Zeilen und Spalten vertauscht, also im Beispiel die Margarinemarken in den Spalten und die Margarinemerkmale in den Zeilen anordnet. Hierin ist ein entscheidender Unterschied zur Faktorenanalyse zu sehen.

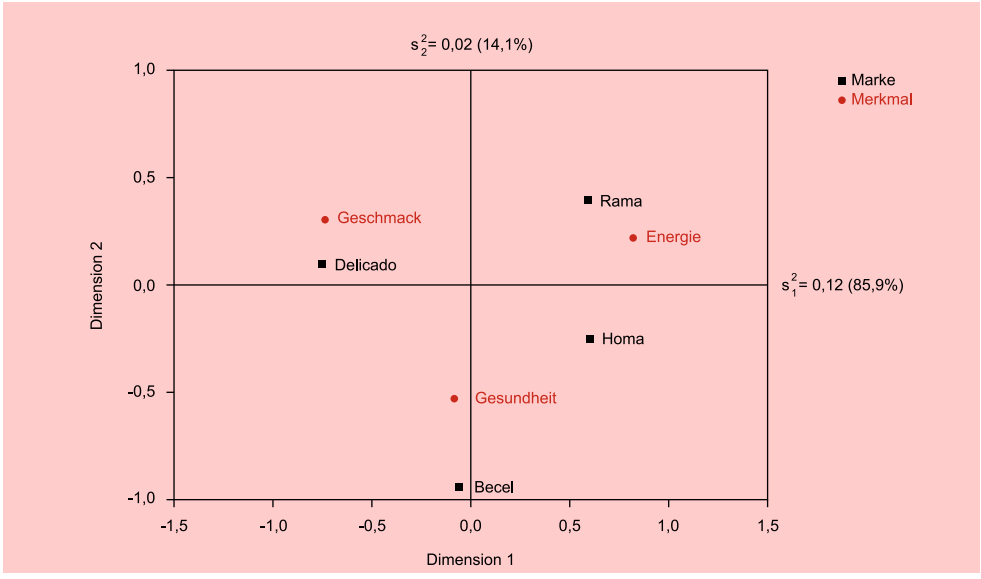


Abbildung 16.2: Korrespondenzanalyse für das Margarinebeispiel (Symmetrische Normalisierung)

16.2 Allgemeine Vorgehensweise

Die rechnerische Durchführung einer Korrespondenzanalyse lässt sich in drei Schritte gliedern:

- Standardisierung der Daten,
- Extraktion der Dimensionen,
- Normalisierung der Koordinaten.

Schritt 1: Standardisierung der Daten

Die gemeinsamen Häufigkeiten einer (I x J)-Kreuztabelle (mit I Zeilen J Spalten) seien durch n_{ij} bezeichnet: n_{ij} = Häufigkeit der Merkmalskombination (i, j).

Dividiert man die Häufigkeiten n_{ij} durch die Fallzahl der Kreuztabelle (Gesamthäufigkeit)

$$n = \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij}, \tag{16.1}$$

so erhält man die relativen Häufigkeiten

$$p_{ij} = n_{ij}/n. \tag{16.2}$$

Die Fallzahl der (4 x 3)-Kreuztabelle in Abbildung 16.1 beträgt 45 und man erhält durch Division die relativen Häufigkeiten in Abbildung 16.3, die jetzt unabhängig von der Fallzahl sind.

	Margarinemerkmale		
Margarinemarken	Geschmack	Energie	Gesundheit
Delicado	0,200	0,067	0,133
Rama	0,067	0,133	0,067
Becel	0,022	0,022	0,044
Homa	0,044	0,111	0,089

Abbildung 16.3: Kreuztabelle mit relativen Häufigkeiten

Aus den Randverteilungen (marginalen Häufigkeiten) einer Kreuztabelle, den Zeilen- und Spaltensummen, lässt sich für jede Merkmalskombination (i, j) eine erwartete Häufigkeit berechnen. Die *erwarteten relativen Häufigkeiten* erhält man durch

Erwartete relative Häufigkeiten

$$\hat{e}_{ij} = \frac{n_{i.} \cdot n_{.j}}{n^2} \quad \text{mit} \quad n_{i.} = \sum_{j=1}^{J} n_{ij} \quad \text{und} \quad n_{.j} = \sum_{i=1}^{I} n_{ij} \tag{16.3}$$

Die erwarteten Häufigkeiten geben an, welche Werte theoretisch bei Unabhängigkeit zwischen Zeilen- und Spaltenmerkmalen zu erwarten wären. In den Abweichungen der beobachteten Häufigkeiten von den erwarteten Häufigkeiten

$$p_{ij} - \hat{e}_{ij} \tag{16.4}$$

zeigt sich der *Informationsgehalt* der Daten. Weichen die beobachteten Werte nicht oder nur wenig von den erwarteten Werten ab, so enthalten sie auch keine oder nur

wenig Information, denn sie könnten dann auch aus den marginalen Häufigkeiten der Kreuztabelle berechnet werden. In diesem Falle wäre auch keine Visualisierung der Zeilen und Spalten der Kreuztabelle möglich, da die Punkte im Zentrum des Korrespondenzraumes einen Klumpen bilden würden.

In der Korrespondenzanalyse werden die relativen Häufigkeiten unter Verwendung der erwarteten relativen Häufigkeiten wie folgt standardisiert:

$$z_{ij} = \frac{p_{ij} - \hat{e}_{ij}}{\sqrt{\hat{e}_{ij}}} \tag{16.5}$$

Abbildung 16.4 zeigt die standardisierten Daten für das Beispiel:

	Margarinemerkmale		
Margarinemarken	Geschmack	Energie	Gesundheit
Delicado	0,183	-0,183	0,000
Rama	-0,075	0,149	-0,075
Becel	-0,043	-0,043	0,086
Homa	-0,130	0,104	0,026

Abbildung 16.4: Standardisierte Daten z_{ij}

Chi-Quadrat-Abweichungen

Die Quadrate der standardisierten Daten werden als Chi-Quadrat-Abweichungen bezeichnet. Mit ihnen lässt sich die *Chi-Quadrat-Statistik* recht einfach wie folgt berechnen:

$$\chi^2 = \sum \frac{(\text{beobachtete Häufigkeit} - \text{erwartete Häufigkeit})^2}{\text{erwartete Häufigkeit}} = n \cdot \sum_i \sum_j z_{ij}^2 \tag{16.6}$$

Die Chi-Quadrat-Statistik misst die Streuung der beobachteten Werte um die erwarteten Werte und kann somit als Maß für die in einer Kreuztabelle enthaltene Streuung oder Information angesehen werden. In der Kontingenzanalyse wird Chi-Quadrat zur Prüfung der Abhängigkeit zwischen Zeilen und Spalten verwendet. Unter der Nullhypothese, dass Zeilen und Spalten voneinander unabhängig sind, ist die Chi-Quadrat-Statistik annähernd chi-quadrat-verteilt. Bei Vorliegen eines großen Wertes von Chi-Quadrat kann die Nullhypothese widerlegt werden (vgl. Kapitel 6: Kontingenzanalyse).

Ein Nachteil von Chi-Quadrat als Maß für die Streuung ist, dass es von der Höhe der Fallzahl der Daten abhängig ist, d. h. man erhält auch hohe Werte für Chi-Quadrat bei Daten mit niedriger Streuung, wenn nur die Fallzahl hinreichend groß ist. In der Kontingenzanalyse wird daher das durch die Fallzahl dividierte Chi-Quadrat verwendet. Es wird als *Inertia* (Trägheit) einer Kreuztabelle bezeichnet:

Inertia

$$T = \frac{\chi^2}{n} = \sum_i \sum_j z_{ij}^2 \tag{16.7}$$

Für unser Beispiel erhält man für die Inertia:

$$T = \frac{6,27}{45} = 0,139$$

Der Wertebereich der Inertia ist begrenzt durch die Anzahl der Zeilen und Spalten einer Kreuztabelle. T kann nicht größer werden als

$$K = Min\{I, J\} - 1 \tag{16.8}$$

Für eine (4 x 3)-Kreuztabelle, wie in unserem Beispiel, beträgt die maximale Inertia damit K=2. Praktisch ist der Wert der Inertia meist viel kleiner als der Maximalwert.

Schritt 2: Extraktion der Dimensionen

Die Aufgabe der Korrespondenzanalyse lässt sich wie folgt präzisieren: Darstellung der Zeilen- und Spalten einer Kreuztabelle in einem gemeinsamen Raum (Korrespondenzraum) mit möglichst geringer Dimensionalität, und zwar so, dass die in den Daten enthaltene Streuung (Information) möglichst weitgehend erhalten bleibt. Dabei muss meist die Sparsamkeit einer Darstellung in wenigen Dimensionen gegen den Verlust an Information abgewogen werden.

Korrespondenzraum

Die maximale Anzahl der Dimensionen für den Korrespondenzraum ist K und entspricht somit der maximalen Inertia. Für unser kleines Beispiel lassen sich daher die Daten in einem zweidimensionalen Raum, wie ihn Abbildung 16.2 zeigt, ohne Informationsverlust darstellen.

Gewöhnlich wird man aber auch bei größeren Kreuztabellen eine zweidimensionale Lösung zwecks anschaulicher Darstellung anstreben. Dies ist dann i. d. R. nicht ohne Informationsverlust möglich. Mit Hilfe der Korrespondenzanalyse ist es aber möglich, diesen Informationsverlust zu minimieren.

Zur Gewinnung der Dimensionen bzw. der Koordinaten der Zeilen- und Spaltenelemente wird die Matrix mit den standardisierten Daten z_{ij} einer *Singulärwertzerlegung* unterzogen.[1] Diese lässt sich in Matrizenschreibweise wie folgt darstellen:

Gewinnung der Koordinaten

$$\mathbf{Z} = \mathbf{U} \cdot \mathbf{S} \cdot \mathbf{V}' \tag{16.9}$$

Dabei bedeuten:

$\mathbf{Z} = (z_{ij})$: (I x J) - Matrix mit den standardisierten Daten
$\mathbf{U} = (u_{ik})$: (I x K) - Matrix für die Zeilenelemente
$\mathbf{V} = (v_{jk})$: (J x K) - Matrix für die Spaltenelemente
$\mathbf{S} = (s_{kk})$: (K x K) - Diagonalmatrix mit den Singulärwerten.

$\mathbf{V}'$ bezeichnet die Transponierte von Matrix $\mathbf{V}$.

Die Matrizen $\mathbf{U}$ und $\mathbf{V}$ sind für unser Beispiel in Abbildung 16.5 dargestellt.

	Dimensionen	
	1	2
Delicado	-0,743	0,170
Rama	0,479	0,545
Becel	-0,028	-0,750
Homa	0,467	-0,335

	Dimensionen	
	1	2
Geschmack	-0,667	0,472
Energie	0,742	0,341
Gesundheit	-0,075	-0,813

Abbildung 16.5: Matrix U (Zeilenelemente) und Matrix V (Spaltenelemente)

[1] Die Singulärwertzerlegung bzw. Singular Value Decomposition (SVD) ähnelt dem Eigenwertverfahren, welches in der Faktorenanalyse für die Zerlegung der Korrelationsmatrix angewendet wird. Im Gegensatz zur Korrelationsmatrix, die quadratisch ist, kann die Matrix $\mathbf{Z}$ auch rechteckig sein. Die Singulärwertzerlegung ist somit ein verallgemeinertes Eigenwertverfahren, welches sich auch auf beliebige nicht quadratische Matrizen anwenden lässt.

Die Matrix **S** enthält in der Diagonalen die Singulärwerte der Dimensionen. Sie lauten hier:

$$s_1 = 0,346 \qquad s_2 = 0,140$$

Trägheitsgewicht

Die quadrierten Singulärwerte sind sog. *Eigenwerte* (Trägheitsgewichte) und liefern ein Maß für die Streuung (Information), die eine Dimension aufnimmt oder repräsentiert. Sie sind in Abbildung 16.2 angegeben. Die Eigenwerte summieren sich zur Inertia:

$$T = \sum_k s_k^2 \tag{16.10}$$

Eigenwertanteil

Der Anteil des quadrierten Singulärwertes einer Dimension an der Inertia ergibt deren *Eigenwertanteil*:

$$EA_k = \frac{s_k^2}{T} \qquad \text{Eigenwertanteil der Dimension k} \tag{16.11}$$

Es ergibt sich hier:

$0,1197/0,1393 = 0,859$ bzw. $85,9\%$	Eigenwertanteil von Dimension 1
$0,0196/0,1393 = 0,141$ bzw. $14,1\%$	Eigenwertanteil von Dimension 2

Singulärwertzerlegung

Die beiden Dimensionen sind orthogonal (rechtwinklig) zueinander und bilden so die Achsen eines rechtwinkligen Koordinatensystems. Durch die Singulärwertzerlegung werden sie derart extrahiert, dass die erste Dimension einen maximalen Anteil der in den Daten vorhandenen Streuung (Information) aufnimmt. Die zweite Dimension nimmt einen maximalen Anteil der noch verbleibenden Streuung auf, usw. Die Wichtigkeit der Dimensionen nimmt somit sukzessiv ab. In Abbildung 16.2 bildet die horizontale Achse die erste Dimension und die vertikale Achse die zweite Dimension. Da hier nur zwei Dimensionen möglich sind, addieren sich die beiden Eigenwertanteile zu 100 %.

Schritt 3: Normalisierung der Koordinaten

Um aus den Matrizen **U** und **V** die endgültigen Koordinaten zu gewinnen, die die grafische Darstellung der Zeilen- und Spaltenelemente in einem gemeinsamen Korrespondenzraum (Biplot) ermöglichen, müssen diese noch *normalisiert* (reskaliert) werden. Dabei werden die Singulärwerte s_k als Gewichte für die Dimensionen und die marginalen relativen Häufigkeiten $p_{i.}$ und $p_{.j}$ zur Gewichtung der Zeilen und Spalten herangezogen:

$$\text{Zeilenpunkte (row points):} \qquad r_{ik} = u_{ik}\sqrt{s_k}/\sqrt{p_{i.}} \tag{16.12}$$

$$\text{Spaltenpunkte (column points):} \qquad c_{jk} = v_{jk}\sqrt{s_k}/\sqrt{p_{.j}} \tag{16.13}$$

Damit ergeben sich die Koordinaten r_{ik} für die Zeilenelemente (Margarinemarken) und c_{jk} für die Spaltenelemente (Margarinemerkmale) in Abbildung 16.6, die Abbildung 16.2 zugrunde liegen.

Row Points	Dimensionen	
r_{ik}	1	2
Delicado	-0,691	0,100
Rama	0,546	0,395
Becel	-0,055	-0,941
Homa	0,555	-0,253

Column Points	Dimensionen	
c_{jk}	1	2
Geschmack	-0,679	0,306
Energie	0,756	0,221
Gesundheit	-0,077	-0,527

Abbildung 16.6: Koordinaten für die Zeilenelemente (Margarinemarken)

Symmetrische Normalisierung

Die hier vorgenommene Normalisierung wird als *symmetrische Normalisierung* bezeichnet. Sie gilt als die klassische Form der Korrespondenzanalyse. Daneben existieren weitere Formen der Normalisierung, wie z. B. die Prinzipal-Normalisierung sowie zwei asymmetrische Formen, die Zeilen-Prinzipal- und die Spalten-Prinzipal-Normalisierung. Im Programm SPSS werden alle diese Formen der Normalisierung als Optionen angeboten. Die verschiedenen Formen der Normalisierung unterscheiden sich dadurch, dass die Inertia unterschiedlich auf die Dimensionen wie auch auf die Zeilen- und Spaltenelemente aufgeteilt wird.

Interpretation

Inhaltliche Interpretation

Zur inhaltlichen Interpretation der Dimensionen kann man sich an den Positionen der Zeilen- und Spaltenelemente im Korrespondenzraum orientieren. Betrachtet man in Abbildung 16.2 für die symmetrische Darstellung die Positionen der drei Merkmale, so sieht man, dass auf der horizontalen Achse insbesondere die Merkmale „Geschmack“ und „Energie“ streuen. Sie sind auf ihr weit links und weit rechts positioniert, während das Merkmal „Gesundheit“ nahezu in der Mitte liegt. Die horizontale Achse repräsentiert also die beiden Merkmale „Geschmack“ und „Energie“, die offenbar gegensätzlich empfunden werden und somit die Polaritäten dieser Achse bilden. Man könnte sie daher als „Geschmack vs. Energie“ benennen. Die vertikale Achse dagegen lässt sich als „Gesundheit“ bezeichnen, da sie nur durch dieses Merkmal geprägt wird.

Zwischen-Gruppen-Distanzen

Bedingt durch die unterschiedlichen Formen der Normalisierung wird die Anwendung und Interpretation der Korrespondenzanalyse nicht gerade erleichtert. Der Vorteil der Korrespondenzanalyse, die Darstellung der Positionen von Zeilen- und Spaltenelementen in einem gemeinsamen Korrespondenzraum (Biplot), bildet auch deren Schwachpunkt. Wenngleich es gängige Praxis ist, auch die Distanzen zwischen den Zeilen- und Spaltenelementen (Zwischen-Gruppen-Distanzen) zu interpretieren, so fehlt hierfür die theoretische Grundlage. Es können sich daher leicht Missverständnisse ergeben. Gerechtfertigt ist (zumindest bei symmetrischer Normalisierung) nur die Interpretation der Distanzen innerhalb der beiden Gruppen, also der Distanzen zwischen den Zeilenelementen und der Distanzen zwischen den Spaltenelementen. Man vermeidet diese Probleme, wenn man getrennte Plots für die Zeilen- und Spaltenelemente erstellt.

16.3 Umsetzung mit SPSS

Casewise- und Weight- Format

Zur Durchführung von Korrespondenzanalysen bietet SPSS im Modul Categories die Prozedur CORRESPONDENCE an. Dabei kann der Nutzer zwischen zwei Datenformaten wählen:

- „Casewise“-Format für eine fallweise Dateneingabe,
- „Weight“-Format für eine aggregierte Dateneingabe.

Das „Casewise“-Format erfordert, dass die Daten fallweise, d. h. einzeln für jede Person, eingegeben werden. Praktischer ist daher das „Weight“-Format, das eine aggregierte Eingabe der Häufigkeiten ermöglicht. Die Häufigkeiten können hierzu aus der Kreuztabelle entnommen werden. Abbildung 16.8 zeigt die Daten des Beispiels im „Weight“-Format.

Mittels der beiden kategorialen Variablen „Margarine-Merkmal“ und „Margarinemarke“ werden die Merkmalskombinationen spezifiziert, die den Zellen der Kreuztabelle entsprechen. Die Häufigkeiten der Kreuztabelle werden mittels der Variable „Fallzahl“ eingegeben. Sie lassen sich sodann den Kombinationen als Gewichte zuordnen. Dazu ist vor Aufruf der Korrespondenzanalyse der Menüpunkt „Daten/Fälle gewichten“ aufzurufen(Abbildung 16.7).

Abbildung 16.7: Dialogfenster „Fälle gewichten“

Die Korrespondenzanalyse wird über den Menüpunkt „Analysieren/ Dimensionsreduktion/ Korrespondenzanalyse“ aufgerufen (Abbildung 16.8).

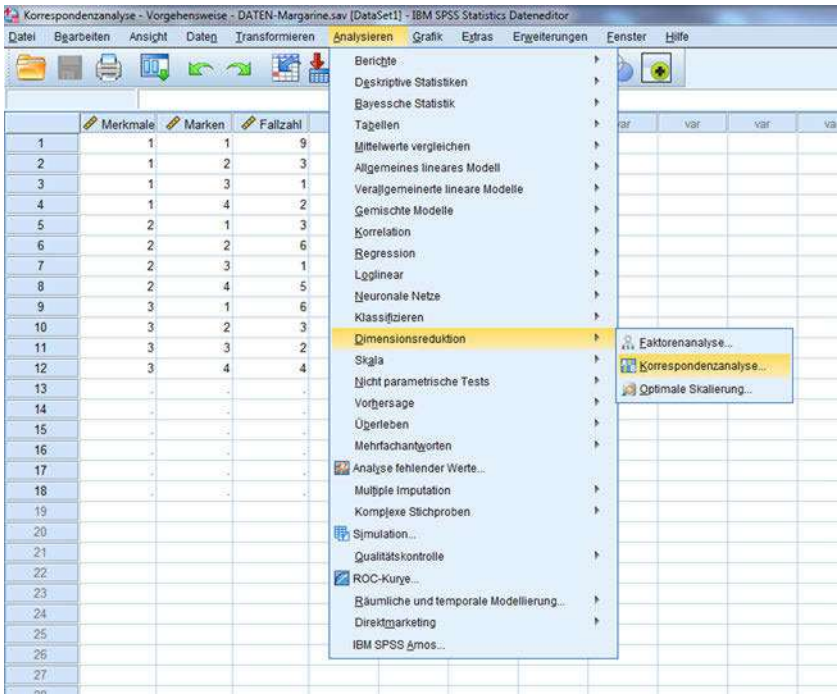


Abbildung 16.8: Daten-Editor mit Daten im „Weight“-Format und Menüpunkt „Dimensionsreduktion / Korrespondenzanalyse“

A Tabellenanhang

K. Backhaus, B. Erichson, W. Plinke, R. Weiber, *Multivariate Analysemethoden*,
https://doi.org/10.1007/978-3-662-56655-8

A.1 t-Tabelle

t-Tabelle

FG	Irrtumswahrscheinlichkeit für den zweiseitigen Test α								
	0,5	0,2	0,1	0,05	0,02	0,01	0,002	0,001	0,0001
1	1,000	3,078	6,314	12,706	31,821	63,657	318,309	636,619	6366,198
2	0,816	1,886	2,920	4,303	6,965	9,925	22,327	31,599	99,993
3	0,765	1,638	2,353	3,182	4,541	5,841	10,215	12,924	28,000
4	0,741	1,533	2,132	2,776	3,747	4,604	7,173	8,610	15,544
5	0,727	1,476	2,015	2,571	3,365	4,032	5,893	6,869	11,178
6	0,718	1,440	1,943	2,447	3,143	3,707	5,208	5,959	9,082
7	0,711	1,415	1,895	2,365	2,998	3,499	4,785	5,408	7,885
8	0,706	1,397	1,860	2,306	2,896	3,355	4,501	5,041	7,120
9	0,703	1,383	1,833	2,262	2,821	3,250	4,297	4,781	6,594
10	0,700	1,372	1,812	2,228	2,764	3,169	4,144	4,587	6,211
11	0,697	1,363	1,796	2,201	2,718	3,106	4,025	4,437	5,921
12	0,695	1,356	1,782	2,179	2,681	3,055	3,930	4,318	5,694
13	0,694	1,350	1,771	2,160	2,650	3,012	3,852	4,221	5,513
14	0,692	1,345	1,761	2,145	2,624	2,977	3,787	4,140	5,363
15	0,691	1,341	1,753	2,131	2,602	2,947	3,733	4,073	5,239
16	0,690	1,337	1,746	2,120	2,583	2,921	3,686	4,015	5,134
17	0,689	1,333	1,740	2,110	2,567	2,898	3,646	3,965	5,044
18	0,688	1,330	1,734	2,101	2,552	2,878	3,610	3,922	4,966
19	0,688	1,328	1,729	2,093	2,539	2,861	3,579	3,883	4,897
20	0,687	1,325	1,725	2,086	2,528	2,845	3,552	3,850	4,837
21	0,686	1,323	1,721	2,080	2,518	2,831	3,527	3,819	4,784
22	0,686	1,321	1,717	2,074	2,508	2,819	3,505	3,792	4,736
23	0,685	1,319	1,714	2,069	2,500	2,807	3,485	3,768	4,693
24	0,685	1,318	1,711	2,064	2,492	2,797	3,467	3,745	4,654
25	0,684	1,316	1,708	2,060	2,485	2,787	3,450	3,725	4,619
26	0,684	1,315	1,706	2,056	2,479	2,779	3,435	3,707	4,587
27	0,684	1,314	1,703	2,052	2,473	2,771	3,421	3,690	4,558
28	0,683	1,313	1,701	2,048	2,467	2,763	3,408	3,674	4,530
29	0,683	1,311	1,699	2,045	2,462	2,756	3,396	3,659	4,506
30	0,683	1,310	1,697	2,042	2,457	2,750	3,385	3,646	4,482
32	0,682	1,309	1,694	2,037	2,449	2,738	3,365	3,622	4,441
34	0,682	1,307	1,691	2,032	2,441	2,728	3,348	3,601	4,405
36	0,681	1,306	1,688	2,028	2,434	2,719	3,333	3,582	4,374
38	0,681	1,304	1,686	2,024	2,429	2,712	3,319	3,566	4,346
40	0,681	1,303	1,684	2,021	2,423	2,704	3,307	3,551	4,321
45	0,680	1,301	1,679	2,014	2,412	2,690	3,281	3,520	4,269
50	0,679	1,299	1,676	2,009	2,403	2,678	3,261	3,496	4,228
55	0,679	1,297	1,673	2,004	2,396	2,668	3,245	3,476	4,196
60	0,679	1,296	1,671	2,000	2,390	2,660	3,232	3,460	4,169
70	0,678	1,294	1,667	1,994	2,381	2,648	3,211	3,435	4,127
80	0,678	1,292	1,664	1,990	2,374	2,639	3,195	3,416	4,096
90	0,677	1,291	1,662	1,987	2,368	2,632	3,183	3,402	4,072
100	0,677	1,290	1,660	1,984	2,364	2,626	3,174	3,390	4,053
120	0,677	1,289	1,658	1,980	2,358	2,617	3,160	3,373	4,025
200	0,676	1,286	1,653	1,972	2,345	2,601	3,131	3,340	3,970
500	0,675	1,283	1,648	1,965	2,334	2,586	3,107	3,310	3,922
1000	0,675	1,282	1,646	1,962	2,330	2,581	3,098	3,300	3,906
ue	0,674	1,282	1,645	1,960	2,326	2,576	3,090	3,291	3,891
FG	0,25	0,1	0,05	0,025	0,01	0,005	0,001	0,0005	0,00005
	Irrtumswahrscheinlichkeit für den einseitigen Test α								

α = Signifikanzniveau (1-Vertrauenswahrscheinlichkeit)
FG = Freiheitsgrade

Abbildung A.1: t-Tabelle

A.2 F-Tabelle

F-Tabelle (Vertrauenswahrscheinlichkeit 0,9)

v_2 \ v_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	10000
1	39,86	49,50	53,59	55,83	57,24	58,20	58,91	59,44	59,86	60,19	60,71	61,22	61,74	62,00	62,26	62,53	62,79	63,06	63,32
2	8,53	9,00	9,16	9,24	9,29	9,33	9,35	9,37	9,38	9,39	9,41	9,42	9,44	9,45	9,46	9,47	9,47	9,48	9,49
3	5,54	5,46	5,39	5,34	5,31	5,28	5,27	5,25	5,24	5,23	5,22	5,20	5,18	5,18	5,17	5,16	5,15	5,14	5,13
4	4,54	4,32	4,19	4,11	4,05	4,01	3,98	3,95	3,94	3,92	3,90	3,87	3,84	3,83	3,82	3,80	3,79	3,78	3,76
5	4,06	3,78	3,62	3,52	3,45	3,40	3,37	3,34	3,32	3,30	3,27	3,24	3,21	3,19	3,17	3,16	3,14	3,12	3,11
6	3,78	3,46	3,29	3,18	3,11	3,05	3,01	2,98	2,96	2,94	2,90	2,87	2,84	2,82	2,80	2,78	2,76	2,74	2,72
7	3,59	3,26	3,07	2,96	2,88	2,83	2,78	2,75	2,72	2,70	2,67	2,63	2,59	2,58	2,56	2,54	2,51	2,49	2,47
8	3,46	3,11	2,92	2,81	2,73	2,67	2,62	2,59	2,56	2,54	2,50	2,46	2,42	2,40	2,38	2,36	2,34	2,32	2,29
9	3,36	3,01	2,81	2,69	2,61	2,55	2,51	2,47	2,44	2,42	2,38	2,34	2,30	2,28	2,25	2,23	2,21	2,18	2,16
10	3,29	2,92	2,73	2,61	2,52	2,46	2,41	2,38	2,35	2,32	2,28	2,24	2,20	2,18	2,16	2,13	2,11	2,08	2,06
11	3,23	2,86	2,66	2,54	2,45	2,39	2,34	2,30	2,27	2,25	2,21	2,17	2,12	2,10	2,08	2,05	2,03	2,00	1,97
12	3,18	2,81	2,61	2,48	2,39	2,33	2,28	2,24	2,21	2,19	2,15	2,10	2,06	2,04	2,01	1,99	1,96	1,93	1,90
13	3,14	2,76	2,56	2,43	2,35	2,28	2,23	2,20	2,16	2,14	2,10	2,05	2,01	1,98	1,96	1,93	1,90	1,88	1,85
14	3,10	2,73	2,52	2,39	2,31	2,24	2,19	2,15	2,12	2,10	2,05	2,01	1,96	1,94	1,91	1,89	1,86	1,83	1,80
15	3,07	2,70	2,49	2,36	2,27	2,21	2,16	2,12	2,09	2,06	2,02	1,97	1,92	1,90	1,87	1,85	1,82	1,79	1,76
16	3,05	2,67	2,46	2,33	2,24	2,18	2,13	2,09	2,06	2,03	1,99	1,94	1,89	1,87	1,84	1,81	1,78	1,75	1,72
17	3,03	2,64	2,44	2,31	2,22	2,15	2,10	2,06	2,03	2,00	1,96	1,91	1,86	1,84	1,81	1,78	1,75	1,72	1,69
18	3,01	2,62	2,42	2,29	2,20	2,13	2,08	2,04	2,00	1,98	1,93	1,89	1,84	1,81	1,78	1,75	1,72	1,69	1,66
19	2,99	2,61	2,40	2,27	2,18	2,11	2,06	2,02	1,98	1,96	1,91	1,86	1,81	1,79	1,76	1,73	1,70	1,67	1,63
20	2,97	2,59	2,38	2,25	2,16	2,09	2,04	2,00	1,96	1,94	1,89	1,84	1,79	1,77	1,74	1,71	1,68	1,64	1,61
21	2,96	2,57	2,36	2,23	2,14	2,08	2,02	1,98	1,95	1,92	1,87	1,83	1,78	1,75	1,72	1,69	1,66	1,62	1,59
22	2,95	2,56	2,35	2,22	2,13	2,06	2,01	1,97	1,93	1,90	1,86	1,81	1,76	1,73	1,70	1,67	1,64	1,60	1,57
23	2,94	2,55	2,34	2,21	2,11	2,05	1,99	1,95	1,92	1,89	1,84	1,80	1,74	1,72	1,69	1,66	1,62	1,59	1,55
24	2,93	2,54	2,33	2,19	2,10	2,04	1,98	1,94	1,91	1,88	1,83	1,78	1,73	1,70	1,67	1,64	1,61	1,57	1,53
25	2,92	2,53	2,32	2,18	2,09	2,02	1,97	1,93	1,89	1,87	1,82	1,77	1,72	1,69	1,66	1,63	1,59	1,56	1,52
26	2,91	2,52	2,31	2,17	2,08	2,01	1,96	1,92	1,88	1,86	1,81	1,76	1,71	1,68	1,65	1,61	1,58	1,54	1,50
27	2,90	2,51	2,30	2,17	2,07	2,00	1,95	1,91	1,87	1,85	1,80	1,75	1,70	1,67	1,64	1,60	1,57	1,53	1,49
28	2,89	2,50	2,29	2,16	2,06	2,00	1,94	1,90	1,87	1,84	1,79	1,74	1,69	1,66	1,63	1,59	1,56	1,52	1,48
29	2,89	2,50	2,28	2,15	2,06	1,99	1,93	1,89	1,86	1,83	1,78	1,73	1,68	1,65	1,62	1,58	1,55	1,51	1,47
30	2,88	2,49	2,28	2,14	2,05	1,98	1,93	1,88	1,85	1,82	1,77	1,72	1,67	1,64	1,61	1,57	1,54	1,50	1,46
40	2,84	2,44	2,23	2,09	2,00	1,93	1,87	1,83	1,79	1,76	1,71	1,66	1,61	1,57	1,54	1,51	1,47	1,42	1,38
60	2,79	2,39	2,18	2,04	1,95	1,87	1,82	1,77	1,74	1,71	1,66	1,60	1,54	1,51	1,48	1,44	1,40	1,35	1,29
120	2,75	2,35	2,13	1,99	1,90	1,82	1,77	1,72	1,68	1,65	1,60	1,55	1,48	1,45	1,41	1,37	1,32	1,26	1,19
10000	2,71	2,30	2,08	1,95	1,85	1,77	1,72	1,67	1,63	1,60	1,55	1,49	1,42	1,38	1,34	1,30	1,24	1,17	1,03

v1 = Zahl der erklärenden Variablen (J)
v2 = Zahl der Freiheitsgrade des Nenners (K - J - 1)

Abbildung A.2: F-Tabelle (Vertrauenswahrscheinlichkeit 0,9)

F-Tabelle (Vertrauenswahrscheinlichkeit 0,95)

v_2 \ v_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	10000
1	161,45	199,50	215,71	224,58	230,16	233,99	236,77	238,88	240,54	241,88	243,91	245,95	248,01	249,05	250,10	251,14	252,20	253,25	254,30
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38	19,40	19,41	19,43	19,45	19,45	19,46	19,47	19,48	19,49	19,50
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79	8,74	8,70	8,66	8,64	8,62	8,59	8,57	8,55	8,53
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,91	5,86	5,80	5,77	5,75	5,72	5,69	5,66	5,63
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,68	4,62	4,56	4,53	4,50	4,46	4,43	4,40	4,37
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,00	3,94	3,87	3,84	3,81	3,77	3,74	3,70	3,67
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,57	3,51	3,44	3,41	3,38	3,34	3,30	3,27	3,23
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,28	3,22	3,15	3,12	3,08	3,04	3,01	2,97	2,93
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	3,07	3,01	2,94	2,90	2,86	2,83	2,79	2,75	2,71
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,91	2,85	2,77	2,74	2,70	2,66	2,62	2,58	2,54
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85	2,79	2,72	2,65	2,61	2,57	2,53	2,49	2,45	2,41
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75	2,69	2,62	2,54	2,51	2,47	2,43	2,38	2,34	2,30
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67	2,60	2,53	2,46	2,42	2,38	2,34	2,30	2,25	2,21
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60	2,53	2,46	2,39	2,35	2,31	2,27	2,22	2,18	2,13
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54	2,48	2,40	2,33	2,29	2,25	2,20	2,16	2,11	2,07
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,42	2,35	2,28	2,24	2,19	2,15	2,11	2,06	2,01
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45	2,38	2,31	2,23	2,19	2,15	2,10	2,06	2,01	1,96
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41	2,34	2,27	2,19	2,15	2,11	2,06	2,02	1,97	1,92
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38	2,31	2,23	2,16	2,11	2,07	2,03	1,98	1,93	1,88
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35	2,28	2,20	2,12	2,08	2,04	1,99	1,95	1,90	1,84
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32	2,25	2,18	2,10	2,05	2,01	1,96	1,92	1,87	1,81
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30	2,23	2,15	2,07	2,03	1,98	1,94	1,89	1,84	1,78
23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,32	2,27	2,20	2,13	2,05	2,01	1,96	1,91	1,86	1,81	1,76
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25	2,18	2,11	2,03	1,98	1,94	1,89	1,84	1,79	1,73
25	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24	2,16	2,09	2,01	1,96	1,92	1,87	1,82	1,77	1,71
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22	2,15	2,07	1,99	1,95	1,90	1,85	1,80	1,75	1,69
27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,31	2,25	2,20	2,13	2,06	1,97	1,93	1,88	1,84	1,79	1,73	1,67
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19	2,12	2,04	1,96	1,91	1,87	1,82	1,77	1,71	1,65
29	4,18	3,33	2,93	2,70	2,55	2,43	2,35	2,28	2,22	2,18	2,10	2,03	1,94	1,90	1,85	1,81	1,75	1,70	1,64
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16	2,09	2,01	1,93	1,89	1,84	1,79	1,74	1,68	1,62
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08	2,00	1,92	1,84	1,79	1,74	1,69	1,64	1,58	1,51
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99	1,92	1,84	1,75	1,70	1,65	1,59	1,53	1,47	1,39
120	3,92	3,07	2,68	2,45	2,29	2,18	2,09	2,02	1,96	1,91	1,83	1,75	1,66	1,61	1,55	1,50	1,43	1,35	1,26
10000	3,84	3,00	2,61	2,37	2,21	2,10	2,01	1,94	1,88	1,83	1,75	1,67	1,57	1,52	1,46	1,40	1,32	1,22	1,03

v1 = Zahl der erklärenden Variablen (J)
v2 = Zahl der Freiheitsgrade des Nenners (K - J - 1)

Abbildung A.3: F-Tabelle (Vertrauenswahrscheinlichkeit 0,95)

F-Tabelle (Vertrauenswahrscheinlichkeit 0,975)

v_2 \ v_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	10000
1	647,79	799,50	864,16	899,58	921,85	937,11	948,22	956,66	963,28	968,63	976,71	984,87	993,10	997,25	1001,41	1005,60	1009,80	1014,02	1018,21
2	38,51	39,00	39,17	39,25	39,30	39,33	39,36	39,37	39,39	39,40	39,41	39,43	39,45	39,46	39,46	39,47	39,48	39,49	39,50
3	17,44	16,04	15,44	15,10	14,88	14,73	14,62	14,54	14,47	14,42	14,34	14,25	14,17	14,12	14,08	14,04	13,99	13,95	13,90
4	12,22	10,65	9,98	9,60	9,36	9,20	9,07	8,98	8,90	8,84	8,75	8,66	8,56	8,51	8,46	8,41	8,36	8,31	8,26
5	10,01	8,43	7,76	7,39	7,15	6,98	6,85	6,76	6,68	6,62	6,52	6,43	6,33	6,28	6,23	6,18	6,12	6,07	6,02
6	8,81	7,26	6,60	6,23	5,99	5,82	5,70	5,60	5,52	5,46	5,37	5,27	5,17	5,12	5,07	5,01	4,96	4,90	4,85
7	8,07	6,54	5,89	5,52	5,29	5,12	4,99	4,90	4,82	4,76	4,67	4,57	4,47	4,41	4,36	4,31	4,25	4,20	4,14
8	7,57	6,06	5,42	5,05	4,82	4,65	4,53	4,43	4,36	4,30	4,20	4,10	4,00	3,95	3,89	3,84	3,78	3,73	3,67
9	7,21	5,71	5,08	4,72	4,48	4,32	4,20	4,10	4,03	3,96	3,87	3,77	3,67	3,61	3,56	3,51	3,45	3,39	3,33
10	6,94	5,46	4,83	4,47	4,24	4,07	3,95	3,85	3,78	3,72	3,62	3,52	3,42	3,37	3,31	3,26	3,20	3,14	3,08
11	6,72	5,26	4,63	4,28	4,04	3,88	3,76	3,66	3,59	3,53	3,43	3,33	3,23	3,17	3,12	3,06	3,00	2,94	2,88
12	6,55	5,10	4,47	4,12	3,89	3,73	3,61	3,51	3,44	3,37	3,28	3,18	3,07	3,02	2,96	2,91	2,85	2,79	2,73
13	6,41	4,97	4,35	4,00	3,77	3,60	3,48	3,39	3,31	3,25	3,15	3,05	2,95	2,89	2,84	2,78	2,72	2,66	2,60
14	6,30	4,86	4,24	3,89	3,66	3,50	3,38	3,29	3,21	3,15	3,05	2,95	2,84	2,79	2,73	2,67	2,61	2,55	2,49
15	6,20	4,77	4,15	3,80	3,58	3,41	3,29	3,20	3,12	3,06	2,96	2,86	2,76	2,70	2,64	2,59	2,52	2,46	2,40
16	6,12	4,69	4,08	3,73	3,50	3,34	3,22	3,12	3,05	2,99	2,89	2,79	2,68	2,63	2,57	2,51	2,45	2,38	2,32
17	6,04	4,62	4,01	3,66	3,44	3,28	3,16	3,06	2,98	2,92	2,82	2,72	2,62	2,56	2,50	2,44	2,38	2,32	2,25
18	5,98	4,56	3,95	3,61	3,38	3,22	3,10	3,01	2,93	2,87	2,77	2,67	2,56	2,50	2,44	2,38	2,32	2,26	2,19
19	5,92	4,51	3,90	3,56	3,33	3,17	3,05	2,96	2,88	2,82	2,72	2,62	2,51	2,45	2,39	2,33	2,27	2,20	2,13
20	5,87	4,46	3,86	3,51	3,29	3,13	3,01	2,91	2,84	2,77	2,68	2,57	2,46	2,41	2,35	2,29	2,22	2,16	2,09
21	5,83	4,42	3,82	3,48	3,25	3,09	2,97	2,87	2,80	2,73	2,64	2,53	2,42	2,37	2,31	2,25	2,18	2,11	2,04
22	5,79	4,38	3,78	3,44	3,22	3,05	2,93	2,84	2,76	2,70	2,60	2,50	2,39	2,33	2,27	2,21	2,14	2,08	2,00
23	5,75	4,35	3,75	3,41	3,18	3,02	2,90	2,81	2,73	2,67	2,57	2,47	2,36	2,30	2,24	2,18	2,11	2,04	1,97
24	5,72	4,32	3,72	3,38	3,15	2,99	2,87	2,78	2,70	2,64	2,54	2,44	2,33	2,27	2,21	2,15	2,08	2,01	1,94
25	5,69	4,29	3,69	3,35	3,13	2,97	2,85	2,75	2,68	2,61	2,51	2,41	2,30	2,24	2,18	2,12	2,05	1,98	1,91
26	5,66	4,27	3,67	3,33	3,10	2,94	2,82	2,73	2,65	2,59	2,49	2,39	2,28	2,22	2,16	2,09	2,03	1,95	1,88
27	5,63	4,24	3,65	3,31	3,08	2,92	2,80	2,71	2,63	2,57	2,47	2,36	2,25	2,19	2,13	2,07	2,00	1,93	1,85
28	5,61	4,22	3,63	3,29	3,06	2,90	2,78	2,69	2,61	2,55	2,45	2,34	2,23	2,17	2,11	2,05	1,98	1,91	1,83
29	5,59	4,20	3,61	3,27	3,04	2,88	2,76	2,67	2,59	2,53	2,43	2,32	2,21	2,15	2,09	2,03	1,96	1,89	1,81
30	5,57	4,18	3,59	3,25	3,03	2,87	2,75	2,65	2,57	2,51	2,41	2,31	2,20	2,14	2,07	2,01	1,94	1,87	1,79
40	5,42	4,05	3,46	3,13	2,90	2,74	2,62	2,53	2,45	2,39	2,29	2,18	2,07	2,01	1,94	1,88	1,80	1,72	1,64
60	5,29	3,93	3,34	3,01	2,79	2,63	2,51	2,41	2,33	2,27	2,17	2,06	1,94	1,88	1,82	1,74	1,67	1,58	1,48
120	5,15	3,80	3,23	2,89	2,67	2,52	2,39	2,30	2,22	2,16	2,05	1,94	1,82	1,76	1,69	1,61	1,53	1,43	1,31
10000	5,03	3,69	3,12	2,79	2,57	2,41	2,29	2,19	2,11	2,05	1,95	1,83	1,71	1,64	1,57	1,49	1,39	1,27	1,04

v1 = Zahl der erklärenden Variablen (J)
v2 = Zahl der Freiheitsgrade des Nenners (K - J - 1)

Abbildung A.4: F-Tabelle (Vertrauenswahrscheinlichkeit 0,975)

F-Tabelle (Vertrauenswahrscheinlichkeit 0,99)

v_2 \ v_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	10000
1	4052,18	4999,50	5403,35	5624,58	5763,65	5858,99	5928,36	5981,07	6022,47	6055,85	6106,32	6157,28	6208,73	6234,63	6260,65	6286,78	6313,03	6339,39	6365,55
2	98,50	99,00	99,17	99,25	99,30	99,33	99,36	99,37	99,39	99,40	99,42	99,43	99,45	99,46	99,47	99,47	99,48	99,49	99,50
3	34,12	30,82	29,46	28,71	28,24	27,91	27,67	27,49	27,35	27,23	27,05	26,87	26,69	26,60	26,50	26,41	26,32	26,22	26,13
4	21,20	18,00	16,69	15,98	15,52	15,21	14,98	14,80	14,66	14,55	14,37	14,20	14,02	13,93	13,84	13,75	13,65	13,56	13,46
5	16,26	13,27	12,06	11,39	10,97	10,67	10,46	10,29	10,16	10,05	9,89	9,72	9,55	9,47	9,38	9,29	9,20	9,11	9,02
6	13,75	10,92	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87	7,72	7,56	7,40	7,31	7,23	7,14	7,06	6,97	6,88
7	12,25	9,55	8,45	7,85	7,46	7,19	6,99	6,84	6,72	6,62	6,47	6,31	6,16	6,07	5,99	5,91	5,82	5,74	5,65
8	11,26	8,65	7,59	7,01	6,63	6,37	6,18	6,03	5,91	5,81	5,67	5,52	5,36	5,28	5,20	5,12	5,03	4,95	4,86
9	10,56	8,02	6,99	6,42	6,06	5,80	5,61	5,47	5,35	5,26	5,11	4,96	4,81	4,73	4,65	4,57	4,48	4,40	4,31
10	10,04	7,56	6,55	5,99	5,64	5,39	5,20	5,06	4,94	4,85	4,71	4,56	4,41	4,33	4,25	4,17	4,08	4,00	3,91
11	9,65	7,21	6,22	5,67	5,32	5,07	4,89	4,74	4,63	4,54	4,40	4,25	4,10	4,02	3,94	3,86	3,78	3,69	3,60
12	9,33	6,93	5,95	5,41	5,06	4,82	4,64	4,50	4,39	4,30	4,16	4,01	3,86	3,78	3,70	3,62	3,54	3,45	3,36
13	9,07	6,70	5,74	5,21	4,86	4,62	4,44	4,30	4,19	4,10	3,96	3,82	3,66	3,59	3,51	3,43	3,34	3,25	3,17
14	8,86	6,51	5,56	5,04	4,69	4,46	4,28	4,14	4,03	3,94	3,80	3,66	3,51	3,43	3,35	3,27	3,18	3,09	3,01
15	8,68	6,36	5,42	4,89	4,56	4,32	4,14	4,00	3,89	3,80	3,67	3,52	3,37	3,29	3,21	3,13	3,05	2,96	2,87
16	8,53	6,23	5,29	4,77	4,44	4,20	4,03	3,89	3,78	3,69	3,55	3,41	3,26	3,18	3,10	3,02	2,93	2,84	2,75
17	8,40	6,11	5,19	4,67	4,34	4,10	3,93	3,79	3,68	3,59	3,46	3,31	3,16	3,08	3,00	2,92	2,83	2,75	2,65
18	8,29	6,01	5,09	4,58	4,25	4,01	3,84	3,71	3,60	3,51	3,37	3,23	3,08	3,00	2,92	2,84	2,75	2,66	2,57
19	8,18	5,93	5,01	4,50	4,17	3,94	3,77	3,63	3,52	3,43	3,30	3,15	3,00	2,92	2,84	2,76	2,67	2,58	2,49
20	8,10	5,85	4,94	4,43	4,10	3,87	3,70	3,56	3,46	3,37	3,23	3,09	2,94	2,86	2,78	2,69	2,61	2,52	2,42
21	8,02	5,78	4,87	4,37	4,04	3,81	3,64	3,51	3,40	3,31	3,17	3,03	2,88	2,80	2,72	2,64	2,55	2,46	2,36
22	7,95	5,72	4,82	4,31	3,99	3,76	3,59	3,45	3,35	3,26	3,12	2,98	2,83	2,75	2,67	2,58	2,50	2,40	2,31
23	7,88	5,66	4,76	4,26	3,94	3,71	3,54	3,41	3,30	3,21	3,07	2,93	2,78	2,70	2,62	2,54	2,45	2,35	2,26
24	7,82	5,61	4,72	4,22	3,90	3,67	3,50	3,36	3,26	3,17	3,03	2,89	2,74	2,66	2,58	2,49	2,40	2,31	2,21
25	7,77	5,57	4,68	4,18	3,85	3,63	3,46	3,32	3,22	3,13	2,99	2,85	2,70	2,62	2,54	2,45	2,36	2,27	2,17
26	7,72	5,53	4,64	4,14	3,82	3,59	3,42	3,29	3,18	3,09	2,96	2,81	2,66	2,58	2,50	2,42	2,33	2,23	2,13
27	7,68	5,49	4,60	4,11	3,78	3,56	3,39	3,26	3,15	3,06	2,93	2,78	2,63	2,55	2,47	2,38	2,29	2,20	2,10
28	7,64	5,45	4,57	4,07	3,75	3,53	3,36	3,23	3,12	3,03	2,90	2,75	2,60	2,52	2,44	2,35	2,26	2,17	2,07
29	7,60	5,42	4,54	4,04	3,73	3,50	3,33	3,20	3,09	3,00	2,87	2,73	2,57	2,49	2,41	2,33	2,23	2,14	2,04
30	7,56	5,39	4,51	4,02	3,70	3,47	3,30	3,17	3,07	2,98	2,84	2,70	2,55	2,47	2,39	2,30	2,21	2,11	2,01
40	7,31	5,18	4,31	3,83	3,51	3,29	3,12	2,99	2,89	2,80	2,66	2,52	2,37	2,29	2,20	2,11	2,02	1,92	1,81
60	7,08	4,98	4,13	3,65	3,34	3,12	2,95	2,82	2,72	2,63	2,50	2,35	2,20	2,12	2,03	1,94	1,84	1,73	1,60
120	6,85	4,79	3,95	3,48	3,17	2,96	2,79	2,66	2,56	2,47	2,34	2,19	2,03	1,95	1,86	1,76	1,66	1,53	1,38
10000	6,64	4,61	3,78	3,32	3,02	2,80	2,64	2,51	2,41	2,32	2,19	2,04	1,88	1,79	1,70	1,59	1,48	1,33	1,05

v1 = Zahl der erklärenden Variablen (J)
v2 = Zahl der Freiheitsgrade des Nenners (K - J - 1)

Abbildung A.5: F-Tabelle (Vertrauenswahrscheinlichkeit 0,99)

F-Tabelle (Vertrauenswahrscheinlichkeit 0,995)

v2 \ v1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	10000
1	16210,72	19999,50	21614,74	22499,58	23055,80	23437,11	23714,57	23925,41	24091,00	24224,49	24426,37	24630,21	24835,97	24939,57	25043,63	25148,15	25253,14	25358,57	25463,18
2	198,50	199,00	199,17	199,25	199,30	199,33	199,36	199,37	199,39	199,40	199,42	199,43	199,45	199,46	199,47	199,47	199,48	199,49	199,50
3	55,55	49,80	47,47	46,19	45,39	44,84	44,43	44,13	43,88	43,69	43,39	43,08	42,78	42,62	42,47	42,31	42,15	41,99	41,83
4	31,33	26,28	24,26	23,15	22,46	21,97	21,62	21,35	21,14	20,97	20,70	20,44	20,17	20,03	19,89	19,75	19,61	19,47	19,33
5	22,78	18,31	16,53	15,56	14,94	14,51	14,20	13,96	13,77	13,62	13,38	13,15	12,90	12,78	12,66	12,53	12,40	12,27	12,15
6	18,63	14,54	12,92	12,03	11,46	11,07	10,79	10,57	10,39	10,25	10,03	9,81	9,59	9,47	9,36	9,24	9,12	9,00	8,88
7	16,24	12,40	10,88	10,05	9,52	9,16	8,89	8,68	8,51	8,38	8,18	7,97	7,75	7,64	7,53	7,42	7,31	7,19	7,08
8	14,69	11,04	9,60	8,81	8,30	7,95	7,69	7,50	7,34	7,21	7,01	6,81	6,61	6,50	6,40	6,29	6,18	6,06	5,95
9	13,61	10,11	8,72	7,96	7,47	7,13	6,88	6,69	6,54	6,42	6,23	6,03	5,83	5,73	5,62	5,52	5,41	5,30	5,19
10	12,83	9,43	8,08	7,34	6,87	6,54	6,30	6,12	5,97	5,85	5,66	5,47	5,27	5,17	5,07	4,97	4,86	4,75	4,64
11	12,23	8,91	7,60	6,88	6,42	6,10	5,86	5,68	5,54	5,42	5,24	5,05	4,86	4,76	4,65	4,55	4,45	4,34	4,23
12	11,75	8,51	7,23	6,52	6,07	5,76	5,52	5,35	5,20	5,09	4,91	4,72	4,53	4,43	4,33	4,23	4,12	4,01	3,91
13	11,37	8,19	6,93	6,23	5,79	5,48	5,25	5,08	4,94	4,82	4,64	4,46	4,27	4,17	4,07	3,97	3,87	3,76	3,65
14	11,06	7,92	6,68	6,00	5,56	5,26	5,03	4,86	4,72	4,60	4,43	4,25	4,06	3,96	3,86	3,76	3,66	3,55	3,44
15	10,80	7,70	6,48	5,80	5,37	5,07	4,85	4,67	4,54	4,42	4,25	4,07	3,88	3,79	3,69	3,58	3,48	3,37	3,26
16	10,58	7,51	6,30	5,64	5,21	4,91	4,69	4,52	4,38	4,27	4,10	3,92	3,73	3,64	3,54	3,44	3,33	3,22	3,11
17	10,38	7,35	6,16	5,50	5,07	4,78	4,56	4,39	4,25	4,14	3,97	3,79	3,61	3,51	3,41	3,31	3,21	3,10	2,99
18	10,22	7,21	6,03	5,37	4,96	4,66	4,44	4,28	4,14	4,03	3,86	3,68	3,50	3,40	3,30	3,20	3,10	2,99	2,87
19	10,07	7,09	5,92	5,27	4,85	4,56	4,34	4,18	4,04	3,93	3,76	3,59	3,40	3,31	3,21	3,11	3,00	2,89	2,78
20	9,94	6,99	5,82	5,17	4,76	4,47	4,26	4,09	3,96	3,85	3,68	3,50	3,32	3,22	3,12	3,02	2,92	2,81	2,69
21	9,83	6,89	5,73	5,09	4,68	4,39	4,18	4,01	3,88	3,77	3,60	3,43	3,24	3,15	3,05	2,95	2,84	2,73	2,62
22	9,73	6,81	5,65	5,02	4,61	4,32	4,11	3,94	3,81	3,70	3,54	3,36	3,18	3,08	2,98	2,88	2,77	2,66	2,55
23	9,63	6,73	5,58	4,95	4,54	4,26	4,05	3,88	3,75	3,64	3,47	3,30	3,12	3,02	2,92	2,82	2,71	2,60	2,49
24	9,55	6,66	5,52	4,89	4,49	4,20	3,99	3,83	3,69	3,59	3,42	3,25	3,06	2,97	2,87	2,77	2,66	2,55	2,43
25	9,48	6,60	5,46	4,84	4,43	4,15	3,94	3,78	3,64	3,54	3,37	3,20	3,01	2,92	2,82	2,72	2,61	2,50	2,38
26	9,41	6,54	5,41	4,79	4,38	4,10	3,89	3,73	3,60	3,49	3,33	3,15	2,97	2,87	2,77	2,67	2,56	2,45	2,33
27	9,34	6,49	5,36	4,74	4,34	4,06	3,85	3,69	3,56	3,45	3,28	3,11	2,93	2,83	2,73	2,63	2,52	2,41	2,29
28	9,28	6,44	5,32	4,70	4,30	4,02	3,81	3,65	3,52	3,41	3,25	3,07	2,89	2,79	2,69	2,59	2,48	2,37	2,25
29	9,23	6,40	5,28	4,66	4,26	3,98	3,77	3,61	3,48	3,38	3,21	3,04	2,86	2,76	2,66	2,56	2,45	2,33	2,21
30	9,18	6,35	5,24	4,62	4,23	3,95	3,74	3,58	3,45	3,34	3,18	3,01	2,82	2,73	2,63	2,52	2,42	2,30	2,18
40	8,83	6,07	4,98	4,37	3,99	3,71	3,51	3,35	3,22	3,12	2,95	2,78	2,60	2,50	2,40	2,30	2,18	2,06	1,93
60	8,49	5,79	4,73	4,14	3,76	3,49	3,29	3,13	3,01	2,90	2,74	2,57	2,39	2,29	2,19	2,08	1,96	1,83	1,69
120	8,18	5,54	4,50	3,92	3,55	3,28	3,09	2,93	2,81	2,71	2,54	2,37	2,19	2,09	1,98	1,87	1,75	1,61	1,43
10000	7,88	5,30	4,28	3,72	3,35	3,09	2,90	2,75	2,62	2,52	2,36	2,19	2,00	1,90	1,79	1,67	1,54	1,37	1,05

v1 = Zahl der erklärenden Variablen (J)
v2 = Zahl der Freiheitsgrade des Nenners (K - J - 1)

Abbildung A.6: F-Tabelle (Vertrauenswahrscheinlichkeit 0,995)

A.3 c-Tabelle nach Cochran

$\alpha = 0{,}05$

k \ v	1	2	3	4	5	6	7	8	9	10	16	36	144	∞
2	0,9985	0,9750	0,9392	0,9057	0,8772	0,8534	0,8332	0,8159	0,8010	0,7880	0,7341	0,6602	0,5813	0,5000
3	0,9669	0,8709	0,7977	0,7457	0,7071	0,6771	0,6530	0,6333	0,6167	0,6025	0,5466	0,4748	0,4031	0,3333
4	0,9065	0,7679	0,6841	0,6287	0,5895	0,5598	0,5365	0,5175	0,5017	0,4884	0,4366	0,3720	0,3093	0,2500
5	0,8412	0,6838	0,5981	0,5441	0,5065	0,4783	0,4564	0,4387	0,4241	0,4118	0,3645	0,3066	0,2513	0,2000
6	0,7808	0,6161	0,5321	0,4803	0,4447	0,4184	0,3980	0,3817	0,3682	0,3568	0,3135	0,2612	0,2119	0,1667
7	0,7271	0,5612	0,4800	0,4307	0,3974	0,3726	0,3535	0,3384	0,3259	0,3154	0,2756	0,2278	0,1833	0,1429
8	0,6798	0,5157	0,4377	0,3910	0,3595	0,3362	0,3185	0,3043	0,2926	0,2829	0,2462	0,2022	0,1616	0,1250
9	0,6385	0,4775	0,4027	0,3584	0,3286	0,3067	0,2901	0,2768	0,2659	0,2568	0,2226	0,1820	0,1446	0,1111
10	0,6020	0,4450	0,3733	0,3311	0,3029	0,2823	0,2666	0,2541	0,2439	0,2353	0,2032	0,1655	0,1308	0,1000
12	0,5410	0,3924	0,3264	0,2880	0,2624	0,2439	0,2299	0,2187	0,2098	0,2020	0,1737	0,1403	0,1100	0,0833
15	0,4709	0,3346	0,2758	0,2419	0,2195	0,2034	0,1911	0,1815	0,1736	0,1671	0,1429	0,1144	0,0889	0,0667
20	0,3894	0,2705	0,2205	0,1921	0,1735	0,1602	0,1501	0,1422	0,1357	0,1303	0,1108	0,0879	0,0675	0,0500
24	0,3434	0,2354	0,1907	0,1656	0,1493	0,1374	0,1286	0,1216	0,1160	0,1113	0,0942	0,0743	0,0567	0,0417
30	0,2929	0,1980	0,1593	0,1377	0,1237	0,1137	0,1061	0,1002	0,0958	0,0921	0,0771	0,0604	0,0457	0,0333
40	0,2370	0,1576	0,1259	0,1082	0,0968	0,0887	0,0827	0,0780	0,0745	0,0713	0,0595	0,0462	0,0347	0,0250
60	0,1737	0,1131	0,0895	0,0765	0,0682	0,0623	0,0583	0,0552	0,0520	0,0497	0,0411	0,0316	0,0234	0,0167
120	0,0998	0,0632	0,0495	0,0419	0,0371	0,0337	0,0312	0,0292	0,0279	0,0266	0,0218	0,0165	0,0120	0,0083
∞	0	0	0	0	0	0	0	0	0	0	0	0	0	0

$\alpha = 0{,}01$

k \ v	1	2	3	4	5	6	7	8	9	10	16	36	144	∞
2	0,9999	0,9950	0,9794	0,9586	0,9373	0,9172	0,8988	0,8823	0,8674	0,8539	0,7949	0,7067	0,6062	0,5000
3	0,9933	0,9423	0,8831	0,8335	0,7933	0,7606	0,7335	0,7107	0,6912	0,6743	0,6059	0,5153	0,4230	0,3333
4	0,9676	0,8643	0,7814	0,7212	0,6761	0,6410	0,6129	0,5897	0,5702	0,5536	0,4884	0,4057	0,3251	0,2500
5	0,9279	0,7885	0,6957	0,6329	0,5875	0,5531	0,5259	0,5037	0,4854	0,4697	0,4094	0,3351	0,2644	0,2000
6	0,8828	0,7218	0,6258	0,5635	0,5195	0,4866	0,4608	0,4401	0,4229	0,4084	0,3529	0,2858	0,2299	0,1667
7	0,8376	0,6644	0,5685	0,5080	0,4659	0,4347	0,4105	0,3911	0,3751	0,3616	0,3105	0,2494	0,1929	0,1429
8	0,7945	0,6152	0,5209	0,4627	0,4226	0,3932	0,3704	0,3522	0,3373	0,3248	0,2779	0,2214	0,1700	0,1250
9	0,7544	0,5727	0,4810	0,4251	0,3870	0,3592	0,3378	0,3207	0,3067	0,2950	0,2514	0,1992	0,1521	0,1111
10	0,7175	0,5358	0,4469	0,3934	0,3572	0,3308	0,3106	0,2945	0,2813	0,2704	0,2297	0,1811	0,1376	0,1000
12	0,6528	0,4751	0,3919	0,3428	0,3099	0,2861	0,2680	0,2535	0,2419	0,2320	0,1961	0,1535	0,1157	0,0833
15	0,5747	0,4069	0,3317	0,2882	0,2593	0,2386	0,2228	0,2104	0,2002	0,1918	0,1612	0,1251	0,0934	0,0667
20	0,4799	0,3297	0,2654	0,2288	0,2048	0,1877	0,1748	0,1646	0,1567	0,1501	0,1248	0,0960	0,0709	0,0500
24	0,4247	0,2871	0,2295	0,1970	0,1759	0,1608	0,1495	0,1406	0,1338	0,1283	0,1060	0,0810	0,0595	0,0417
30	0,3632	0,2412	0,1913	0,1635	0,1454	0,1327	0,1232	0,1157	0,1100	0,1054	0,0867	0,0658	0,0480	0,0333
40	0,2940	0,1915	0,1508	0,1281	0,1135	0,1033	0,0957	0,0898	0,0853	0,0816	0,0668	0,0503	0,0363	0,0250
60	0,2151	0,1371	0,1069	0,0902	0,0796	0,0722	0,0668	0,0625	0,0594	0,0567	0,0461	0,0344	0,0245	0,0167
120	0,1225	0,0759	0,0585	0,0489	0,0429	0,0387	0,0357	0,0334	0,0316	0,0302	0,0242	0,0178	0,0125	0,0083
∞	0	0	0	0	0	0	0	0	0	0	0	0	0	0

v = Anzahl der Freiheitsgrade für s_z^2

k = Anzahl der Varianzen

entnommen aus: Sachs, L.; 1999, S.615

Abbildung A.7: c-Tabelle nach Cochran

A.4 χ^2-Tabelle

Chi-Quadrat-Tabelle

α / FG	0,99	0,975	0,95	0,9	0,8	0,7	0,5	0,3	0,2	0,1	0,05	0,025	0,01	0,001	0,0001
1	0,00016	0,00098	0,00393	0,0158	0,064	0,148	0,455	1,07	1,64	2,71	3,84	5,02	6,63	10,83	15,14
2	0,0201	0,0506	0,1026	0,2107	0,446	0,713	1,39	2,41	3,22	4,61	5,99	7,38	9,21	13,82	18,42
3	0,115	0,216	0,352	0,584	1,01	1,42	2,37	3,66	4,64	6,25	7,81	9,35	11,34	16,27	21,11
4	0,297	0,484	0,711	1,06	1,65	2,19	3,36	4,88	5,99	7,78	9,49	11,14	13,28	18,47	23,51
5	0,554	0,831	1,15	1,61	2,34	3,00	4,35	6,06	7,29	9,24	11,07	12,83	15,09	20,52	25,74
6	0,872	1,24	1,64	2,20	3,07	3,83	5,35	7,23	8,56	10,64	12,59	14,45	16,81	22,46	27,86
7	1,24	1,69	2,17	2,83	3,82	4,67	6,35	8,38	9,80	12,02	14,07	16,01	18,48	24,32	29,88
8	1,65	2,18	2,73	3,49	4,59	5,53	7,34	9,52	11,03	13,36	15,51	17,53	20,09	26,12	31,83
9	2,09	2,70	3,33	4,17	5,38	6,39	8,34	10,66	12,24	14,68	16,92	19,02	21,67	27,88	33,72
10	2,56	3,25	3,94	4,87	6,18	7,27	9,34	11,78	13,44	15,99	18,31	20,48	23,21	29,59	35,56
11	3,05	3,82	4,57	5,58	6,99	8,15	10,34	12,90	14,63	17,28	19,68	21,92	24,72	31,26	37,37
12	3,57	4,40	5,23	6,30	7,81	9,03	11,34	14,01	15,81	18,55	21,03	23,34	26,22	32,91	39,13
13	4,11	5,01	5,89	7,04	8,63	9,93	12,34	15,12	16,98	19,81	22,36	24,74	27,69	34,53	40,87
14	4,66	5,63	6,57	7,79	9,47	10,82	13,34	16,22	18,15	21,06	23,68	26,12	29,14	36,12	42,58
15	5,23	6,26	7,26	8,55	10,31	11,72	14,34	17,32	19,31	22,31	25,00	27,49	30,58	37,70	44,26
16	5,81	6,91	7,96	9,31	11,15	12,62	15,34	18,42	20,47	23,54	26,30	28,85	32,00	39,25	45,92
17	6,41	7,56	8,67	10,09	12,00	13,53	16,34	19,51	21,61	24,77	27,59	30,19	33,41	40,79	47,57
18	7,01	8,23	9,39	10,86	12,86	14,44	17,34	20,60	22,76	25,99	28,87	31,53	34,81	42,31	49,19
19	7,63	8,91	10,12	11,65	13,72	15,35	18,34	21,69	23,90	27,20	30,14	32,85	36,19	43,82	50,80
20	8,26	9,59	10,85	12,44	14,58	16,27	19,34	22,77	25,04	28,41	31,41	34,17	37,57	45,31	52,39
21	8,90	10,28	11,59	13,24	15,44	17,18	20,34	23,86	26,17	29,62	32,67	35,48	38,93	46,80	53,96
22	9,54	10,98	12,34	14,04	16,31	18,10	21,34	24,94	27,30	30,81	33,92	36,78	40,29	48,27	55,52
23	10,20	11,69	13,09	14,85	17,19	19,02	22,34	26,02	28,43	32,01	35,17	38,08	41,64	49,73	57,07
24	10,86	12,40	13,85	15,66	18,06	19,94	23,34	27,10	29,55	33,20	36,42	39,36	42,98	51,18	58,61
25	11,52	13,12	14,61	16,47	18,94	20,87	24,34	28,17	30,68	34,38	37,65	40,65	44,31	52,62	60,14
26	12,20	13,84	15,38	17,29	19,82	21,79	25,34	29,25	31,79	35,56	38,89	41,92	45,64	54,05	61,66
27	12,88	14,57	16,15	18,11	20,70	22,72	26,34	30,32	32,91	36,74	40,11	43,19	46,96	55,48	63,16
28	13,56	15,31	16,93	18,94	21,59	23,65	27,34	31,39	34,03	37,92	41,34	44,46	48,28	56,89	64,66
29	14,26	16,05	17,71	19,77	22,48	24,58	28,34	32,46	35,14	39,09	42,56	45,72	49,59	58,30	66,15
30	14,95	16,79	18,49	20,60	23,36	25,51	29,34	33,53	36,25	40,26	43,77	46,98	50,89	59,70	67,63
35	18,51	20,57	22,47	24,80	27,84	30,18	34,34	38,86	41,78	46,06	49,80	53,20	57,34	66,62	74,93
40	22,16	24,43	26,51	29,05	32,34	34,87	39,34	44,16	47,27	51,81	55,76	59,34	63,69	73,40	82,06
50	29,71	32,36	34,76	37,69	41,45	44,31	49,33	54,72	58,16	63,17	67,50	71,42	76,15	86,66	95,97
60	37,48	40,48	43,19	46,46	50,64	53,81	59,33	65,23	68,97	74,40	79,08	83,30	88,38	99,61	109,50
80	53,54	57,15	60,39	64,28	69,21	72,92	79,33	86,12	90,41	96,58	101,88	106,63	112,33	124,84	135,78
100	70,06	74,22	77,93	82,36	87,95	92,13	99,33	106,91	111,67	118,50	124,34	129,56	135,81	149,45	161,32
120	86,92	91,57	95,70	100,62	106,81	111,42	119,33	127,62	132,81	140,23	146,57	152,21	158,95	173,62	186,33
150	112,67	117,98	122,69	128,28	135,26	140,46	149,33	158,58	164,35	172,58	179,58	185,80	193,21	209,26	223,11
200	156,43	162,73	168,28	174,84	183,00	189,05	199,33	209,99	216,61	226,02	233,99	241,06	249,45	267,54	283,06

α = Signifikanzniveau
FG = Zahl der Freiheitsgrade (DR)

Abbildung A.8: χ^2-Tabelle

A.5 Durbin-Watson-Tabelle

Vertrauenswahrscheinlichkeit 0,95

K	J=1		J=2		J=3		J=4		J=5	
	d^+_u	d^+_o	d^+_u	d^+_o	d^+_u	d^+_o	d^+_u	d^+_o	d^+_u	d^+_o
15	1,08	1,36	0,95	1,54	0,82	1,75	0,69	1,97	0,56	2,21
16	1,10	1,37	0,98	1,54	0,86	1,73	0,74	1,93	0,62	2,15
17	1,13	1,38	1,02	1,54	0,90	1,71	0,78	1,90	0,67	2,10
18	1,16	1,39	1,05	1,53	0,93	1,69	0,82	1,87	0,71	2,06
19	1,18	1,40	1,08	1,53	0,97	1,68	0,86	1,85	0,75	2,02
20	1,20	1,41	1,10	1,54	1,00	1,68	0,90	1,83	0,79	1,99
21	1,22	1,42	1,13	1,54	1,03	1,67	0,93	1,81	0,83	1,96
22	1,24	1,43	1,15	1,54	1,05	1,66	0,96	1,80	0,86	1,94
23	1,26	1,44	1,17	1,54	1,08	1,66	0,99	1,79	0,90	1,92
24	1,27	1,45	1,19	1,55	1,10	1,66	1,01	1,78	0,93	1,90
25	1,29	1,45	1,21	1,55	1,12	1,66	1,04	1,77	0,95	1,89
26	1,30	1,46	1,22	1,55	1,14	1,65	1,06	1,76	0,98	1,88
27	1,32	1,47	1,24	1,56	1,16	1,65	1,08	1,76	1,01	1,86
28	1,33	1,48	1,26	1,56	1,18	1,65	1,10	1,75	1,03	1,85
29	1,34	1,48	1,27	1,56	1,20	1,65	1,12	1,74	1,05	1,84
30	1,35	1,49	1,28	1,57	1,21	1,65	1,14	1,74	1,07	1,83
31	1,36	1,50	1,30	1,57	1,23	1,65	1,16	1,74	1,09	1,83
32	1,37	1,50	1,31	1,57	1,24	1,65	1,18	1,73	1,11	1,82
33	1,38	1,51	1,32	1,58	1,26	1,65	1,19	1,73	1,13	1,81
34	1,39	1,51	1,33	1,58	1,27	1,65	1,21	1,73	1,15	1,81
35	1,40	1,52	1,34	1,58	1,28	1,65	1,22	1,73	1,16	1,80
36	1,41	1,52	1,35	1,59	1,29	1,65	1,24	1,73	1,18	1,80
37	1,42	1,53	1,36	1,59	1,31	1,66	1,25	1,72	1,19	1,80
38	1,43	1,54	1,37	1,59	1,32	1,66	1,26	1,72	1,21	1,79
39	1,43	1,54	1,38	1,60	1,33	1,66	1,27	1,72	1,22	1,79
40	1,44	1,54	1,39	1,60	1,34	1,66	1,29	1,72	1,23	1,79
45	1,48	1,57	1,43	1,62	1,38	1,67	1,34	1,72	1,29	1,78
50	1,50	1,59	1,46	1,63	1,42	1,67	1,38	1,72	1,34	1,77
55	1,53	1,60	1,49	1,64	1,45	1,68	1,41	1,72	1,38	1,77
60	1,55	1,62	1,51	1,65	1,48	1,69	1,44	1,73	1,41	1,77
65	1,57	1,63	1,54	1,66	1,50	1,70	1,47	1,73	1,44	1,77
70	1,58	1,64	1,55	1,67	1,52	1,70	1,49	1,74	1,46	1,77
75	1,60	1,65	1,57	1,68	1,54	1,71	1,51	1,74	1,49	1,77
80	1,61	1,66	1,59	1,69	1,56	1,72	1,53	1,74	1,51	1,77
85	1,62	1,67	1,60	1,70	1,57	1,72	1,55	1,75	1,52	1,77
90	1,63	1,68	1,61	1,70	1,59	1,73	1,57	1,75	1,54	1,78
95	1,64	1,69	1,62	1,71	1,60	1,73	1,58	1,75	1,56	1,78
100	1,65	1,69	1,63	1,72	1,61	1,74	1,59	1,76	1,57	1,78

K = Zahl der Beobachtungen
J = Zahl der Regressoren
d^+_u = unterer Grenzwert des Unschärfebereichs
d^+_o = oberer Grenzwert des Unschärfebereichs

entnommen aus: Durbin, J./Watson, G.S., 1951, S.173

Abbildung A.9: Durbin-Watson-Tabelle (Vertrauenswahrscheinlichkeit 0,95)

Vertrauenswahrscheinlichkeit 0,975

K	J=1		J=2		J=3		J=4		J=5	
	d^+_u	d^+_o	d^+_u	d^+_o	d^+_u	d^+_o	d^+_u	d^+_o	d^+_u	d^+_o
15	0,95	1,23	0,83	1,40	0,71	1,61	0,59	1,84	0,48	2,09
16	0,98	1,24	0,86	1,40	0,75	1,59	0,64	1,80	0,53	2,03
17	1,01	1,25	0,90	1,40	0,79	1,58	0,68	1,77	0,57	1,98
18	1,03	1,26	0,93	1,40	0,82	1,56	0,72	1,74	0,62	1,93
19	1,06	1,28	0,96	1,41	0,86	1,55	0,76	1,72	0,66	1,90
20	1,08	1,28	0,99	1,41	0,89	1,55	0,79	1,70	0,70	1,87
21	1,10	1,30	1,01	1,41	0,92	1,54	0,83	1,69	0,73	1,84
22	1,12	1,31	1,04	1,42	0,95	1,54	0,86	1,68	0,77	1,82
23	1,14	1,32	1,06	1,42	0,97	1,54	0,89	1,67	0,80	1,80
24	1,16	1,33	1,08	1,43	1,00	1,54	0,91	1,66	0,83	1,79
25	1,18	1,34	1,10	1,43	1,02	1,54	0,94	1,65	0,86	1,77
26	1,19	1,35	1,12	1,44	1,04	1,54	0,96	1,65	0,88	1,76
27	1,21	1,36	1,13	1,44	1,06	1,54	0,99	1,64	0,91	1,75
28	1,22	1,37	1,15	1,45	1,08	1,54	1,01	1,64	0,93	1,74
29	1,24	1,38	1,17	1,45	1,10	1,54	1,03	1,63	0,96	1,73
30	1,25	1,38	1,18	1,46	1,12	1,54	1,05	1,63	0,98	1,73
31	1,26	1,39	1,20	1,47	1,13	1,55	1,07	1,63	1,00	1,72
32	1,27	1,40	1,21	1,47	1,15	1,55	1,08	1,63	1,02	1,71
33	1,28	1,41	1,22	1,48	1,16	1,55	1,10	1,63	1,04	1,71
34	1,29	1,41	1,24	1,48	1,17	1,55	1,12	1,63	1,06	1,70
35	1,30	1,42	1,25	1,48	1,19	1,55	1,13	1,63	1,07	1,70
36	1,31	1,43	1,26	1,49	1,20	1,56	1,15	1,63	1,09	1,70
37	1,32	1,43	1,27	1,49	1,21	1,56	1,16	1,62	1,10	1,70
38	1,33	1,44	1,28	1,50	1,23	1,56	1,17	1,62	1,12	1,70
39	1,34	1,44	1,29	1,50	1,24	1,56	1,19	1,63	1,13	1,69
40	1,35	1,45	1,30	1,51	1,25	1,57	1,20	1,63	1,15	1,69
45	1,39	1,48	1,34	1,53	1,30	1,58	1,25	1,63	1,21	1,69
50	1,42	1,50	1,38	1,54	1,34	1,59	1,30	1,64	1,26	1,69
55	1,45	1,52	1,41	1,56	1,37	1,60	1,33	1,64	1,30	1,69
60	1,47	1,54	1,44	1,57	1,40	1,61	1,37	1,65	1,33	1,69
65	1,49	1,55	1,46	1,59	1,43	1,62	1,40	1,66	1,36	1,69
70	1,51	1,57	1,48	1,60	1,45	1,63	1,42	1,66	1,39	1,70
75	1,53	1,58	1,50	1,61	1,47	1,64	1,45	1,67	1,42	1,70
80	1,54	1,59	1,52	1,62	1,49	1,65	1,47	1,67	1,44	1,70
85	1,56	1,60	1,53	1,63	1,51	1,65	1,49	1,68	1,46	1,71
90	1,57	1,61	1,55	1,64	1,53	1,66	1,50	1,69	1,48	1,71
95	1,58	1,62	1,56	1,65	1,54	1,67	1,52	1,69	1,50	1,71
100	1,59	1,63	1,57	1,65	1,55	1,67	1,53	1,70	1,51	1,72

K = Zahl der Beobachtungen
J = Zahl der Regressoren
d^+_u = unterer Grenzwert des Unschärfebereichs
d^+_o = oberer Grenzwert des Unschärfebereichs

entnommen aus: Durbin, J./Watson, G.S., 1951, S.174

Abbildung A.10: Durbin-Watson-Tabelle (Vertrauenswahrscheinlichkeit 0,975)

A.6 q-Werte-Tabelle

df des Nenners	p%	Spannweite 2	3	4	5	6	7	8	9	10	11	12
1	5	18,00	27,00	32,80	37,10	40,40	43,10	45,40	47,40	49,10	50,60	52,00
	1	90,00	135,00	164,00	186,00	202,00	216,00	227,00	237,00	246,00	253,00	260,00
2	5	6,09	8,30	9,80	10,90	11,70	12,40	13,00	13,50	14,00	14,40	14,70
	1	14,00	19,00	22,30	24,70	26,60	28,20	29,50	30,70	31,70	32,60	33,40
3	5	4,50	5,91	6,82	7,50	8,04	8,48	8,85	9,18	9,46	9,72	9,95
	1	8,26	10,60	12,20	13,30	14,20	15,00	15,60	16,20	16,70	17,10	17,50
4	5	3,93	5,04	5,76	6,29	6,71	7,05	7,35	7,60	7,83	8,0,	8,21
	1	6,51	8,12	9,17	9,96	10,60	11,10	11,50	11,90	12,30	12,60	12,80
5	5	3,64	4,60	5,22	5,67	6,03	6,33	6,58	6,80	6,99	7,17	7,32
	1	5,70	6,97	7,80	8,42	8,91	9,32	9,67	9,97	10,20	10,50	10,70
6	5	3,46	4,34	4,90	5,31	5,63	5,89	6,12	6,32	6,49	6,65	6,79
	1	5,24	6,33	7,03	7,56	7,97	8,32	8,61	8,87	9,10	9,30	9,49
7	5	3,34	4,16	4,69	5,06	5,36	5,61	5,82	6,00	6,16	6,30	6,43
	1	4,95	5,92	6,54	7,01	7,37	7,68	7,94	8,17	8,37	8,55	8,71
8	5	3,26	4,04	4,53	4,89	5,17	5,40	5,60	5,77	5,92	6,05	6,18
	1	4,74	5,63	6,20	6,63	6,96	7,24	7,47	7,68	7,87	8,03	8,13
9	5	3,20	3,95	4,42	4,76	5,02	5,24	5,43	5,60	5,74	5,87	5,98
	1	4,60	5,43	5,96	6,35	6,66	6,91	7,13	7,32	7,49	7,65	7,78
10	5	3,15	3,88	4,33	4,65	4,91	5,12	5,30	5,46	5,60	5,72	5,83
	1	4,48	5,27	5,77	6,14	6,43	6,67	6,87	7,05	7,21	7,36	7,48
11	5	3,11	3,82	4,26	4,57	4,82	5,03	5,20	5,35	5,49	5,61	5,71
	1	4,39	5,14	5,62	5,97	6,25	6,48	6,67	6,84	6,99	7,13	7,26
12	5	3,08	3,77	4,20	4,51	4,75	7,95	5,12	5,27	5,40	5,51	5,62
	1	4,32	5,04	5,50	5,84	6,10	6,32	6,51	6,67	6,81	6,94	7,06
13	5	3,06	3,73	4,15	4,45	4,69	4,88	5,05	5,19	5,32	5,43	5,53
	1	4,25	4,96	5,40	5,73	5,98	6,19	6,37	6,53	6,87	6,79	6,90
14	5	3,03	3,70	4,11	4,41	4,64	4,83	4,99	5,13	5,25	5,36	5,46
	1	4,21	4,69	5,32	5,63	5,88	6,08	6,26	6,41	6,54	6,66	6,77
16	5	3,00	3,65	4,05	4,33	4,56	4,74	4,90	5,03	5,15	5,26	5,46
	1	4,13	4,78	5,19	5,49	5,72	5,92	6,08	6,22	6,35	6,46	6,56
18	5	2,97	3,61	4,00	4,28	4,49	4,67	4,82	4,96	5,07	5,17	5,27
	1	4,07	4,70	5,09	5,38	5,60	5,79	5,94	6,08	6,20	6,31	6,41
20	5	2,95	3,58	3,96	4,23	4,45	4,62	4,77	4,90	5,01	5,11	5,20
	1	4,03	4,64	5,02	5,29	5,51	5,69	5,84	5,97	6,09	6,19	6,29
24	5	2,92	3,53	3,90	4,17	4,37	4,54	4,68	4,81	4,92	4,01	5,10
	1	3,95	4,54	4,91	5,17	5,37	5,54	5,69	5,81	5,92	6,02	6,11
30	5	2,89	3,49	3,84	4,10	4,30	4,46	4,60	4,72	4,83	4,92	5,00
	1	3,89	4,45	4,80	5,05	5,24	5,40	5,54	5,56	5,76	5,85	5,93
40	5	2,86	3,44	3,79	4,04	4,23	4,39	4,52	4,63	4,74	4,82	4,91
	1	3,82	4,37	4,70	4,93	5,11	5,27	5,39	5,50	5,60	5,69	5,77
60	5	2,83	3,40	3,74	3,98	4,16	4,31	4,44	4,55	4,65	4,73	4,81
	1	3,76	4,28	4,60	4,82	4,99	5,13	5,25	5,36	5,45	5,53	5,60
120	5	2,80	3,36	3,69	3,92	4,10	4,24	4,36	4,48	4,56	4,64	4,72
	1	37,00	4,20	4,50	4,71	4,87	5,01	5,12	5,21	5,30	5,38	5,44
		2,77	3,31	3,63	3,86	4,03	4,17	4,29	4,39	4,47	4,55	4,62
		3,64	4,12	4,40	4,60	4,76	4,88	4,99	5,08	5,16	5,23	5,29

df = Zahl der Freiheitsgrade
p = Signifikanzniveau in %

entnommen aus: Fröhlich, Werner D./Becker, Johannes: Forschungsstatistik, 6. Aufl., Bonn 1972, S.547

Abbildung A.11: q-Werte-Tabelle

Stichwortverzeichnis