Universität Potsdam

Dieter Holtmann

Grundlegende multivariate Modelle der sozialwissenschaftlichen Datenanalyse

3., veränderte Auflage

Universitätsverlag Potsdam

Dieter Holtmann
Grundlegende multivariate Modelle der sozialwissenschaftlichen Datenanalyse

Dieter Holtmann

Grundlegende multivariate Modelle der sozialwissenschaftlichen Datenanalyse

3., veränderte Auflage

Universitätsverlag Potsdam

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

Universitätsverlag Potsdam 2010
http://info.ub.uni-potsdam.de/verlag.htm

Am Neuen Palais 10, 14469 Potsdam
Tel.: +49 (0)331 977 4623 / Fax: 3474
E-Mail: verlag@uni-potsdam.de

Herausgeber: Prof. Dr. Dieter Holtmann, Wirtschafts- und Sozialwissenschaftliche Fakultät der Universität Potsdam

3., veränderte Auflage
Online veröffentlicht auf dem Publikationsserver der Universität Potsdam
URL http://pub.ub.uni-potsdam.de/volltexte/2010/4593/
URN urn:nbn:de:kobv:517-opus-45931
http://nbn-resolving.org/urn:nbn:de:kobv:517-opus-45931

Zugleich gedruckt erschienen im Universitätsverlag Potsdam
ISBN 978-3-86956-084-7

Die jeweils aktuelle Auflage ist abrufbar unter
http://opus.kobv.de/ubp/abfrage_collections.php?coll_id=716

Vorwort

Zur adäquaten Analyse sozialwissenschaftlicher Phänomene ist die Anwendung multivariater Modelle hilfreich, die die Analyse von Zusammenhängen und Abhängigkeiten zwischen vielen Merkmalen ermöglichen.

Als grundlegende Modelle werden im folgenden Band behandelt: Die Elaboration von Zusammenhängen lässt sich durch Teilgruppenvergleich (Tabellenanalyse) auf nominalem Messniveau und durch partielle Korrelation auf metrischem Messniveau durchführen. In der multiplen Regression wird die Variation eines interessierenden Phänomens auf die Variation einer Reihe von Erklärungsfaktoren zurückgeführt. Die wichtigsten Interpretationshilfen dabei sind der Anteil der erklärten Varianz und die Effekte. In der Pfadanalyse werden alle Mechanismen herausgearbeitet, durch deren Zusammenwirken die Höhe jedes statistischen Zusammenhangs bestimmt wird: Direkte und indirekte Kausaleffekte, scheinkausale Komponenten und Assoziationseffekte. In der Varianzanalyse wird die Variation eines interessierenden Phänomens auf Haupteffekte und Interaktionseffekte einer Reihe von Erklärungsfaktoren zurückgeführt.

Der vorliegende Band liegt meiner Lehrveranstaltung „Grundlegende multivariate Modelle der sozialwissenschaftlichen Datenanalyse“ zugrunde (vergleiche http://www.uni-potsdam.de/u/soziologie/methoden/mitarbeiter/holtmann/index.htm). Für die zahlreichen Anregungen von Studierenden und Mitarbeitern möchte ich mich herzlich bedanken. An der vorliegenden Fassung wirkten mit: Claudia Buchheister (Kap. 1), Silke Grau (Kap.2), Markus Seyfried (Kap. 3.1), Michael Mutz (Kap. 3.2) und Silke Hans (Kap. 4). In der dritten Auflage wurden einige Erläuterungen ergänzt. Zu früheren Fassungen waren die Anmerkungen u.a. von Uwe Hasenbein, Michael Schüler, Christian Dössel und Tilo Görl sehr hilfreich. Meinen Mitarbeitern, die mit mir zusammen dieses Lehrprogramm umsetzen, sowie Annett Wadewitz und Anja Nitsche für die Gestaltung des vorliegenden Buches gilt an dieser Stelle mein besonderer Dank.

Prof. Dr. Dieter Holtmann
(Sozialwissenschaften / Methoden der empirischen Sozialforschung)

Inhaltsverzeichnis

1. Überblick über die multivariaten Modelle der sozialwissenschaftlichen Datenanalyse 1
1.1 Charakterisierung verschiedener Datenanalyseverfahren 1
1.2 Abhängigkeiten vs. Zusammenhänge (Asymmetrische und symmetrische Fragestellungen) 3
1.3 Erforderliches Messniveau der Variablen 4
1.4 Charakterisierung einiger Verfahren durch Diagramme 5
1.5 Gegenüberstellung der Logik der Verfahren der multiplen Regression, der Faktorenanalyse und der Varianzanalyse 8
1.6 Bewertung 9
1.7 Pragmatische Abgrenzung von grundlegenden und fortgeschrittenen multivariaten Modellen sowie Aufbau des vorliegenden Bandes 11
Literaturverzeichnis 12

2. Kausalanalyse mit Tabellenanalyse und partieller Korrelation 13
2.1 Kausalanalyse und statistischer Kausalbegriff 13
2.2 Tabellenanalyse 15
2.2.1 Ein Beispiel für Korrelation ohne Kausalität (Scheinkorrelation/scheinkausale Korrelation) 16
2.2.2 Zerlegung der Vier-Felder-Tafel an einem einführenden Beispiel 17
2.2.2.1 Log-lineare Modellierung des einführenden Beispiels 21
2.2.3 Die Grundgleichung (Zerlegungsformel für Maßzahlen) 24
2.2.4 Typologie von Kausalstrukturen mit drei Variablen 29
2.2.4.1 Typen mit: [xy : z] = [xy : ¬z] („Scheinkorrelation“, Intervenierende Variable, Suppressor, Distorter) 29
2.2.4.2 Zerlegungsformel am Beispiel eines Suppressor- sowie Distorter-Phönomens 32
2.2.4.3 Vorzeichenregel nach Davis für Suppressor- und Distorter Phänomene 39
2.2.4.4 Typen mit: [xy : z] ≠ [xy ¬z] (Spezifikation) 43
2.2.4.5 Conjoint influence 43
2.2.4.6 Verschiedene typologische Effekte und Interaktionseffekte 45
2.2.4.6.1 Varianzanalytische Interpretation von Rosenbergs Mobilitäts-Beispiel 46
2.2.4.7 Interaktion, Spezifikation und typologische Effekte aus Sicht der Varianzanalyse 54
2.2.4.8 Kausale Interpretation von Zusammenhängen 65
2.3 Partielle Korrelation 67
Literaturverzeichnis 73

3. Multiple Regressionsanalyse und Pfadanalyse 75
3.1 Multiple Regressionsanalyse 75
3.1.1 Das Grundprinzip der einfachen Regression, geometrische Interpretation und Matrixschreibweise 77
3.1.2 Die Regressionskoeffizienten b und beta 79
3.1.3 Gleichungsansatz der multiplen Regression und Matrixschreibweise 81
3.1.4 Multipler Korrelationseffekt R 84
3.1.5 Interpretation der Koeffizienten 86
3.1.6 Schrittweise Regression 88
3.1.7 Zuordnung der gesamten erklärten Varianz zu den Prädiktoren 89
3.1.7.1 Einführung in die Problemstellung 89
3.1.7.2 Charakterisierung der Koeffizienten mit Hilfe von Residuen 90

3.1.7.3 Zwei Zerlegungen von Multiple R^2 94
3.1.7.4 Darstellung der erklärten Varianz durch Kovarianzen und Effekte 95
3.1.8 Interaktion in der Regression 96
3.1.9 Anknüpfung an die Tabellenanalyse: Regressionsanalyse der Lebenszufriedenheit 97
3.1.10 Logistische Regression 103
3.1.11 Statistische Inferenz 105
3.1.12 Die Verletzung der Modellannahmen 109
3.1.13 Beispiel für die Regressionsanalyse 111
3.1.14 Die multivariate Regression 120
3.1.15 „Weighted least squares“ 120
Literaturverzeichnis 123

3.2 Pfadanalyse 124
3.2.1 Ein klassisches Beispiel von Blau und Duncan 124
3.2.2 Kausale Ordnung und Rekursivität 125
3.2.3 Vollständiges Modell und unvollständiges Modell 126
3.2.3.1 Kausale Geschlossenheit eines Modells gegenüber weiteren Einflussfaktoren 128
3.2.3.2 Pfaddiagramm und Effekte für das vollständige Modell mit 2, 3 und 4 Variablen 129
3.2.3.3 Unvollständiges Modell 133
3.2.3.4 Standardisierte oder unstandardisierte Koeffizienten? 133
3.2.4 Die vier Mechanismen zur Erklärung einer Korrelation 134
3.2.5 Anwendungsbeispiel: Parteienwahl in Abhängigkeit von Parteiidentifikation und Einstellungen 136
3.2.6 Multiple R^2 in der Pfadanalyse 138
3.2.6.1 Zentrale Konzepte der statistischen Analyse gemäß der Pfadanalyse 139
3.2.7 Effekte und Erklärungskraft in der Pfadanalyse 139
3.2.7.1 Korrelierte Effekte (Multiple Regression) 139
3.2.7.2 Indirekte Effekte (Pfadanalyse) 141
3.2.7.3 Problematisierung der Pfadanalyse 141
3.2.7.4 Vergleich von Gesamtzusammenhang und bereinigtem Zusammenhang: Typologie und Zerlegung von R^2 142
3.2.8 Partielle Korrelation oder Pfadkoeffizient? 158
3.2.9 Relative Bedeutung von Multiple R^2 und den Effekten für die Erklärung 160
Literaturverzeichnis 161

4. Varianzanalyse und Kovarianzanalyse 163
4.1 Einfache Varianzanalyse als Verallgemeinerung des t-Tests 164
4.1.1 Varianzzerlegung 164
4.1.2 Signifikanztest 167
4.1.3 Einfache Varianzanalyse und t-Test 170
4.1.4 Anteil erklärter Varianz als Deskription 170
4.2 Zweifache Varianzanalyse 171
4.2.1 Gleiche Zellenhäufigkeiten 171
4.2.2 Ungleiche Zellenhäufigkeiten 174
4.3 Dreifache Varianzanalyse 177
4.4 Die einfache Varianzanalyse als Spezialfall der multiplen Regression 179
4.5 Die zweifache Varianzanalyse als Spezialfall der multiplen Regression mit Interaktionstermen 181
4.5.1 Beispiel für die zweifache Varianzanalyse mit ungleichen Zellenhäufigkeiten: Untersuchung der Lebenszufriedenheit 183

4.6 Unterschiedliche Codierungen in der Varianzanalyse 192
4.6.1 Codierung durch Dichotomien in der einfachen Varianzanalyse 192
4.6.2 Effekt-Codierung in der einfachen Varianzanalyse 193
4.6.3 Effekt-Codierung in der zweifachen Varianzanalyse 194
4.7 Die Design-Matrix 195
4.8 Kovarianzanalyse 198
4.8.1 Kovarianzzerlegung 199
4.8.1.1 Kovarianzzerlegung nach einer nominalen unabhängigen Variablen 199
4.8.1.2 Anwendung der Kovarianzzerlegung: Aggregatdaten und Mehrebenenanalyse 201
4.8.1.3 Die Kovarianz- und Korrelationszerlegung nach einer nominalen unabhängigen Variablen als Spezialfall einer allgemeinen Kovarianz- und Korrelationszerlegung nach metrischen Variablen 204
4.9 Kontrastgruppenanalyse (tree analysis) 211
4.10 Anwendungsbeispiel zur Varianzanalyse: Vergleich der Erklärungskraft verschiedener Berufsstruktur- und Klassenmodelle für die Bundesrepublik Deutschland 218
4.10.1 Probleme des Modellvergleichs und Kriterien zur Beurteilung der Erklärungskraft 218
4.10.1.1 Indikatoren für die Hierarchie der materiellen Lage 219
4.10.1.2 Indikatoren für den ideologischen Standort (Bewusstsein) 219
4.10.2 Vergleich der Erklärungskraft der verschiedenen Berufsstruktur- und Klassenmodelle 220
4.10.3 Graphische Darstellung der verschiedenen Berufsstruktur- und Klassenmodelle 223
4.10.4 Berufsstrukturmodell auf Basis der bundesdeutschen Sozialstatistik nach Einkommen und Bewusstseins-Index 223
Literaturverzeichnis 230

Anhang: Multiple Regressionsanalyse mit Hilfe von Determinanten 232

Sachregister 237

1. Überblick über die multivariaten Modelle der sozialwissenschaftlichen Datenanalyse

Im Folgenden werden die wichtigsten **multivariaten Modelle** der sozialwissenschaftlichen Datenanalyse, d.h. die Modelle, die Zusammenhänge und Abhängigkeiten von mehr als zwei Merkmalen berücksichtigen, kurz charakterisiert und systematisiert.

1.1 Charakterisierung verschiedener Datenanalyseverfahren

Den Übergang von der Analyse des Zusammenhangs je zweier Variablen/Faktoren zur multivariaten (= Mehrvariablen-) Analyse bildet die **Tabellenanalyse**, bei der im einfachsten Fall der Einfluss eines Drittfaktors auf den Zusammenhang zweier Variablen bei nominalem oder ordinalem Messniveau untersucht wird. Auf metrischem Messniveau entspricht dem die Berechnung der partiellen Korrelation. Sowohl in der Tabellenanalyse als auch in der partiellen Korrelation braucht man sich nicht auf die Kontrolle eines Drittfaktors zu beschränken, sondern kann mehrere „Dritt"-Faktoren gleichzeitig kontrollieren. Als Alternative zur Tabellenanalyse kann man die Regressionsanalyse verwenden – bei nominalem oder ordinalem Messniveau wird die Regression mit Dummyvariablen als Repräsentanten von Merkmalsausprägungen ausgeführt.

In der multiplen **Regressionsanalyse** wird eine metrische abhängige Variable (Prädikand) durch zwei oder mehr metrische unabhängige Variablen (Prädiktoren) statistisch erklärt. Die **Pfadanalyse** besteht in der mehrfachen Anwendung der Regression, indem mit jeder Variablen eine Regression auf alle Variablen durchgeführt wird, die dieser Variablen kausal vorangehen. Es muss also eine kausale Ordnung der Variablen für das Verfahren vorgegeben werden. Nicht-rekursive Modelle lassen in Erweiterung der Pfadanalyse auch Wechselwirkungen zwischen Variablen zu.

In der **Varianzanalyse** wird die Streuung der abhängigen metrischen Variablen in Bestandteile zerlegt, die der Variation der nominalen unabhängigen Variablen und ihrer Interaktionen zugerechnet werden können bzw. als unerklärter Rest interpretierbar sind. Die einfache Varianzanalyse, bei der eine unabhängige Variable als Prädiktor auftritt, ist ein Spezialfall der multiplen Regression mit Dummyvariablen als Prädiktoren, die mehrfache Varianzanalyse ist ein Spezialfall der multiplen Regression mit Interaktionstermen.

Die **„tree analysis"** entsteht durch mehrfache Anwendung der Varianzanalyse jeweils auf Dichotomien (= Variablen mit zwei Ausprägungen). Sie liefert eine Baumdarstellung, bei der die Varianz der abhängigen Variablen dadurch erklärt wird, dass sie nach der jeweils besten Prädiktor-Dichotomie zerlegt wird.

Die **Faktorenanalyse** ist eine Art multiple Regression einer Vielzahl manifester (direkt beobachtbarer) Variablen auf wenige latente Dimensionen (= Faktoren). Auf diese Weise wird eine Vereinfachung der Struktur erzielt. Die Korrelationszusammenhänge zwischen mehreren Variablen werden ohne größeren Informationsverlust durch wenige Faktoren reproduziert. Da die gefundenen Faktoren i.d.R. hypothetische sind, kann an sie die Forderung der statistischen Unabhängigkeit (Orthogonalität) gestellt werden. Eine bessere inhaltliche Interpretation der Faktoren erfordert jedoch oftmals oblique (nicht orthogonale) Faktoren. Die Faktorenanalyse dient u.a. der graphischen Darstellung von Zusammenhangsstrukturen. Ein wichtiger Aspekt ist dabei auch die Bestimmung der erforderlichen Anzahl von Dimensionen.
Die **kanonische Korrelation** ist eine Verallgemeinerung der multiplen Regression (oder der Faktorenanalyse) auf zwei Mengen von Variablen: Aus jeder der beiden Variablenmengen werden diejenigen Linearkombinationen bestimmt, welche die maximale Korrelation liefern. Auf diese Weise soll der Zusammenhang von zwei Mengen von Variablen untersucht werden.

Die **Diskriminanzanalyse** ist ein Klassifikationsverfahren, bei dem die Zuordnung von Fällen oder Variablen zu Gruppen mit Hilfe von Klassifikationsgleichungen für jede der Gruppen durchgeführt wird. Diese Gleichungen erhält man mit Hilfe der Varianzanalyse. Die Fälle werden denjenigen Gruppen zugeordnet, für die sie die höchste Wahrscheinlichkeit haben. Die Variablen, mit denen die Gruppenzugehörigkeit vorhergesagt wird, haben metrisches Messniveau. Die Gruppenzugehörigkeit selbst ist die abhängige Variable und hat nominales Messniveau. Man kann die Diskriminanzanalyse auch als kanonische Korrelation der Ausgangsvariablen und der Variablen der Gruppeneinteilung behandeln. Ein Anwendungsgebiet ist z.B. die Analyse unterschiedlicher Wählergruppen nach soziodemographischen Merkmalen.

Die **multidimensionale Skalierung** (MDS) liefert einen „smallest space" zur Darstellung der Analyseeinheiten – ähnlich wie die Faktorenanalyse. In der multidimensionalen Skalierung liegt das Schwergewicht häufig auf der Konfiguration von Punkten für die Analyseeinheiten und der graphischen Darstellung der Konfiguration, in der Faktorenanalyse auf der Interpretation der Faktoren und ihrer Erklärungskraft. Im Gegensatz zur Faktorenanalyse braucht die Ausgangsinformation über Ähnlichkeitsmaße nur auf ordinalem Messniveau vorzuliegen und liefert dennoch metrische Ergebnisse. Die MDS dient u.a. zur Rekonstruktion von Wahrnehmungsräumen z.B. für Parteipräferenzen.

Die **Clusteranalyse** ordnet die Analyseeinheiten ebenfalls aufgrund von Ähnlichkeitsmaßen in Gruppen. Sie kann unmittelbar ansetzen bei den Ähnlichkeitsmaßen oder von einer Konfiguration von Punkten für die Analyseeinheiten ausgehen, wie sie eine vorhergehende Faktorenanalyse oder multidimensionale Skalierung liefert. Im Unterschied zur Faktorenanalyse wird nicht eine Bündelung von Variablen, sondern von Analyseeinheiten angestrebt.

Der χ^2-Test zur Untersuchung des Zusammenhangs zweier nominaler Variablen mittels einer Kontingenztafel kann auf k (> 2) Variablen erweitert werden. Ferner kann die Varianzanalyse auch bei einer nominalen abhängigen Variablen (mit l Ausprägungen) angewendet werden, wenn man l - 1 Ausprägungen durch Dichotomien (= formal metrische Variablen) darstellt. Der erste Ansatz ermöglicht die Analyse von Zusammenhängen, der letztere von Abhängigkeiten. Da häufig mit logarithmierten Variablen gearbeitet wird, spricht man auch von **log-linearen Modellen**.

Mit Hilfe der **Korrespondenzanalyse** lassen sich die Zusammenhänge zwischen nicht-metrischen Merkmalen in einer graphischen Darstellung dadurch visualisieren, dass Unähnlichkeiten zwischen Verteilungen bzw. Profilen als Distanzen repräsentiert werden.

Die **Konfigurationsfrequenzanalyse** verwendet die Kontingenztafel, um (über- oder unterproportional besetzte) Typen zu ermitteln. Man kann sie auch als spezielles Cluster-Verfahren ansehen.

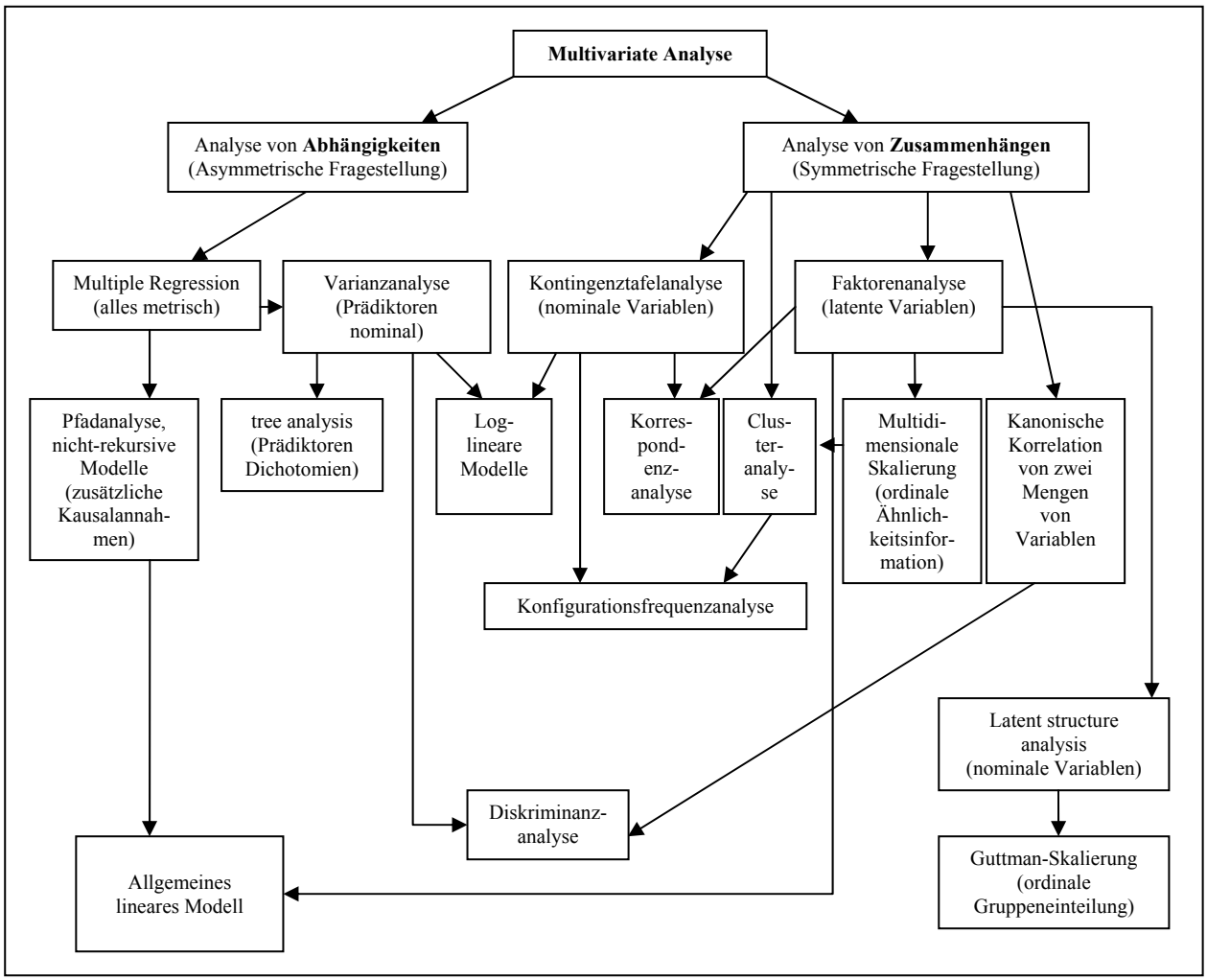
Die **„latent structure analysis"** sucht aufgrund dichotomer Information eine Entmischung in latente, zugrunde liegende Gruppen. Man kann die „latent structure analysis" als eine Art Faktorenanalyse für Dichotomien ansehen, wovon wiederum die Guttman-Skalierung ein Spezialfall für ordinale Gruppeneinteilung ist. Als Entmischungsmodell ist sie ein spezieller Fall solcher Modelle, die auch in der Clusteranalyse eine Rolle spielen.

Das **Allgemeine Lineare Modell** verbindet das lineare Kausalmodell mit einer Messfehlertheorie, die wie die Faktorenanalyse ansetzt: Die beobachteten Indikatoren werden als lineare Funktionen der zugrunde liegenden „wahren" Dimensionen betrachtet. Es wird nicht die deterministische Faktorenanalyse verwendet, sondern die statistische Maximum-Likelihood-Faktorenanalyse. (Bei der Theorieüberprüfung handelt es sich um eine konfirmatorische Faktorenanalyse, ansonsten um eine explorative.) Die Abhängigkeiten unter den latenten Variablen stehen im Zentrum des Interesses, wobei die Faktorenanalyse das Messmodell für die latenten Konstrukte beiträgt.

1.2 Abhängigkeiten vs. Zusammenhänge (Asymmetrische und symmetrische Fragestellungen)

Die Datenanalyseverfahren lassen sich zum größten Teil danach unterscheiden, ob sie Abhängigkeiten oder Zusammenhänge untersuchen. Dies ist in Abbildung 1-1 zusammengefasst. Das Schema umfasst auch die Beziehungen zwischen den Verfahren, die schon in den kurzen Charakterisierungen der Verfahren angesprochen worden sind.

Abbildung 1-1: Zusammenhänge zwischen den multivariaten Analyseverfahren



Diese Zuordnung ordnet drei Verfahren als Kombination beider Konzepte ein:

Log-lineare Ansätze gibt es sowohl zur Analyse von Abhängigkeiten (wie die Varianzanalyse) als auch von Zusammenhängen (wie die Kontingenztafelanalyse).

Die Diskriminanzanalyse ist eher ein Verfahren zur Untersuchung von Abhängigkeiten, wobei die (gesuchte) Gruppeneinteilung die abhängige Variable ist (bzw. Funktionen, die die Gruppen repräsentieren wie z.B. die Wähler einer bestimmten Partei). Die unabhängigen Variablen (z.B. sozialstrukturelle Merkmale von SPD-Wählern, CDU-Wählern etc.) dienen der Charakterisierung bzw. Vorhersage der verschiedenen Teilgruppen (z.B. Wählergruppen). Man kann eine Diskriminanzanalyse auch als kanonische Korrelation zwischen den Ausgangsvariablen und den Variablen für die Gruppeneinteilung (Dichotomien) durchführen. In diesem Sinne spielt auch der Aspekt der Analyse von Zusammenhängen eine Rolle.

Das allgemeine lineare Modell ist eher ein Modell zur Analyse von Abhängigkeiten. Der Aspekt der Gleichbehandlung von Variablen (d.h. nicht nach abhängig oder unabhängig zu unterscheiden) kommt nur durch die Faktorenanalyse herein, bei der jeweils mehrere Indikatoren eine Dimension repräsentieren. Die Beziehungen zwischen den Indikatoren werden in gleicher Weise benutzt, um diese latenten Variablen zu ermitteln.

Allgemein stehen aber Abhängigkeiten und Zusammenhänge in einer engen Beziehung: So lässt sich aus der einfachen Regression (als Konzept der Abhängigkeit) der Korrelationskoeffizient (als Zusammenhangsmaß) ableiten. Die Analyse von Abhängigkeiten scheint mir deshalb tendenziell als Konzept ursprünglicher zu sein, ebenso wie empirische Phänomene der ersten Art mir tendenziell häufiger zu sein scheinen.

Dass man die Faktorenanalyse unter die Verfahren zur Analyse von Zusammenhängen subsumiert, liegt daran, dass man bei den Indikatoren nicht zwischen abhängig und unabhängig unterscheiden kann. Letzteres aber ist darin begründet, dass die Indikatoren als kausal abhängig von den wahren Dimensionen (= Faktoren) angesehen werden. In diesem Sinne ist die Faktorenanalyse ein Verfahren zur Untersuchung von Abhängigkeiten. Bei diesem weiteren Begriff von Abhängigkeit würde im Wesentlichen nur das symmetrische Konzept der statistischen Unabhängigkeit als Verfahren zur Analyse von Zusammenhängen im strengen Sinne übrig bleiben.

Eine scharfe Trennung in Verfahren zur Analyse von Abhängigkeiten und von Zusammenhängen ist also nicht immer möglich.

1.3 Erforderliches Messniveau der Variablen

1) Alle Variablen sind **metrisch**.
   a) Multiple Regression; Pfadanalyse; nicht rekursive Modelle; Rozebooms Zusammenhangsmaß für n Variablen.
   b) Zusätzliche Verwendung von latenten Variablen: Faktorenanalyse; kanonische Korrelation; allgemeines lineares Modell.

2) **Gemischt:** Die Variablen müssen bei den folgenden Ansätzen nicht alle metrisch sein.

   a) Abhängige Variable metrisch:
   Varianzanalyse (unabhängige Variablen nominal); tree analysis (unabhängige Variablen Dichotomien).

   b) Diskriminanzanalyse: Die eigentliche abhängige Variable, die Gruppeneinteilung, ist nominal; bei der Klassifikation wird aber jede Gruppe durch eine metrische abhängige Variable vertreten; die Ausgangsvariablen sind metrisch.

   c) Clusteranalyse: Die Gruppeneinteilung (= nominal) wird bestimmt aufgrund von Ausgangsvariablen, die ein gleiches, aber beliebiges Messniveau haben.

   d) Multidimensionale Skalierung: Aus ordinaler Information über Ähnlichkeitsmaße wird metrische Information gewonnen.

   e) Guttman-Skalierung: Die latente ordinale Gruppeneinteilung wird aufgrund von dichotomer Information ermittelt.

3) Alle Variablen sind **nominal**.

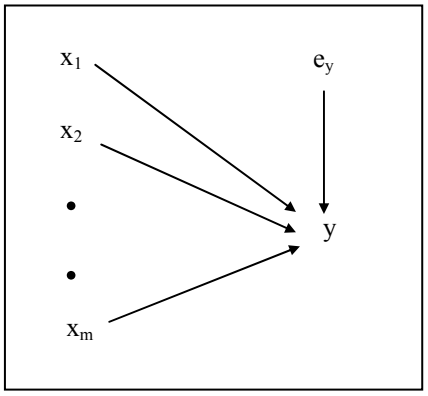
   a) Kontingenztafelanalyse; log-lineare Modelle; Konfigurationsfrequenzanalyse.

   b) Zusätzliche Verwendung von latenten Variablen: Latent structure analysis (Die latente Gruppeneinteilung (nominal) wird aufgrund von dichotomer Information ermittelt.)

1.4 Charakterisierung einiger Verfahren durch Diagramme

Multiple Regression

Abbildung 1-2: Multiple Regression



($x_1,...,x_m$ = unabhängige Variablen; y = abhängige Variable; e_y = error term (Fehlerterm), dem die nicht erklärte Varianz zugeschrieben wird.)

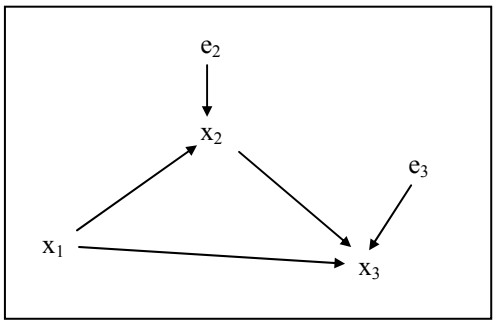
Die Pfeile stehen für Einflussrichtungen. Die **Varianzanalyse** hat als Spezialfall der multiplen Regression den gleichen formalen Aufbau.

Die **logistische Regression** hat die gleiche Struktur, aber die abhängige Variable ist dichotom und wird deshalb zuerst logarithmisch transformiert.

Die **Diskriminanzanalyse** hat formal den gleichen Aufbau, wobei die abhängige Variable aber eine Gruppeneinteilung (nominales Merkmal) ist, die durch metrische Prädiktoren vorhergesagt wird.

Pfadanalyse

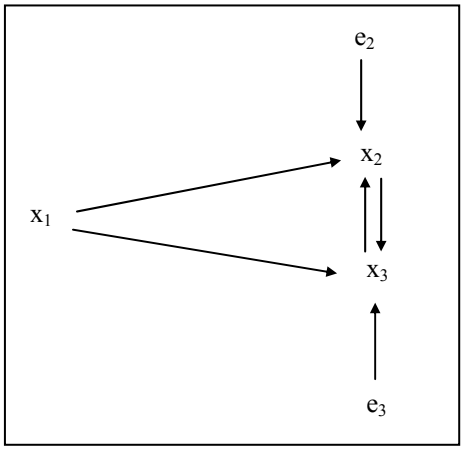
Abbildung 1-3: Pfadanalyse



Es handelt sich hier um eine Regression von x_2 auf x_1 und eine anschließende Regression von x_3 auf x_1 und x_2.

Nicht-rekursive Modelle

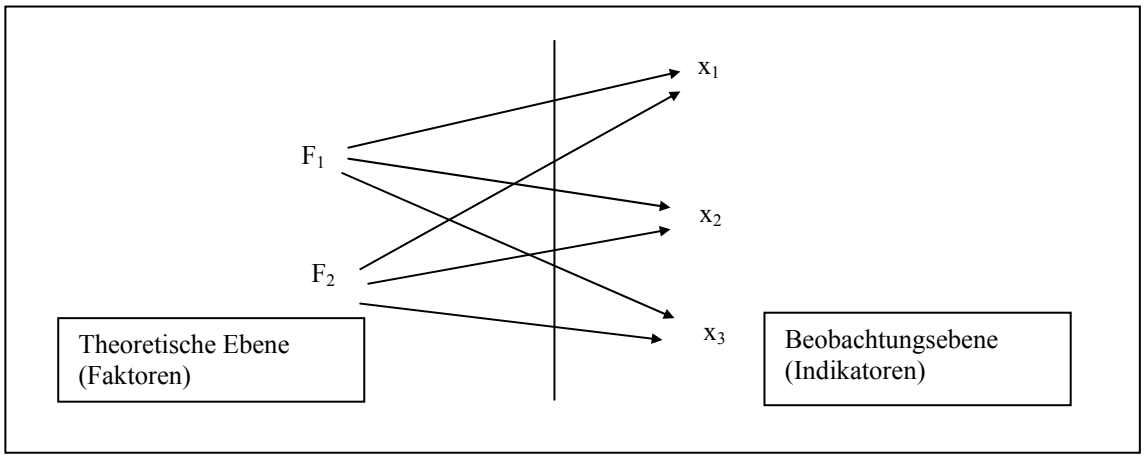
Abbildung 1-4: Nicht-rekursive Modelle



Zwischen x_2 und x_3 gibt es eine Wechselwirkung.

Faktorenanalyse

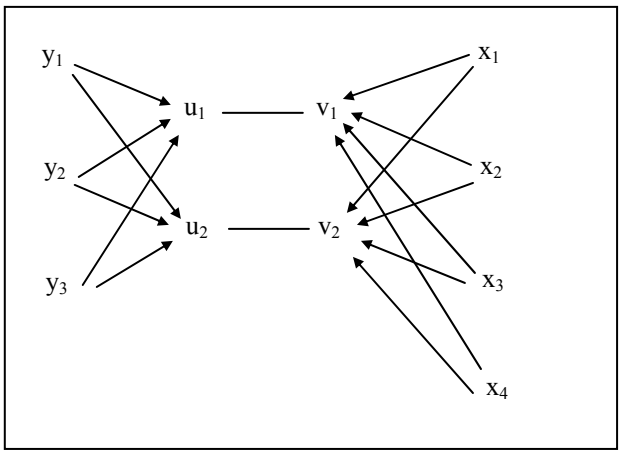
Abbildung 1-5: Faktorenanalyse



Die latenten Faktoren F_i schlagen sich in den manifesten Indikatoren x_j nieder.

Kanonische Korrelation

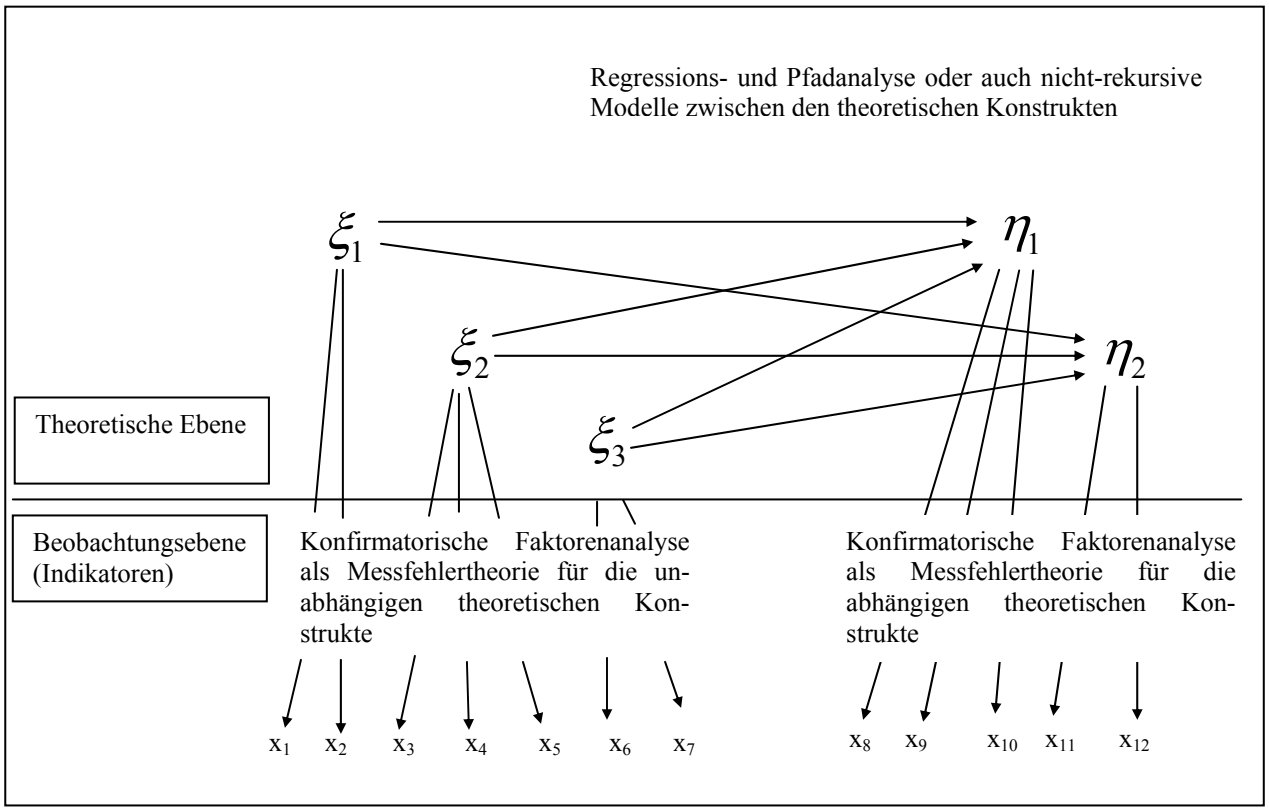
Abbildung 1-6: Kanonische Korrelation



(u_1, v_1) und (u_2, v_2) sind die ersten beiden Paare von kanonischen Faktoren.

Allgemeines lineares Modell (z.B. LISREL oder AMOS)

Abbildung 1-7: Allgemeines lineares Modell (z.B. LISREL oder AMOS)



Aus beobachteten Variablen lässt sich nicht die Existenz nicht beobachteter, latenter Variablen ableiten. (Es geht vielmehr um die Verträglichkeit von Modellen mit empirischen Daten.)

1.5 Gegenüberstellung der Logik der Verfahren der multiplen Regression, der Faktorenanalyse und der Varianzanalyse

Variablen	Modellansatz	Interpretation
Multiple Regression		
Metrische Variablen; eine abhängige Variable y, unabhängige Variablen $x_1, ..., x_m$.	$y = \beta_1 x_1 + ... + \beta_m x_m$ y, x_i bekannt; β_i = Beta-Koeffizienten werden bestimmt. Die x_i können miteinander korrelieren.	Multiple R^2 = Anteil der durch die Regression erklärten Varianz von y; r^2_{y,x_i} = Anteil der durch Prädiktor x_i insgesamt erklärten Varianz von y; Anteil der von x_i unabhängig von den anderen Prädiktoren erklärten Varianz ($= r^2_{y,x_i - \hat{x}_i}$, s.u.)
Faktorenanalyse		
Metrische Variablen; gleichgestellte Variablen $Z_1, ..., Z_n$ (= Indikatoren)	$Z_i = a_{i1}F_1 + ... a_{im}F_m + d_iU_i$ Die Z_i, i = 1, ..., n, sind bekannt; die Faktoren F_j, U_i werden dadurch charakterisiert, dass die Ladungen a_{ij} bestimmt werden. Die Faktoren F_j, U_i sind untereinander unkorreliert.	a^2_{ij} = Anteil der Varianz von z_i, der durch Faktor F_j erklärt wird; d^2_i = Anteil der Restvarianz; $\sum_{i=1}^{n} a^2_{ij}$ = Anteil der Varianzen von $Z_1, ...,$ Z_n, der durch den Faktor F_j erklärt wird.
Varianzanalyse		
Eine metrische abhängige Variable y, m nominale unabhängige Variablen A, B, ... m = 1: Einfache Varianzanalyse m = 2: Zweifache Varianzanalyse, etc.	Streuungszerlegung m = 1 (einfache Varianzanalyse): $ss_y = ss_A + ss_{error}$ ($y_{ij} - \mu = \alpha_i + \varepsilon_{ij}$)	$R^2 = \frac{ss_A}{ss_y}$ = Anteil der Varianz von y, die durch A erklärt wird.
	m = 2 (zweifache Varianzanalyse): $ss_y = ss_A + ss_B + ss_{AB} + ss_{error}$, falls A und B statistisch unabhängig sind.	$\frac{ss_y - ss_{error}}{ss_y}$ = Anteil der Varianz von y, die durch A und B insgesamt erklärt wird. $\frac{ss_{AB}}{ss_y}$ = Anteil der Varianz von y, die durch Interaktion von A und B erklärt wird. $\frac{ss_B}{ss_y}$ = Anteil der Varianz von y, die durch B erklärt wird (für A entsprechend).

1.6 Bewertung

Die **multiple lineare Regression** dürfte das wichtigste Datenanalyseverfahren sein. Das bei der Regressionsanalyse benutzte Konzept der kleinsten Quadrate, um den Fehler einer linearen Schätzung zu messen, ist auch Ausgangspunkt verschiedener anderer Verfahren.

Ein zweites grundlegendes Konzept ist die **Faktorenanalyse**, die eine Regression auf latente, zugrunde liegende Faktoren ist.

Multiple Regression	**Faktorenanalyse**
Korrelierte Regressoren	Unkorrelierte Faktoren
Beobachtete Regressoren	Latente Faktoren

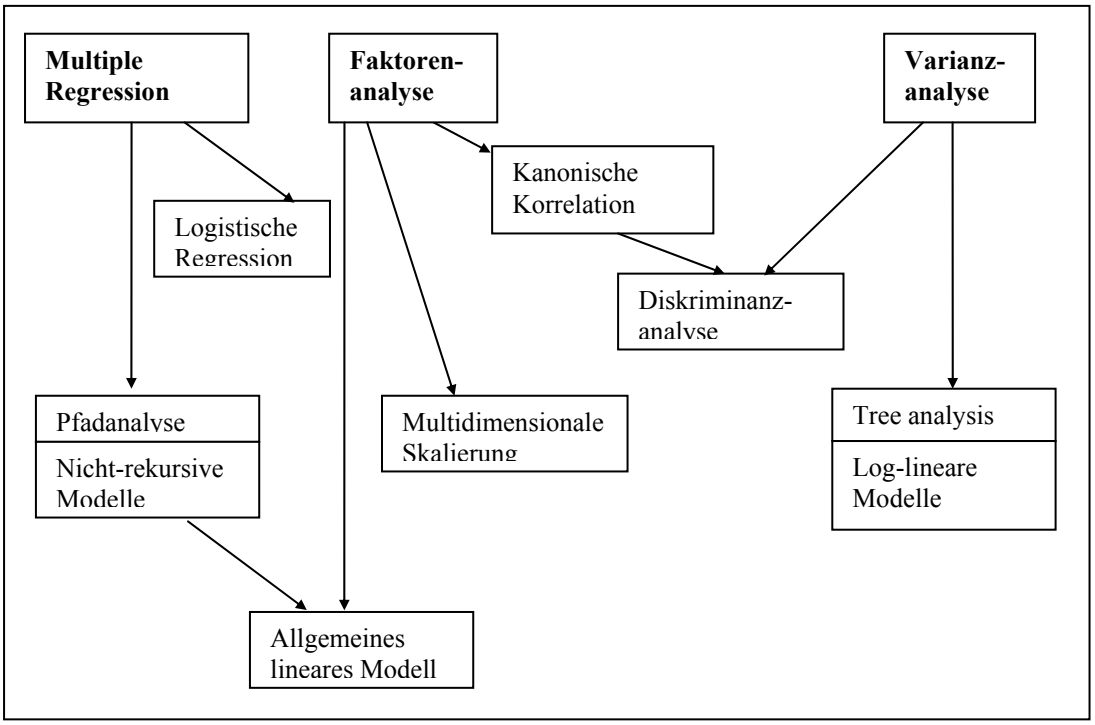
Während die multiple Regression eine möglichst gute Erklärung der Struktur des Zusammenhangs einer beobachteten abhängigen Variablen mit beobachteten unabhängigen Variablen anstrebt, werden in der Faktorenanalyse nicht beobachtete Dimensionen gesucht, die die beobachteten Daten möglichst gut im statistischen Sinne erklären sollen. Da die Faktoren einerseits ein „künstliches Produkt“ sind, insofern sie ja nicht einfach beobachtet werden, kann man andererseits solche Faktoren suchen, die besonders wünschenswerte Eigenschaften haben. Unter dem Gesichtspunkt einer möglichst einfachen Darstellung der Daten strebt man deshalb (i.a.) unkorrelierte Faktoren an. Die beiden Aspekte hängen also zusammen: Weil es sich um „Kunstprodukte“ handelt, kann man von ihnen gewünschte Eigenschaften verlangen. Empirisch beobachtete Variablen, über die man in der Regressionsanalyse nicht hinausgeht, korrelieren dagegen in der Regel.

Als drittes grundlegendes Konzept ist die **Varianzanalyse** zu nennen. Konzeptuell ist sie in ihrer einfachsten Form Bestandteil der multiplen Regressionsanalyse und allgemein ein Spezialfall der multiplen Regressionsanalyse für nominale unabhängige Variablen. Es handelt sich aber keineswegs um einen einfachen Spezialfall, da in der Varianzanalyse in der Regel Interaktionen untersucht werden[1], was schon eine Regressionsanalyse mit mindestens zwei verschiedenen Prädiktor-Mengen und deren Interaktionstermen beinhaltet.

Die meisten anderen Verfahren sind abgeleitet aus diesen drei grundlegenden Konzepten (vgl. Abbildung 1-8).

[1] Man sollte nicht dem Verfahren überlassen, ob man Interaktionen untersucht. Dies lässt sich z.B. auch in der Regressionsanalyse berücksichtigen, dort aber ist es viel weniger verbreitet als in der Varianzanalyse, wo dies „automatisch“ geschieht.

Abbildung 1-8: Multiple Regression, Faktorenanalyse und Varianzanalyse als grundlegende Konzepte



Die multiple Regression, Varianzanalyse und Faktorenanalyse sowie die kanonische Korrelation und Diskriminanzanalyse würde ich als „klassische" Datenanalyseverfahren bezeichnen. Relevante neuere Entwicklungen sind die log-linearen Modelle zur Analyse von Nominal-Daten, die logistische Regression, die multidimensionale Skalierung zur Analyse von Ordinal-Daten, die neuere Rezeption des (ansonsten älteren) Verfahrens der Pfadanalyse, die Entwicklung nicht-rekursiver Modelle und die Entwicklung des allgemeinen linearen Modells.

Wie bereits in der Übersicht in Abbildung 1-1 dargestellt, ist auch die Analyse von Kontingenztafeln ein elementares Konzept. Die χ^2-Analyse für die Kontingenz von 2 Merkmalen gehört (nach der hier benutzten Definition: multivariat = Verwendung von mehr als 2 abhängigen/unabhängigen Variablen) noch nicht zur multivariaten Analyse. Die Erweiterung dieses Konzepts auf mehr als zwei Merkmale erlaubt Aussagen darüber, ob es signifikante Interaktionseffekte gibt. Obwohl die Analyse nominaler Daten in starker Entwicklung ist (vgl. das Jahrbuch „Sociological Methodology" und die Zeitschrift „Sociological Methods and Research"), erscheinen mir die Analyse- und Interpretationsmöglichkeiten als weniger anschaulich als die metrischen Ansätze. Die neueren Entwicklungen der multidimensionalen Skalierung in Nachfolge der Faktorenanalyse findet man vor allem in der Zeitschrift „Psychometrika".

Die Anschlussfähigkeit an die Theoriebildung in den Sozialwissenschaften versprechen vor allem die linearen Kausalmodelle zu leisten. Diese Modellierungen der Sozialwissenschaftler werden insbesondere auch von den Ökonometrikern verfolgt, wo die häufige Voraussetzung, dass es sich um metrische Variablen handelt, am unproblematischsten ist. Die fruchtbarsten Entwicklungen scheinen mir in der Weiterentwicklung pfadanalytischer Modelle mit beobachteten oder auch latenten Variablen zu liegen, da in diesen Modellierungen als Erklärung die Mechanismen von direkten und indirekten Kausalbeiträgen herausgearbeitet werden.

1.7 Pragmatische Abgrenzung von grundlegenden und fortgeschrittenen multivariaten Modellen sowie Aufbau des vorliegenden Bandes

Im vorliegenden Band zu den grundlegenden multivariaten Modellen werden **Tabellenanalyse**, **multiple Regression**, **Pfadanalyse** und **Varianzanalyse** (mit einer abhängigen Variablen und mehreren unabhängigen Variablen) als grundlegende Modelle behandelt. In einem anderen Band werden die „fortgeschrittenen" multivariaten Modelle dargestellt und diskutiert.

Pragmatisch als „fortgeschritten" werden die Ansätze angesehen, zu deren Lösung man ein mathematisches Eigenwertproblem lösen muss und die deshalb hier gemeinsam behandelt werden: **Faktorenanalyse**, **kanonische Korrelation** sowie **Diskriminanzanalyse** und **multivariate Varianzanalyse**.

Die log-linearen und verwandte **Modelle für nicht-metrische Daten** sind später entwickelt worden als die metrischen, was ein Hinweis auf ihre geringere Anschaulichkeit ist. Andererseits haben sie den Vorzug, geringere Anforderungen an das vorausgesetzte Messniveau zu stellen. Die **Korrespondenzanalyse** ist geeignet zur Analyse des Zusammenhangs von nicht-metrischen Daten, wobei an Grundüberlegungen der Faktorenanalyse angeknüpft wird.

Schließlich sind **Clusteranalyse**, **multidimensionale Skalierung** und ähnliches i.a. zusätzliche Gesichtspunkte für die Analyse, weshalb sie im folgenden Band ebenfalls in einem Kapitel behandelt werden.

Die **Mehrebenenanalyse** und die **Analyse zeitbezogener Daten** beinhalten weitere vertiefende Analyseansätze. Der Mehrebenenanalyse liegt die Idee zu Grunde, eine abhängige Variable auf individueller Ebene durch Effekte auf verschiedenen Ebenen zu erklären. Die Analyse zeitabhängiger Daten bringt den Gesichtspunkt der Dynamik sozialen Wandels mit ein. Betrachten die vorangegangenen Kapitel eher statische Daten zu einem Zeitpunkt, stehen hier zeitliche Veränderungen im Mittelpunkt, was ebenfalls im Band über die „fortgeschrittenen" multivariaten Modelle der sozialwissenschaftlichen Datenanalyse behandelt wird.

Literaturverzeichnis

Anderson, T.W., 2003[3]: *An Introduction to Multivariate Statistical Analysis*. New York: Wiley InterScience.

Backhaus, K., Erichson, B., Plinke, W., Weiber, R., 2008[12]: *Multivariate Analysemethoden*. Eine anwendungsorientierte Einführung. Berlin: Springer.

Dixon, W.J., Brown, M.B. (Hg.), 1979[2]: *BMDP: Biomedical computer programs*. Berkeley: University of California Press.

Gaensslen, H., Schubö, W., 1976[2]: *Einfache und komplexe statistische Analyse*. München: UTB.

Holm, K. (Hg.), 1975 f.: *Die Befragung*. 6 Bände. München: UTB.

Küchler, M., 1979: *Multivariate Analyseverfahren*. Stuttgart: Teubner.

Litz, H.P., 2000: *Multivariate statistische Methoden.* München: Oldenbourg.

Morrison, D.F., 2002[4]: *Multivariate Statistical Methods*. Belmont: Duxbury Press.

Nie, N.H. et al., 1975[2]: *Statistical package for the social sciences* (SPSS). New York: McGraw-Hill.

Opp, K.-D., Schmidt, P., 1976: *Einführung in die Mehrvariablenanalyse*. Grundlagen der Formulierung und Prüfung komplexer sozialwissenschaftlicher Aussagen. Reinbek bei Hamburg: Rowohlt.

Rao, C.R., 1973[2]: *Linear Statistical Inference and Its Applications*. New York: Wiley InterScience.

Stuart, A. et al., 2004[6]: *Kendall's Advanced Theory of Statistics*. 3 Bände. London: Arnold.

Van Koolwijk, J., Wieken-Mayser, M., (Hg.), 1986: *Kausalanalyse*. Techniken der empirischen Sozialforschung. Band 8. München, Wien: Oldenbourg.

Van Koolwijk, J., Wieken-Mayser, M., (Hg.), 1974 f.: *Techniken der empirischen Sozialforschung*. 8 Bände. München, Wien: Oldenbourg.

Van de Geer, J.P., 1971: Introduction to Multivariate Analysis for the Social Sciences. San Francisco: Freeman.

2. Kausalanalyse mit Tabellenanalyse und partieller Korrelation

2.1 Kausalanalyse und statistischer Kausalbegriff

Im Unterschied zum Verstehen begreifen kausale Erklärungen bestimmte Erscheinungen als Wirkungen bestimmter Ursachen. Man differenziert zwischen deterministischen und stochastischen kausalen Zusammenhängen. Wenn bei bestimmten Ursachen immer das Auftreten bestimmter Wirkungen behauptet wird, spricht man von deterministischen Aussagen. Stochastisch heißt, dass aufgrund bestimmter Ursachen bestimmte Wirkungen mit einer bestimmten Wahrscheinlichkeit eintreten.

Deterministisch: Immer: Wenn A, dann B.
(Gesetz) (All-Aussage) ($\forall x: Ax \Rightarrow Bx$)
(Ein Teil der naturwissenschaftlichen Aussagen).

Stochastisch: Wenn A, dann mit hoher Wahrscheinlichkeit B. („A begünstigt B“)
(Oder: „strukturiert“)

(Empirische Regelmäßigkeit)

Als Kriterien für die Existenz einer (stochastischen) **Kausalbeziehung** zwischen zwei Merkmalen x und y schlagen Hyman (1955) und Lazarsfeld (1955) vor:

(1) Es gibt einen statistischen Zusammenhang.
(2) x geht y kausal voran, wobei die zeitliche Reihenfolge nicht hinreichend ist.
(3) Es handelt sich nicht um eine „Scheinkorrelation“. D.h. die festgestellte Korrelation muss auch bei Einführung von Dritt- bzw. Kontrollvariablen Bestand haben.

Ad 1) Diese Bedingung ist zu eng, da sie die scheinbare Nicht-Kausalität vernachlässigt. Die Korrelation zwischen x und y wird erst in den Teilgruppen, die nach der Kontrollvariablen z gebildet werden, sichtbar.

Ad 2) Während die statistische Beziehung symmetrisch ist, bildet die Kausalbeziehung einen asymmetrischen Wirkungszusammenhang (kausale Ordnung). Im Experiment ist die kausale Ordnung offensichtlich, da die unabhängige Variable variiert wird (als Stimulus) und der Effekt (Response) gemessen wird.
Die kausale Ordnung lässt sich nicht auf die zeitliche Ordnung reduzieren. Deshalb muss man sich vor dem Fehlschluss „post hoc ergo propter hoc“ (von der zeitlichen Reihenfolge auf die kausale Ordnung) hüten.

Ad 3) Im Experiment werden die Untersuchungs- und Kontrollgruppe nach einem Zufallsverfahren (Wahrscheinlichkeitsauswahl) gebildet, sodass der Einfluss von anderen Variablen als der manipulierten unabhängigen Variablen kontrolliert wird.
In der Tabellenanalyse wird zur Simulation des Experiments versucht, einige relevante Variablen zu kontrollieren. Man kann nie sicher sein, alle relevanten Variablen erfasst zu haben.

Allgemein liegt Kausalität im statistischen Sinn dann vor, wenn die Variationen der Variablen x auch Variationen in der Variablen y erzeugen und alle anderen möglichen Einflussfaktoren konstant gehalten werden. *Unter bestimmten Umständen* können also Korrelationen kausal interpretiert werden.

Kausalität und Experiment

Charakteristisch für ein Experiment ist die Manipulierbarkeit der unabhängigen Variablen und die Kontrolle sonstiger Faktoren, die alternativ neben der unabhängigen Variablen als Störfaktoren auf die abhängige Variable einwirken können. Diese Kontrolle erfolgt durch Rekrutierung von äquivalenten Gruppen (Untersuchungsgruppe und Kontrollgruppe) bezüglich wesentlicher Merkmale (mögliche Störfaktoren). Dabei kommen zwei Techniken zur Anwendung:

1) **Randomisierung:** Die Untersuchungsgruppe und die Kontrollgruppe sind Zufallsstichproben, d.h. mittels einer Wahrscheinlichkeitsauswahl gezogen. Systematische Unterschiede in der Gruppenzusammensetzung werden auf diese Weise vermieden. Deshalb kann man mit großer Wahrscheinlichkeit vermuten, dass alle Drittvariablen in beiden Gruppen gleichartig wirken und der Effekt in der abhängigen Variablen auf den Stimulus zurückzuführen ist.

2) **Matching bzw. Parallelisierung:** Einteilung in Untersuchungs- und Kontrollgruppe derart, dass für einige relevante Variablen die Zuordnung in die beiden Gruppen quasi per Münzwurf durchgeführt wird. Bezüglich bestimmter Merkmalsausprägungen werden jeweils „gleiche" Personen der Untersuchungs- und Kontrollgruppe zugeordnet. Wird z.B. die Gruppenzusammensetzung so „gematcht", dass hinsichtlich der Merkmale Geschlecht und Bildung die Gruppen „parallel" sind, so werden der geschlechts- und bildungsspezifische Störeinfluss auf die abhängige Variable kontrolliert. Bei der Randomisierung gilt dann für alle relevanten Testfaktoren z, beim Matching nur für die berücksichtigten Testfaktoren z: [x z] = 0.[2] Damit haben mögliche Störfaktoren keinen Zusammenhang mit der Gruppeneinteilung.

Deshalb ist das Experiment am geeignetesten, Kausalbeziehungen zu untersuchen. In den Sozialwissenschaften sind Experimente jedoch selten durchführbar. Ferner lassen sich Laborergebnisse nicht einfach auf natürliche Umgebungen verallgemeinern.
Das Experiment ist jedoch das Modell, an dem die anderen Vorgehensweisen gemessen werden, auch wenn es selbst kaum benutzt wird.

<u>Beispiel für ein Experiment</u>

y = Vorurteilsmessung

	Messung t_1	Film t_2	Messung t_3
Experimentalgruppe G_1	$\bar{y}_{G_1,t_1}$	ja	$\bar{y}_{G_1,t_3}$
Kontrollgruppe G_2	$\bar{y}_{G_2,t_1}$	nein	$\bar{y}_{G_2,t_3}$

$\bar{y}_{G_1,t_1} = \bar{y}_{G_2,t_1}$

Effekt des Films:
$\bar{y}_{G_1,t_3} - \bar{y}_{G_2,t_3}$

[2] [xz] steht für eine Messung des Zusammenhangs von x und z, wobei offengelassen werden soll, um welche Maßzahl es sich handelt.

Die häufigste Simulation eines Experiments ist die Tabellenanalyse (und auf metrischem Messniveau die partielle Korrelation).

Den Übergang vom Experiment zur Tabellenanalyse bildet das Ex-post-facto-Experiment: Es wird nur eine Messung durchgeführt, und zwar nach dem Stimulus. In diesem Fall fehlt die Manipulierbarkeit der unabhängigen Variablen. Es besteht die Gefahr des Fehlschlusses von der zeitlichen Reihenfolge auf die kausale Ordnung („post hoc ergo propter hoc"). Wie bei der Tabellenanalyse lässt sich dafür argumentieren, dass viele relevante Variablen in den Sozialwissenschaften nicht manipulierbar sind.

Folgende beide Techniken, Testfaktoren statistisch konstant zu halten, sind besonders wichtig:

1) Tabellenanalyse bzw. Teilgruppenvergleich für nominales Messniveau.
2) Partielle Korrelation für metrisches Messniveau.

2.2 Tabellenanalyse

Die Tabellenanalyse (elaboration; Einführung von Drittfaktoren) beinhaltet den Versuch, durch nachträgliche Homogenisierung mittels statistischer Manipulation des Datenmaterials unabhängige Variablen mit Kausalwirkung zu finden. Im Gegensatz zum Experiment, das durch die Tabellenanalyse approximiert wird, sind die unabhängigen Variablen in der Tabellenanalyse nicht einfach einzeln zu verändern. Die Merkmale treten in Kombinationen auf. („Qualities are blockbooked", Rosenberg 1968).
Auch die zweite entscheidende Bedingung des Experiments, dass die Gruppenbildung nämlich nach einem Zufallsverfahren durchgeführt wird, sodass die Gruppen austauschbar sind und deshalb der Einfluss sonstiger Faktoren kontrolliert wird[3], ist in der Tabellenanalyse nicht erfüllt. In der Tabellenanalyse wird deshalb der Einfluss von einigen relevanten Drittfaktoren getestet.
Die Tabellenanalyse ist ein „Quasi-Experiment", da Teilgruppen verglichen werden, die nicht unter Kontrolle des Forschers entstanden sind. Es besteht die Gefahr von unechten Korrelationen (wegen des möglichen Einflusses von Drittfaktoren). In der Tabellenanalyse wird auf nominalem Messniveau versucht, mögliche Störfaktoren (invalidating factors) zu kontrollieren.

Einwände gegen die Tabellenanalyse sind:
1) Es werden nur einige Variablen kontrolliert, Zusammenhänge werden also nur eingeschränkt getestet.
2) Die Anwendbarkeit ist dadurch begrenzt, dass bei Aufspalten in Teiltabellen die Besetzungszahlen gegen Null gehen. Andererseits beinhaltet dies eine differenzierte Analyse durch Berücksichtigung von Interaktionen.
3) Leitet man die Theorie erst aus den Daten ab, so handelt es sich um eine post factum Interpretation, welche zu flexibel und durch die Daten nicht falsifizierbar ist. Allerdings bietet sich hier nach wie vor die Möglichkeit, die split-half-Methode zu verwenden, also in der einen Hälfte der Daten Hypothesen zu generieren und sie mit der anderen Hälfte zu überprüfen.

Auf der anderen Seite sind einige Variablen kaum (Schicht, Religion, Stadt/Land), andere gar nicht zu manipulieren (Alter, Geschlecht, nationale, ethnische und soziale Herkunft, Stellung in der Geschwisterfolge). Ferner können die Daten im Gegensatz zum Experiment aus einer natürlichen Umgebung kommen. (Es stellt sich die Frage, ob methodisch einwandfreie Experimentalergebnisse außerhalb des Labors überhaupt Geltung haben (externe Validität).) Letztlich hat man trotz der Einwände in vielen Fällen gar keine andere Möglichkeit als eine Ex-Post- Kausalanalyse durchzuführen.

[3] Die „ceteris paribus"-Klausel ist erfüllt, wenn die übrigen Einflussfaktoren konstant bleiben.

So verwendete beispielsweise E. Durkheim (1897) in seiner Selbstmord-Studie „Le Suicide“ Faktoren, die die Selbstmordrate beeinflussen.
Nach Durkheim fällt mit wachsender Integration die Selbstmordrate. Als Indikatoren werden benutzt:

1) Religion (Juden sind stärker (untereinander) integriert als Katholiken, Katholiken stärker als Protestanten),
2) Familienstand (Verheiratete sind stärker integriert als Unverheiratete),
3) Stadt/Land (Auf dem Land ist die Integration stärker als in der Stadt.).

Während für die Darstellung von Ergebnissen Tabellen sehr geeignet sind, insbesondere weil sie leichter zu verstehen sind, eignen sich für die Analyse komplexer Zusammenhänge lineare statistische Verfahren wohl noch besser, die allerdings metrisches Messniveau voraussetzen.

2.2.1 Ein Beispiel für Korrelation ohne Kausalität (Scheinkorrelation/scheinkausale Korrelation)

Tabelle 2-1: Beispiel nach Hirschi und Selvin
(Stichprobe von Jugendlichen)

		x: Kirchenbesuch Ja	x: Kirchenbesuch Nein
y: Delinquenz	Ja	44 %	56 %
	Nein	56 %	44 %
		100 %	100 %

$[xy] \neq 0$

z_1: Alter ≤ 14

		x: Kirchenbesuch Ja	x: Kirchenbesuch Nein
y: Delinquenz	Ja	33 %	33 %
	Nein	67 %	67 %
		100 %	100 %

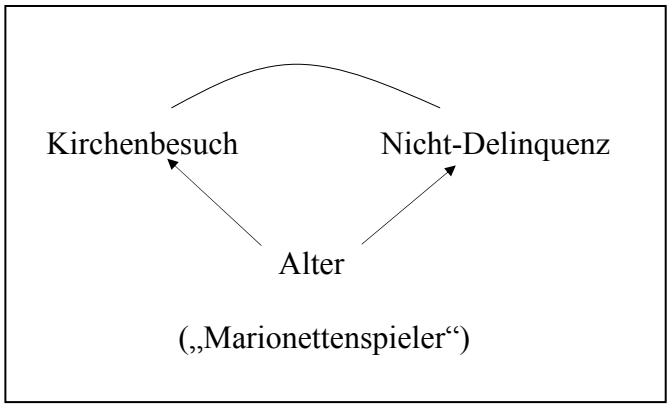
$[xy: z_1] = 0$

z_2: Alter > 14

		x: Kirchenbesuch Ja	x: Kirchenbesuch Nein
y: Delinquenz	Ja	67 %	67 %
	Nein	33 %	33 %
		100 %	100 %

$[xy: z_2] = 0$

Abbildung 2-1: „Scheinkorrelation“



„Scheinkorrelation“ (Tatsächliche Korrelation, aber nicht kausal zu interpretieren): Von den untersuchten Jugendlichen sind die Jüngeren eher Kirchenbesucher und gleichzeitig nicht delinquent, ohne dass zwischen den beiden letztgenannten Merkmalen eine Kausalbeziehung vorliegt. Das Alter (als „Marionettenspieler“) verursacht die Kovariation.

Bleibt der Zusammenhang in dem „Test“ der Einführung einer dritten Variablen erhalten, so ist der Zusammenhang vorläufig „verifiziert“ – genauer: hat sich bis auf weiteres bewährt –, ansonsten falsifiziert.

Ein solches Einführen von Test- oder Kontrollvariablen ist der klassische Ansatz der „elaboration analysis“ oder Tabellenanalyse.

Eine bekannte Einführung ist die Arbeit von Morris Rosenberg (1968), aus der auch die folgenden Beispiele zur Tabellenanalyse stammen, sofern nicht anders vermerkt.

2.2.2 Zerlegung der Vier-Felder-Tafel an einem einführenden Beispiel

Ein Beispiel zur Einführung (nach Mayntz/Holm/Hübner 1971: 203 f.):
Wie ist es möglich, dass ein Zusammenhang bei Aufgliederung nach einem Drittfaktor verschwindet?
Nach einer industriesoziologischen Untersuchung (H. Zeisel: Probleme der Aufschlüsselung. In: R. König (Hrsg.): Das Interview. Köln 1957: 305) wurde ein Zusammenhang zwischen dem Familienstand der Arbeiterinnen (x) und der Häufigkeit ihres Fernbleibens von der Arbeit (y) festgestellt: Verheiratete Frauen bleiben häufiger der Arbeit fern als Ledige (vgl. Tabelle ***2-2***).

x	➔	z	➔	y
Familienstand		Hausarbeit		Fernbleiben vom Betrieb
(led./verh.)		(wenig/viel)		(wenig/viel)

Tabelle 2-2:

		Familienstand (x)		
		ledig	verheiratet	
Häufigkeit des Fern-bleibens v. Betrieb (y)	wenig	1000	600	1600
	viel	600	1000	1600
		1600	1600	3200

Differenz der Kreuzprodukte [xy] = 1000^2 - 600^2 (Anzahl der konkordanten Paare minus Anzahl der diskordanten Paare)
[xy] > 0 ([xy] noch keine normierte Maßzahl)
Kausalmodell: x → y

Nun könnte man z.B. die Hypothese aufstellen, dass der Umfang der Hausarbeit der Frauen deren betriebliche Anwesenheit beeinflusst. Als Test dieser Hypothese wird der Umfang der Hausarbeit als Kontrollvariable (z) eingeführt. Es ergeben sich - jeweils für „viel“ und „wenig“ Hausarbeit - zwei Teiltabellen (vgl. Tabelle 2-3).

Tabelle 2-3:

z_1
(wenig Hausarbeit)

	led.	verh.	
wenig	900	300	1200
viel	300	100	400
	1200	400	1600

Kein Zusammenhang zwischen x und y:
Je ein Viertel der Ledigen und der Verheirateten bleiben viel fern.
[xy : z_1] = 0

z_2
(viel Hausarbeit)

	led.	verh.	
wenig	100	300	400
viel	300	900	1200
	400	1200	1600

Kein Zusammenhang zwischen x und y:
Je drei Viertel der Ledigen und der Verheirateten bleiben viel fern.
[xy : z_2] = 0

(Technisch liegt es also an den Randverteilungen, inhaltlich an den Beziehungen zu dem Drittfaktor, dass der Zusammenhang in den Teilgruppen verschwindet.)

Die beiden bedingten Korrelationen ergeben Null.
Es handelt sich hier um ein Beispiel einer intervenierenden Variablen z (vgl. dazu genauer 2.2.4).
Differenzierteres Kausalmodell:

x → z → y

(Es gibt einen **indirekten** Kausaleffekt von x auf y, aber **keinen direkten**.)

z ist eine intervenierende Variable.

Familienstand (x) strukturiert Belastung mit Hausarbeit. Belastung mit Hausarbeit strukturiert Fernbleiben vom Betrieb.

Tabelle 2-4:

		Familienstand (x)		
		ledig	verheiratet	
Hausarbeit (z)	wenig	1200	400	1600
	viel	400	1200	1600
		1600	1600	3200

[xz] > 0

Tabelle 2-5:

		Hausarbeit (z)		
		wenig	viel	
Häufigkeit des Fernbleibens vom Betrieb (y)	wenig	1200	400	1600
	viel	400	1200	1600
		1600	1600	3200

[zy] > 0

Zur Entdeckung eines Zusammenhangs zwischen zwei Merkmalen reicht eine zweidimensionale Tabelle (Kreuztabellierung). Führt man einen Drittfaktor (Testfaktor) ein, so benötigt man eine dreidimensionale Tabelle. Die Anzahl der Dimensionen der Tabelle entspricht also der Anzahl der Merkmale, während die Anzahl der Ausprägungen dieser Merkmale beliebig sein kann.

Die Zusammenhänge beim Einführen einer dritten Variablen lassen sich jedoch am einfachsten in dem elementarsten Fall von Dichotomien (also 2 x 2-Tabellen) untersuchen, weshalb in diesem Kapitel 2.2 immer von Vier-Felder-Tafeln ausgegangen wird (vgl. Tabelle 2-6).

Tabelle 2-6:

		x	
		x_1	x_2
y	y_1	a	b
	y_2	c	d

Eine Vier-Felder-Tafel lässt sich durch die Einführung einer (dichotomen) Kontrollvariablen z in zwei Teil-Tabellen zerlegen[4].

Tabelle 2-7:

z_1:

		x	
		x_1	x_2
y	y_1	a_1	b_1
	y_2	c_1	d_1

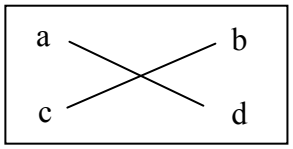
z_2:

		x	
		x_1	x_2
y	y_1	a_2	b_2
	y_2	c_2	d_2

Dabei gilt, dass sich die Zellbesetzungen der Teiltabellen immer zur Zellbesetzung der Gesamttabelle addieren:
$a_1 + a_2 = a$
$b_1 + b_2 = b$
$c_1 + c_2 = c$
$d_1 + d_2 = d$

Verwendet man als Zusammenhangsmaß [xy] die **Differenz der Kreuzprodukte**[5], so erhält man:
$[xy] = ad - bc$

Kreuzprodukte:



Und für die beiden Teiltabellen ergeben sich die bedingten Zusammenhänge:

$[xy : z_1] = a_1d_1 - b_1c_1, \quad [xy : z_2] = a_2d_2 - b_2c_2$

Es wird sich zeigen, dass sich nicht nur die Gesamttabelle in zwei Tabellen, sondern dass sich auch der Gesamtzusammenhang in verschiedene Komponenten zerlegen lässt. Dabei ist die Beziehung etwas komplizierter als bei der einfachen Addition von zwei Teiltabellen zu einer Gesamttabelle.

Für diese Beziehung braucht man noch die Zusammenhänge von x und y mit dem Drittfaktor z (vgl. Tabelle 2-8).

Tabelle 2-8:

		x	
		x_1	x_2
z	z_1	$a_1 + c_1$	$b_1 + d_1$
	z_2	$a_2 + c_2$	$b_2 + d_2$

		z	
		z_1	z_2
y	y_1	$a_1 + b_1$	$a_2 + b_2$
	y_2	$c_1 + d_1$	$c_2 + d_2$

[4] Bei metrischem Messniveau erfüllt die partielle Korrelation die gleiche Funktion.
[5] Die eigentlichen Maßzahlen für die Vierfeldertafel unterscheiden sich genau durch die Normierung dieser Differenz.

Das Kreuzprodukt lautet:
$[xz] = (a_1 + c_1)(b_2 + d_2) - (a_2 + c_2)(b_1 + d_1)$ $\quad [zy] = (a_1 + b_1)(c_2 + d_2) - (c_1 + d_1)(a_2 + b_2)$

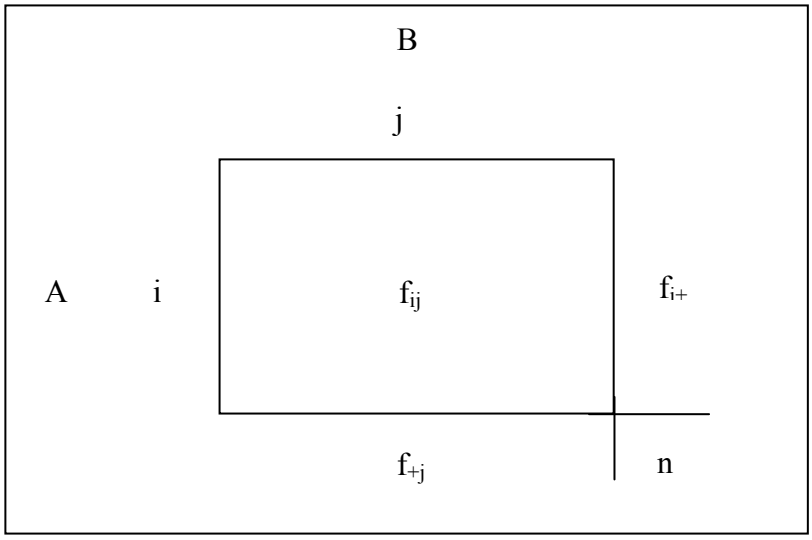
Es wird in (2.2.3) gezeigt werden, dass dann die Beziehung besteht:
$[xy] = \alpha [xy : z_1] + \beta [xy : z_2] + \gamma [xz] [yz]$, wobei α, β, γ positive Koeffizienten sind.

2.2.2.1 Log-lineare Modellierung des einführenden Beispiels

Das einführende Beispiel für eine intervenierende Variable beinhaltete, dass die Belastung durch Hausarbeit in der Kausalbeziehung zwischen Familienstand und Fernbleiben im Betrieb insofern interveniert, dass ledige Frauen stärker belastet sind durch Hausarbeit und dass stärker durch Hausarbeit belastete Frauen häufiger dem Betrieb fern bleiben müssen. Dieses Beispiel soll nun dadurch modelliert werden, welche der möglichen Interaktionen zwischen diesen drei Merkmalen berücksichtigt werden müssen, um die beobachteten (Kombinations-) Daten zu reproduzieren. Da man eine sparsame Modellierung anstrebt, will man nur die zwingend notwendige Interaktionen herausarbeiten.

Log-lineare Modelle zur Analyse von Kreuztabellen

Abbildung 2-2: Kreuztabelle zweier Merkmale A und B:



(Notation: f wie frequencies)

Das Konzept der statistischen Unabhängigkeit $f_{ij} = n \cdot \frac{f_{i+}}{n} \cdot \frac{f_{+j}}{n}$ ist eigentlich ein multiplikatives Konzept. Durch Logarithmieren lässt sich die Unabhängigkeit additiv ausdrücken:

$$\ln f_{ij} = \ln n + \ln \frac{f_{i+}}{n} + \ln \frac{f_{+j}}{n}$$

Der Soziologe und Statistiker Leo A. Goodman (University of Chicago) hat zur Analyse von Kontingenztafeln die log-lineare Modellierung entwickelt (1970, 1972 etc.).

Mit f_{ij} werden die beobachteten Häufigkeiten bezeichnet, mit F_{ij} die unter der Annahme des Modells zu erwartenden Häufigkeiten.
Die Idee dieser Modellierung besteht nun darin, dass die logarithmischen Häufigkeiten log F_{ij} sich aus dem additiven Zusammenwirken eines Bezugspunktes θ (theta) und der Effekte der Randverteilungen λ_i^A bzw. λ_j^B (lambda) bzw. der Effekte der Interaktion λ_{ij}^{AB} darstellen lassen.

Multiplikativ:

$$\underbrace{f_{ij} = n}_{\text{Stichproben-umfang}} \cdot \underbrace{\frac{f_{i+}}{n}}_{\text{Effekt von } A_i} \cdot \underbrace{\frac{f_{+j}}{n}}_{\text{Effekt von } B_j} \cdot \underbrace{\frac{f_{ij}/n}{f_{i+}/n \cdot f_{+j}/n}}_{\text{Effekt der Interaktion } A_iB_j \text{ (Abweichung von der Unabhängigkeit)}}$$

Additiver Modellansatz für die Grundgesamtheit: $\ln F_{ij} = \theta + \lambda_i^A + \lambda_j^B + \lambda_{ij}^{AB}$

Das Modell der statistischen Unabhängigkeit lässt sich bei dieser Modellierung also dadurch charakterisieren, dass $\lambda_{ij}^{AB} = 0$ (für alle i, j).

Beispiel: Fernbleiben vom Betrieb

Die pfadanalytische Interpretation des Beispiels lautete:

Familienstand (A) strukturiert Belastung mit Hausarbeit (B).
Belastung mit Hausarbeit (B) strukturiert Fernbleiben vom Betrieb (C).

Drei Merkmale lassen sich z.B. in der Form einer dreidimensionalen Tabelle (A, B, C) analysieren. Das (vollständige) log-lineare Modell würde lauten:

$$\ln F_{ijk} = \theta + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{ik}^{AC} + \lambda_{jk}^{BC} + \lambda_{ijk}^{ABC}$$

Das vollständige oder saturierte Modell umfasst also alle denkmöglichen Parameter und ist nur eine (evtl. sogar zu umfangreiche) Umformung der Daten.

Die Idee der log-linearen Modellierung besteht nun darin, eine „sparsame“ Modellierung vorzunehmen, d.h. ein so einfaches Modell wie nur möglich, wobei Einfachheit bedeutet, mit Parametern möglichst geringer Ordnung auszukommen (λ_i^B ist z.B. 1. Ordnung, λ_{ij}^{AB} ist 2. Ordnung, λ_{ijk}^{ABC} ist 3. Ordnung).

Es lässt sich nun zeigen, dass in dem Beispiel die Randverteilungen A, B, C, (AB), (BC) hinreichend sind („suffiziente Statistiken“), um die Daten des Beispiels perfekt zu reproduzieren.

Tabelle 2-9:

Belastung mit Hausarbeit (B)	**Familienstand (A)**	**Fernbleiben (C)**	
		C_1 wenig	**C_2 viel**
B_1 wenig	A_1 ledig	900	300
	A_2 verheiratet	300	100
B_2 viel	A_1 ledig	100	300
	A_2 verheiratet	300	900

Die Parameter des log-linearen Modells werden mit Hilfe der Maximum-Likelihood-Methode (vgl. den Band zur Deskriptiv- und Inferenzstatistik) geschätzt, dies ist gerade die Leistung des Modells von Goodman.

Tabelle 2-10: Schätzungen der log-linearen Parameter für das Beispiel

Theta (Mean) 5,704

Effekt		Lambda
A	A_1	0
	A_2	0
B	B_1	0
	B_2	0
C	C_1	0
	C_2	0
AB	A_1B_1	.549
	A_1B_2	-.549
	A_2B_1	-.549
	A_2B_2	.549
BC	B_1C_1	.549
	B_1C_2	-.549
	B_2C_1	-.549
	B_2C_2	.549

D.h. es gibt keine Haupteffekte, sondern nur die beiden Interaktionseffekte (AB) und (BC). Zusammenhänge ungleich Null liegen nur vor für die Beziehung Familienstand und Belastung sowie die Beziehung Belastung und Fernbleiben.

Ein Beispiel für die perfekte Anpassung des Modells als Illustration, dass die gemäß dem Modell zu erwartenden Häufigkeiten F_{ijk} den beobachteten Häufigkeiten f_{ijk} genau entsprechen:

$$\ln F_{A_1,B_2,C_1} = \theta + \lambda_1^A + \lambda_2^B + \lambda_1^C + \lambda_{12}^{AB} + \lambda_{21}^{BC}$$
$$= 5{,}704 + 0 + 0 + 0 + (-0{,}549) + (-0{,}549)$$
$$= 4{,}606$$

$$F_{A_1,B_2,C_1} = e^{4{,}606} = 100 = f_{A_1,B_2,C_1}$$

2.2.3 Die Grundgleichung (Zerlegungsformel für Maßzahlen)

In einer sogenannten Grundgleichung lässt sich die Beziehung zwischen den Zusammenhangsmaßen der Teiltabellen und der Gesamttabelle beschreiben.
Es gibt drei bekanntere Versionen der Zerlegungsformel: von Yule für Delta, von Kendall und Lazarsfeld für die Differenz der Kreuzprodukte und von Davis für die Maßzahl Q (von Yule).

1) Zerlegung für **Delta** (nach Yule):

$$\delta_{xy} = \delta_{xy:z_1} + \delta_{xy:z_2} + \frac{n}{n_{z_1} n_{z_2}} \delta_{xz} \delta_{yz}$$

Delta ist allerdings noch keine normierte Maßzahl.

2) Der „klassische Ansatz“ von Lazarsfeld für die **Differenz der Kreuzprodukte** |ij|.
Die Ausgangsdaten seien in der Form von Anteilen statt absoluter Anzahlen gegeben $\left(p_{ij} = \frac{n_{ij}}{n} \; statt \; n_{ij} \right)$, was eine implizite Normierung ist.

Nach Lazarsfeld (1966) erhält man für den Drittfaktor $(k, \bar{k})$:

$$|ij| = \frac{|ij:k|}{p_k} + \frac{|ij:\bar{k}|}{p_{\bar{k}}} + \frac{|ik|\,|jk|}{p_k\, p_{\bar{k}}}$$

Beweis:

Ausgangs-matrix:

p_{ij}	$p_{i\bar{j}}$	p_i
$p_{\bar{i}j}$	$p_{\bar{i}\bar{j}}$	p_j
p_j	$p_{\bar{j}}$	1

Die Differenz der Kreuzprodukte $|ij| := \det\begin{pmatrix} p_{ij} & p_{i\bar{j}} \\ p_{\bar{i}j} & p_{\bar{i}\bar{j}} \end{pmatrix} := \begin{vmatrix} p_{ij} & p_{i\bar{j}} \\ p_{\bar{i}j} & p_{\bar{i}\bar{j}} \end{vmatrix}$ lässt sich nach Determinantenregeln durch Addition der 1. zur 2. Zeile und anschließend der 1. zur 2. Spalte berechnen als:

$$|ij| = \begin{vmatrix} p_{ij} & p_{i\bar{j}} \\ p_j & p_{\bar{j}} \end{vmatrix} = \begin{vmatrix} p_{ij} & p_i \\ p_j & 1 \end{vmatrix} = p_{ij} - p_i \cdot p_j$$

Entsprechend: $|ij:k| = \begin{vmatrix} p_{ijk} & p_{i\bar{j}k} \\ p_{\bar{i}jk} & p_{\bar{i}\bar{j}k} \end{vmatrix} = \begin{vmatrix} p_{ijk} & p_{ik} \\ p_{jk} & p_k \end{vmatrix}$ *und* $|ij:\bar{k}|$.

Lazarsfelds Beweisidee für einen allgemeineren Fall lautet in diesem Spezialfall:

$$|ij:k| = p_{ijk}\,p_k - p_{ik}\,p_{jk},\; also: p_{ijk} = \frac{|ij:k| + p_{ik}\,p_{jk}}{p_k}$$

Entsprechend: $p_{ij\bar{k}} = \frac{|ij:\bar{k}| + p_{i\bar{k}}\,p_{j\bar{k}}}{p_{\bar{k}}}$

Mit Hilfe von $p_{ij} = p_{ijk} + p_{ij}$ folgt dann:

$$|ij| = \begin{vmatrix} p_{ij} & p_i \\ p_j & 1 \end{vmatrix} = \begin{vmatrix} \frac{|ij:k|}{P_k} + \frac{|ij:\bar{k}|}{p_{\bar{k}}} + \frac{p_{ik}\,p_{jk}}{p_k} + \frac{p_{i\bar{k}}\,p_{j\bar{k}}}{p_{\bar{k}}} & p_i \\ p_j & 1 \end{vmatrix}$$

$$= \frac{|ij:k|}{p_k} + \frac{|ij:\bar{k}|}{p_{\bar{k}}} + \frac{p_{ik}p_{jk}}{p_k} + \frac{p_{i\bar{k}}p_{j\bar{k}}}{p_{\bar{k}}} - p_i p_j$$

Ferner: $p_{ik}p_{jk}p_{\bar{k}} + p_{i\bar{k}}p_{j\bar{k}}p_k - p_i p_j p_k p_{\bar{k}}$

$$= p_{ik}p_{jk} - p_{ik}p_{jk}p_k + p_i p_j p_k - p_i p_{jk} p_k - p_j p_{ik} p_k$$
$$+ p_{ik}p_{jk}p_k - p_i p_j p_k + p_i p_j p_k^2$$
$$= p_{ik}(p_{jk} - p_j p_k) - p_i p_k (p_{jk} - p_j p_k) = |ik| \cdot |jk|$$

∟

Statt dieser „klassischen Lösung“ spricht für die Zerlegungsformel für die **Kovarianz**, dass diese allgemeiner verwendbar ist und in anderen Kontexten eine große Rolle spielt (vgl. Punkt 2.3 und Kap. 4.8.1).

⌜

Daraus erhält man die Zerlegung für $*$ durch folgende Überlegung:

$$|ij| = \frac{n_{ij}}{n} - \frac{n_i}{n}\frac{n_j}{n} = \frac{\delta_{ij}}{n} \qquad (= \text{Kovarianz})$$

Ferner: $|ij:k| = \frac{n_{ijk}}{n}\frac{n_k}{n} - \frac{n_{ik}}{n}\frac{n_{jk}}{n} \quad und: \delta_{ij:k} = n_{ijk} - \frac{n_{ik}n_{jk}}{n_k}$

Also: $|ij:k| = \frac{n_k}{n^2}\delta_{ij:k}$

Lazarsfelds Gleichung führt somit für $*$ zu:

$$\frac{\delta_{ij}}{n} = \frac{\delta_{ij:k}}{n} + \frac{\delta_{ij:\bar{k}}}{n} + \frac{\frac{\delta_{ik}}{n}\frac{\delta_{jk}}{n}}{\frac{n_k}{n}\frac{n_{\bar{k}}}{n}}$$

Also: $\delta_{ij} = \delta_{ij:k} + \delta_{ij:\bar{k}} + \frac{n}{n_k n_{\bar{k}}}\delta_{ik}\delta_{jk}$

∟

Die Zerlegungsformel für δ ist wohl bzgl. der Normierung die einfachste, d.h. noch einfacher als für die Kovarianz s_{ij} (da $s_{ij} = \frac{\delta}{n}$):

$$\left(n_{ij} - \frac{n_i n_j}{n}\right) = \left(n_{ijk} - \frac{n_{ik}n_{jk}}{n_k}\right) + \left(n_{ij\bar{k}} - \frac{n_{i\bar{k}}n_{j\bar{k}}}{n_{\bar{k}}}\right) + \frac{n}{n_k n_{\bar{k}}}\left(n_{ik} - \frac{n_i n_k}{n}\right)\left(n_{jk} - \frac{n_j n_k}{n}\right)$$

3) Davis' Zerlegungsformel für die Maßzahl **Q** (von Yule):

$Q_{xy} = W_1 * Q_{xy\,:\,TIED\,z} + W_2 * Q_{xy\,:\,DIFF\,z}$

Hierbei sind W_1, W_2 Gewichte, und zwar die Anteile der Paare, die in z übereinstimmen bzw. differieren. Nach Davis (1971) ist
$Q_{xy : TIED\ z}$ = Partial,
$Q_{xy : DIFF\ z}$ = Differential.

Unter Partial versteht man eine gewichtete Summe zweier bedingter Koeffizienten, sodass man erhält:
$Q_{xy} = W_{1a} * Q_{xy : z} + W_{1b} * Q_{xy : NOT\ z} + W_2 * Q_{xy : DIFF\ z}$

Wie schon im Band zur Deskriptiv- und Inferenzstatistik erörtert wurde, ist Q ein sehr grobes Maß. Andererseits erhöht der von Davis gefundene Zusammenhang das theoretische Verständnis der Zerlegung eines Gesamtzusammenhangs.

Insgesamt:
0) Allgemeine Form der Zerlegung:
$[x, y] = \alpha\,[x, y : z_1] + \beta\,[x, y : z_2] + \gamma\,[x, z]\,[y, z]$

1) Da die Differenz der Kreuzprodukte die Basis von Delta, Phi, Kovarianz, Prozentsatzdifferenz und Yules Q ist, gibt es für alle diese Kennzahlen Zerlegungsformeln, welche sich nur durch die Koeffizienten unterscheiden. In diesen Zerlegungsformeln wird jeweils der Gesamtzusammenhang zweier Variablen zerlegt in einen bedingten Zusammenhang (Zusammenhang unter der Bedingung des Testfaktors z) und in das Produkt der Beziehungen zwischen den beiden Variablen und dem Testfaktor.

2) Mit einer solchen Grundgleichung kann man dann jedes spezielle Beispiel untersuchen.

Beispiel:
Da δ nicht normiert ist, soll nun eine analoge Zerlegung für die Maßzahl Φ **(Phi)** formuliert werden.

$$\delta = \frac{\sqrt{S_1 S_2 S_3 S_4}}{n}\,\Phi$$

$$\left[\text{Weil } \Phi = \frac{ad - bc}{\sqrt{S_1 S_2 S_3 S_4}}\right]$$

$$\Phi(x,y) = \frac{n}{n_{z_1}} \frac{\sqrt{n_{x_1:z_1} \cdot n_{x_2:z_1} \cdot n_{y_1:z_1} \cdot n_{y_2:z_1}}}{\sqrt{n_{x_1} \cdot n_{x_2} \cdot n_{y_1} \cdot n_{y_2}}} \Phi(xy : z_1)$$

$$+ \frac{n}{n_{z_2}} \frac{\sqrt{n_{x_1:z_1} \cdot n_{x_2:z_2} \cdot n_{y_1:z_2} \cdot n_{y_2:z_2}}}{\sqrt{n_{x_1} \cdot n_{x_2} \cdot n_{y_1} \cdot n_{y_2}}} \Phi(xy : z_2)$$

$$+ \Phi(xz)\,\Phi(yz)$$

Tabelle 2-11: Beispiel

		Familienstand x		
Fernbleiben im Betrieb y		1000	600	1600
		600	1000	1600
		1600	1600	

$\Phi_{xy} = 0{,}25$

$$\left(\text{Nämlich}: \Phi_{xy} = \frac{1000^2 - 600^2}{\sqrt{1600^4}} \right)$$

Familienstand x

z Haus-arbeit		led.	verh.	
	wenig	1200	400	1600
	viel	400	1200	1600
		1600	1600	3200

$\Phi_{xz} = 0{,}5$

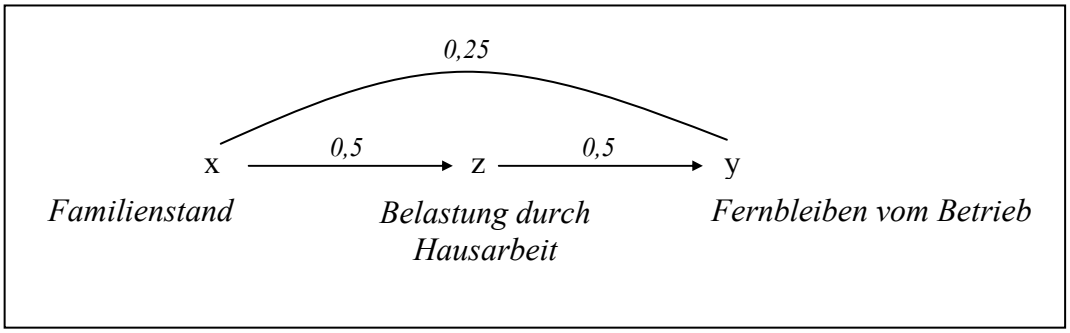
Hausarbeit z

y Fern-bleiben im Betrieb		led.	verh.	
	wenig	1200	400	1600
	viel	400	1200	1600
		1600	1600	3200

$\Phi_{yz} = 0{,}5$

$$\underbrace{\Phi_{xy}}_{0{,}25} = \alpha \cdot \underbrace{\Phi\ (xy : z_1)}_{0} + \beta \cdot \underbrace{\Phi\ (xy : z_2)}_{0} + \underbrace{\Phi\ (xz)}_{0{,}5}\ \underbrace{\Phi\ (yz)}_{0{,}5}$$

Abbildung 2-3: Pfadmodell
(Hier: Kausalkette)



- Es gibt keinen direkten Kausaleffekt, sondern nur einen indirekten.
  Bei Egalisierung der Belastung durch Hausarbeit gäbe es keinen Kausaleffekt von Familienstand auf Fernbleiben im Betrieb mehr.

- Als intervenierende Variablen kommen nur Faktoren z in Frage, die sehr hoch mit x und y korrelieren.

  (Denn ein Produkt von Zahlen $|\Phi| < 1$ wird ja kleiner.) („Scheinkorrelation“: analog)

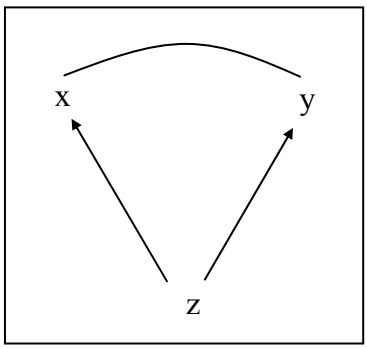
2.2.4 Typologie von Kausalstrukturen mit drei Variablen

2.2.4.1 Typen mit: [xy : z] = [xy : ¬z] („Scheinkorrelation“, Intervenierende Variable, Suppressor, Distorter)

Durch ¬z (lies: non z) wird angezeigt, dass das Attribut z nicht vorhanden ist.

1) „**Scheinkorrelation**“, d.h. tatsächliche Korrelation, welche aber nur scheinbar einen kausalen Zusammenhang ausdrückt. (Andere Bezeichnungen sind: Scheinkausale Korrelation, Störvariable, Explanation, explaining away a relation.)
   Die Kausalstruktur ist in Form eines Diagramms dargestellt (Abbildung 2-4).

Abbildung 2-4: „Scheinkorrelation“



   x und y korrelieren, weil sie beide durch z verursacht werden.
   Idealtypische statistische Beziehungen sind: $[xy] \neq 0$ und $[xy : z] = [xy : \neg z] = 0$
   (Nach der Zerlegungsformel gilt also: $[xy] = \alpha \cdot [xz] \cdot [yz]$ z muss also mit x und y korrelieren.)
   Falls $[xy] > 0$, so: sign [xz] = sign [yz];
   Falls $[xy] < 0$, so: sign [xz] = - sign [yz].

   Beispiele:
   a) x = Anzahl der Feuerwehrpumpen; y = Höhe des Schadens; z = Größe des Feuers.
   b) x = Anzahl der Störche; y = Anzahl der Kinder; z = Stadt/Land. (Auf dem Land gibt es mehr Störche und mehr Geburten.)
   c) Die Sterberate (y) ist im Krankenhaus (x) höher als außerhalb. z = Gesundheitszustand.
   d) (Hirschi und Selvin)
      x = mittlere Kinder; y = Delinquenz; z = Kinderzahl der Herkunftsfamilie. (Mit der Kinderzahl n wächst auch der Anteil $\frac{n-2}{n} = 1 - \frac{2}{n}$ der mittleren Kinder.)

2) **Intervenierende Variable** (Andere Bezeichnungen: Interpretation, understanding, links in the causal chain.)

   Kausalstruktur: $x \rightarrow z \rightarrow y$

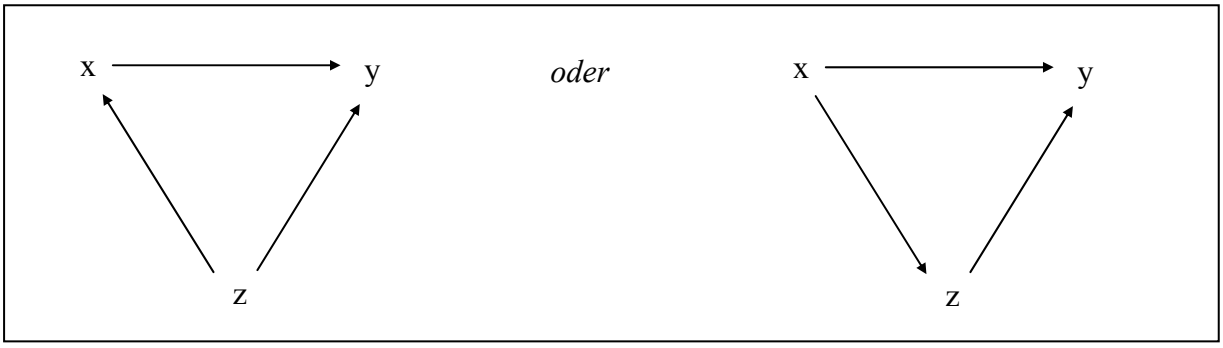
Eine Änderung in x induziert eine Änderung in z und *dadurch* eine Änderung in y. Wird also z konstant gehalten, so verschwindet der Zusammenhang von x und y. Die Ausgangsbeziehung zwischen x und y ist nur durch die zwischengeschaltete z-Variable vermittelt und kann deshalb nicht als direkt kausal interpretiert werden.
Mit weiteren intervenierenden Variablen ließe sich die Kausalkette erweitern.

Statistische Beziehungen:
$[xy] \neq 0, \quad [xy : z] = [xy : \neg z] = 0.$
(Nach der Zerlegungsformel gilt also: $[xy] = \alpha \cdot [xz] \cdot [yz]$)
Voraussetzung ist, dass x und y miteinander korrelieren.
Aufgrund der statistischen Beziehungen sind Scheinkorrelation und intervenierende Variable nicht zu unterscheiden. Nur die kausale Ordnung ist in den beiden Fällen verschieden.
Im o.g. Beispiel von Mayntz et al. (1971) für eine Stichprobe von Arbeiterinnen ist die intervenierende Variable für den Zusammenhang zwischen dem Familienstand (x) und dem betrieblichen Absentismus (y) der Umfang häuslicher Arbeit der Frauen (z).

Beispiele:
a) Intervenierende Variable bereichern die Theorie, da sie das Verständnis erhöhen: Protestanten sind nach Durkheim stärker selbstmordgefährdet, aber weshalb? Die intervenierende Variable z ist der Grad der Integration. Wären Katholiken nicht stärker integriert, so wären sie nicht weniger gefährdet.
b) Verheiratete sind weniger selbstmordgefährdet, aber an welchem Aspekt liegt das? Der Hauptaspekt z ist das Vorhandensein von Kindern (mit der entsprechenden Verantwortung).
c) (Zeisel; Mayntz et al.)
Verheiratete Frauen bleiben eher dem Betrieb fern als unverheiratete. Die intervenierende Variable z ist der Umfang der Hausarbeit.
d) Frauen scheinen autoritärer zu sein als Männer. Wird jedoch die Bildung (z) kontrolliert, so verschwindet der Zusammenhang. (Die weibliche Geschlechtsrolle führt zu einer Benachteiligung in Bildung und Ausbildung.)
e) (Davis)
Nach Homans führt räumliche Nähe zu Zuneigung. Dies wird vermittelt durch das Ausmaß der Interaktionen.

3) **Suppressorvariable** (Dämpfende Variable)
Die „wahre Stärke“ der Relation wird hier durch den Testfaktor T unterdrückt.
Die Abbildung 2-5 stellt die Kausalstruktur dar.

Abbildung 2-5: Suppressorvariable

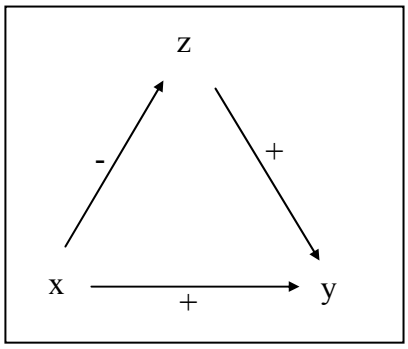


Dem liegen folgende statistische Beziehungen zugrunde: Nach Einführung des Drittfaktors z wird der Zusammenhang zwischen x und y größer als vorher. War [xy] = 0, so handelt es sich im folgenden Sinn um das Gegenteil einer Scheinkorrelation: Während bei einer Scheinkorrelation der ursprüngliche Zusammenhang bei Einführung des Drittfaktors verschwindet, entsteht hier erst ein Zusammenhang nach Einführen des Drittfaktors. Es handelt sich also um eine scheinbare Nicht-Korrelation.

Beispiele:

a) Das Einkommen x eines Haushalts hat scheinbar keinen Einfluss auf den Milchkonsum y. Dies liegt jedoch an der Anzahl der Kinder: Mit steigendem Einkommen fällt die Anzahl der Kinder, während mit der Anzahl der Kinder der Milchkonsum steigt.

Abbildung 2-6: Beispiel für Suppressorvariable



[xy] = 0

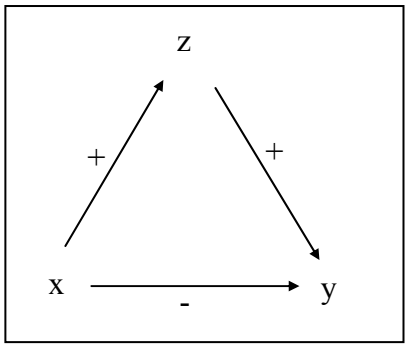
Wird die Anzahl der Kinder statistisch kontrolliert, so wird der positive Zusammenhang zwischen Einkommen und Milchkonsum sichtbar. Der positive direkte und der negative indirekte Effekt heben sich gegenseitig auf, sodass der Gesamteffekt 0 ist.
Der indirekte Effekt ist negativ: sign [xz] = - sign [yz]
Der bereinigte Effekt ist positiv: [xy : z] = [xy : ¬z] > 0

b) In der schon erwähnten Arbeit von Durkheim (1897) wurde eine starke Integration der jüdischen Bevölkerungsgruppe nachgewiesen, jedoch nur eine durchschnittliche Selbstmordrate (y). Dies liegt daran, dass Juden eher in Städten wohnen (z_1) oder eher intellektuelle Berufe ausüben (z_2), wodurch die Suizid-Gefährdung steigt.

4) **Distorter Variable** (Verzerrende Variable).
Es handelt sich hierbei nur um einen graduellen Unterschied zur Suppressorvariablen: Der Einfluss des Drittfaktors kann so stark sein, dass nach Einführen der Drittvariablen ein zu dem ursprünglichen entgegengesetzter Zusammenhang sichtbar wird.

Beispiel: Verheiratete (x) sind scheinbar stärker selbstmordgefährdet (y). Dies liegt jedoch an dem Einfluss des Alters (z): Die Variable z wirkt auf beide Variablen gleichzeitig. Ältere sind eher verheiratet und eher selbstmordgefährdet. Kontrolliert man das Alter, so wird der entgegengesetzte Zusammenhang zwischen x und y sichtbar: Verheiratete sind weniger selbstmordgefährdet (wie es nach der Integrationsthese zu erwarten war).

Abbildung 2-7: Beispiel für Distortervariable



[xy] > 0

Der Effekt von z ist also größer als der direkte Effekt von x auf y, sodass der Gesamteffekt von z dominiert wird.
Der indirekte Effekt ist positiv: sign [xz] = sign [yz]
Der bereinigte Effekt ist negativ: [xy : z] = [xy : ¬z] < 0

2.2.4.2 Zerlegungsformel am Beispiel eines Suppressor- sowie Distorter-Phönomens

Zerlegungsformel am Beispiel eines **Suppressor-Phänomens**

Rosenberg (1968) entnimmt einer Untersuchung von Arnold M. Rose folgendes Beispiel: „What do you think of having Jews on the union staff?“
Es stellt sich heraus, dass Jüngere (≤ 29 Jahre) mit wenig Gewerkschaftssozialisation (< 4 Jahre) dies zu 56,4 % eher neutral sehen, während die mittlere Altersgruppe (30 - 49 Jahre) mit wenig Gewerkschaftssozialisation dies nur zu 37,1 % und die höhere Altersgruppe (≥ 50 Jahre) mit wenig Gewerkschaftssozialisation dies nur zu 38,4 % neutral sehen. Um nur mit Dichotomien zu arbeiten, fasse ich die letzten beiden Gruppen zusammen, da sie sich ja auch ähnlich verhalten.

Tabelle 2-12:

		Dauer Gewerkschaftsmitglied (x)		
		< 4 Jahre	≥4 Jahre	
Jews on union staff (y)	Neutral	62	129	191
	Nicht neutral	64	127	191
		126	256	382

$$[yx] = 62 \cdot 127 - 64 \cdot 129 = -382$$

$$\delta_{yx} = [yx]/n = -1$$

$$s_{yx} = [xy]/n^2 = -0{,}003$$

$$\Phi_{yx} = [yx]/34304 = -0{,}011$$

Die Sozialisation in der Gewerkschaft (operationalisiert durch die Dauer der Gewerkschaftsmitgliedschaft) haben Jüngere systematisch weniger erfahren als Ältere, andererseits könnten Jüngere „toleranter“ (hier gemessen als Neutralität) sein. Deshalb wird Alter (z) als Testfaktor eingeführt, da eine Scheinkomponente aufgrund des Alters vermutet werden kann.

Tabelle 2-13:

Alter ($z = z_1$)
$\leq$ 29 Jahre

		in Gewerkschaft (x)		
		< 4 Jahre	≥ 4 Jahre	
Jews on union staff (y)	Neutral	44	32	76
	Nicht neutral	34	19	53
		78	51	129

$[yx : z_1] = -252$
$\delta_{yx:z_1} = [yx : z_1] / n_1 = -1{,}953$
$s_{yx:z_1} = [yx : z_1] / n_1^{\,2} = -0{,}015$
$\Phi_{yx:z_1} = [yx : z_1] / 4003 = -0{,}063$

Alter ($z = z_2$)
$\geq$ 30 Jahre

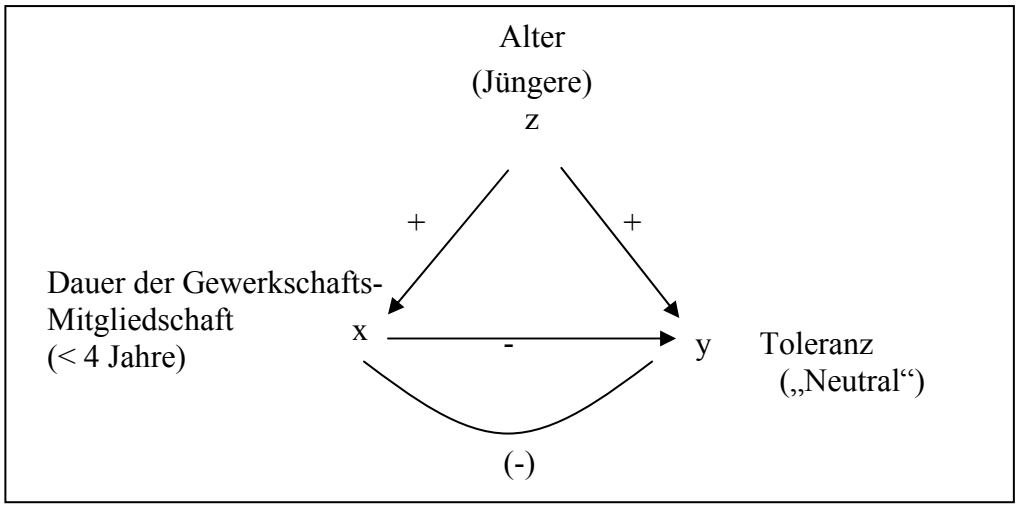
		in Gewerkschaft (x)		
		< 4 Jahre	≥ 4 Jahre	
Jews on union staff (y)	Neutral	18	97	115
	Nicht neutral	30	108	138
		48	205	253

$[yx : z_2] = -966$
$\delta_{yx:z_2} = [yx : z_2] / n_2 = -3{,}818$
$s_{yx:z_2} = [yx : z_2] / n_2^{\,2} = -0{,}015$
$\Phi_{yx:z_2} = [yx : z_2] / 12496 = -0{,}077$

In den Teilgruppen ergibt sich also eine sehr ähnliche Beziehung: $[xy : z_1] \approx [xy : z_2]$

Diese Ergebnisse sind nur möglich wegen der Beziehungen der Ausgangsvariablen x und y zu dem Drittfaktor z.

Abbildung 2-8: Beispiel für Suppressor-Phänomen



Bilanz:

gesamt	direkt	spurious
(-)	-	+

Tabelle 2-14:

		in Gewerkschaft (x)		
		< 4 Jahre	≥ 4 Jahre	
Alter (z)	≤ 29 Jahre	78	51	129
	≥ 30 Jahre	48	205	253
		126	256	382

$[xz] = 13542$
$\delta_{xz} = [xz]/n == 35{,}450$
$s_{xz} = [xz]/n^2 = 0{,}093$
$\Phi_{xz} = [xz]/32446 = 0{,}417$

Tabelle 2-15:

		Alter (z)		
		≤ 29 Jahre	≥ 30 Jahre	
Jews on union staff (y)	Neutral	76	115	191
	Nicht neutral	53	138	191
		129	253	382

$[yz] = 4393$
$\delta_{yz} = [yz]/n = 11{,}500$
$s_{yz} = [yz]/n^2 = 0{,}030$
$\Phi_{yz} = [yz]/34506$

Die Zerlegung für $\overline{\delta_{yx}}$ (Dies ist die klassische Version mit $\alpha = \beta = 1$.)

$$\delta_{yx} = \delta_{yx:z_1} + \delta_{yx:z_2} + \frac{n}{n_1 n_2}\delta_{xz}\delta_{yz}$$

$$-1 = -1{,}953 - 3{,}818 + \frac{382}{129 \cdot 253}35{,}450 \cdot 11{,}500$$

Die Zerlegung für [yx] (Dies ist die rechnerisch einfachste Version.)

$$[yx] = \frac{n}{n_1}[yx : z_1] + \frac{n}{n_2}[yx : z_2] + \frac{1}{n_1 n_2}[xz] \quad [yz]$$

$$-382 = \frac{382}{129}(-252) + \frac{382}{253}(-966) + \frac{1}{129 \cdot 253}13542 \cdot 4393$$

Die Zerlegung für die Kovarianz s_{yx} (Die Kovarianz hat hierbei die günstigsten Eigenschaften, s.u.)

$$s_{yx} = \frac{n_1}{n} s_{yx:z_1} + \frac{n_2}{n} s_{yx:z_2} + \frac{s_{xz} s_{yz}}{\frac{n_1}{n} \frac{n_2}{n}} = s_z^2$$

$$-0{,}003 = \frac{129}{382}(-0{,}015) + \frac{253}{382}(-0{,}015) + \frac{0{,}093 \cdot 0{,}030}{\frac{129}{382}\frac{253}{382}}$$

Die Zerlegung für Phi (Phi ist die wichtigste Maßzahl für die Vierfeldertafel.)

$$\Phi(x,y) = \frac{n}{n_{z_1}} \frac{\sqrt{n_{x_1:z_1} \cdot n_{x_2:z_1} \cdot n_{y_1:z_1} \cdot n_{y_2:z_1}}}{\sqrt{n_{x_1} \cdot n_{x_2} \cdot n_{y_1} \cdot n_{y_2}}} \Phi(xy:z_1)$$

$$+ \frac{n}{n_{z_2}} \frac{\sqrt{n_{x_1:z_2} \cdot n_{x_2:z_2} \cdot n_{y_1:z_2} \cdot n_{y_2:z_2}}}{\sqrt{n_{x_1} \cdot n_{x_2} \cdot n_{y_1} \cdot n_{y_2}}} \Phi(xy:z_2)$$

$$+ \Phi(xz)\Phi(yz)$$

$$-0{,}011 = \frac{382}{129} \frac{\sqrt{78 \cdot 51 \cdot 76 \cdot 53}}{\sqrt{126 \cdot 256 \cdot 191 \cdot 191}} (-0{,}063)$$

$$+ \frac{382}{253} \frac{\sqrt{48 \cdot 205 \cdot 115 \cdot 138}}{\sqrt{126 \cdot 256 \cdot 191 \cdot 191}} (-0{,}077)$$

$$+ 0{,}127 \cdot 0{,}417$$

Zerlegungsformel am Beispiel eines **Distorter-Phänomens**

Rosenberg (1968) illustriert das Distorter-Phänomen an folgendem Beispiel, das ich für die Illustration der Zerlegungsformel verwenden möchte.
Die Arbeiterschicht scheint eine stärkere Affinität zu den Bürgerrechten zu haben. Wenn man aber Ethnie kontrolliert, wird das Gegenteil sichtbar: Die Mittelschicht befürwortet die Bürgerrechte stärker, der falsche Eindruck kommt nur dadurch zustande, dass „Schwarze“ überproportional Arbeiter und gleichzeitig überproportional für die Bürgerrechte sind.

Tabelle 2-16:

		Social class (x)		
		Middle class	Working class	
Civil rights score (y)	High	44	54	98
	Low	76	66	142
		120	120	240

$[yx] = -1200$
$\delta_{yx} = [yx]/n = -5$
$s_{yx} = [yx]/n^2 = -0{,}021$
$\Phi_{yx} = [yx]/14156 = -0{,}085$

Tabelle 2-17:

Ethnie ($z = z_1$)
„Schwarze“

		Social class (x)		
		Middle class	Working class	
Civil rights score (y)	High	14	50	64
	Low	6	50	56
		20	100	120

$[yx : z_1] = 400$
$\delta_{yx:z_1} = [yx : z_1]/n_1 = 3{,}333$
$s_{yx:z_1} = [yx : z_1]/n_1^2 = 0{,}028$
$\Phi_{yx:z_1} = [yx : z_1]/2677 = 0{,}149$

Ethnie ($z = z_2$)
„Weiße“

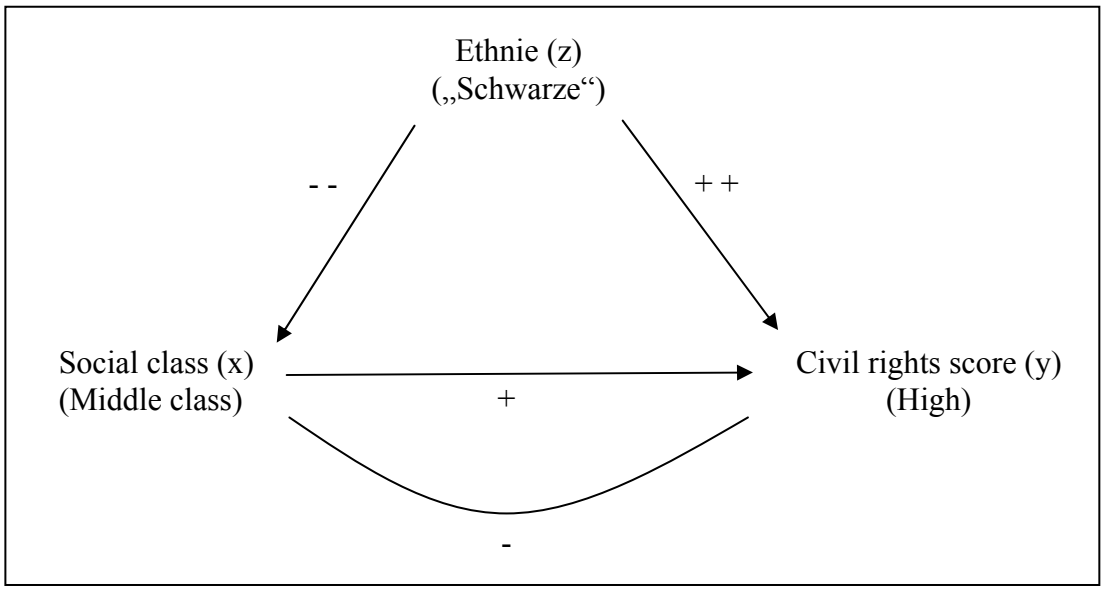
		Social class (x)		
		Middle class	Working class	
Civil rights score (y)	High	30	4	34
	Low	70	16	86
		100	20	120

$[yx : z_2] = 200$
$\delta_{yx:z_2} = [yx : z_2]/n_2 = 1{,}667$
$s_{yx:z_2} = [yx : z_2]/n_2^2 = 0{,}014$
$\Phi_{yx:z_2} = [yx : z_2]/2418 = 0{,}083$

In den beiden Teilgruppen ergibt sich bei etwas großzügiger Betrachtung eine ähnliche Beziehung: $[xy : z_1] \approx [xy : z_2]$

Es liegt also wieder an den Beziehungen der Ausgangsvariablen x und y zu dem Drittfaktor z.

Abbildung 2-9: Beispiel für Distorter-Phänomen



Bilanz:

gesamt	direkt	spurious
-	+	--

Tabelle 2-18:

		Social class (x)		
		Middle class	Working class	
Ethnie (z)	„Schwarze“	20	100	120
	„Weiße“	100	20	120
		120	120	240

$[xz] = -9600$

$\delta_{xz} = [xz]/n = -40$

$s_{xz} = [xz]/n^2 = -0{,}167$

$\Phi_{xz} = [xz]/14400 = -0{,}667$

Tabelle 2-19:

		Ethnie (z)		
		„Schwarze“	„Weiße “	
Civil rights score (y)	High	64	34	98
	Low	56	86	142
		120	120	240

$[yz] = 3600$

$\delta_{yz} = [yz]/n = 15$

$s_{yz} = [yz]/n^2 = 0{,}063$

$\Phi_{yz} = [yz]/14156 = 0{,}254$

Die Zerlegung für δ_{yx}

$$\delta_{yx} = \delta_{yxz_1} + \delta_{yxz_2} + \frac{n}{n_1 n_2}\delta_{xz}\delta_{yz}$$

$$-5 = 3{,}333 + 1{,}667 + \frac{240}{120 \cdot 120} \quad (-40)\cdot 15$$

Die Zerlegung für [yx]

$$[yx] = \frac{n}{n_1}[yx : z_1] + \frac{n}{n_2}[yx : z] + \frac{1}{n_1 n_2}[\mathrm{xz}]\,[yz]$$

$$-1200 = \frac{240}{120}400 + \frac{240}{120}200 + \frac{1}{120 \cdot 120}(-9600)\cdot 3600$$

Die Zerlegung für die Kovarianz s_{xy}

$$s_{yx} = \frac{n_1}{n}s_{yx:z_1} + \frac{n_2}{n}s_{yx:z_2} + \frac{s_{xz}s_{yz}}{\frac{n_1}{n}\frac{n_2}{n}}$$

$$-0{,}021 = \frac{120}{240}0{,}028 + \frac{120}{240}0{,}014 + \frac{(-0{,}167)\cdot 0{,}063}{\frac{120}{240}\frac{120}{240}}$$

Die Zerlegung für Phi

$$\Phi(x,y) = \frac{n}{n_{z_1}}\frac{\sqrt{n_{x_1:z_1}\cdot n_{x_2:z_1}\cdot n_{y_1:z_1}\cdot n_{y_2:z_1}}}{\sqrt{n_{x_1}\cdot n_{x_2}\cdot n_{y_1}\cdot n_{y_2}}}\Phi(xy : z_1)$$

$$+\frac{n}{n_{z_1}}\frac{\sqrt{n_{x_1:z_2}\cdot n_{x_2:z_2}\cdot n_{y_1:z_2}\cdot n_{y_2:z_2}}}{\sqrt{n_{x_1}\cdot n_{x_2}\cdot n_{y_1}\cdot n_{y_2}}}\Phi(xy : z_2)$$

$$+\Phi(xz)\Phi(yz)$$

$$-0{,}085 = \frac{240}{120}\frac{\sqrt{20\cdot 100\cdot 64\cdot 56}}{\sqrt{120\cdot 120\cdot 98\cdot 142}}0{,}149$$

$$+\frac{240}{120}\frac{\sqrt{100\cdot 20\cdot 34\cdot 86}}{120\cdot 120\cdot 98\cdot 142}0{,}083$$

$$+0{,}254\cdot(-0{,}667)$$

2.2.4.3 Vorzeichenregel nach Davis für Suppressor- und Distorter Phänomene

Suppressor Variable („Dämpfend“)

Bei Einführung von z wird die Beziehung zwischen x und y größer
a) Falls das Einführen von z eine positive Korrelation verstärkt, so muss gelten: sign [xz] = -sign [yz]
b) Falls das Einführen von z eine negative Korrelation verstärkt, so muss gelten: sign [xz] = sign [yz]

Distorter Variable („Verzerrend“)

Bei Einführung von z wird ein zu dem ursprünglichen Zusammenhang [xy] entgegengesetztes Vorzeichen in $[xy:z_1]$ und $[xy:z_2]$ sichtbar:
a) Falls [xy] > 0:
sign [xz] = sign [yz]
b) Falls [xy] < 0:
sign [xz] = -sign [yz]

Mit „Bilanzen“ formuliert:

Tabelle 2-20:

	Gesamt-zusammenhang	Direkter Kausaleffekt	Indirekter Kausaleffekt oder Spurious
Suppressor Variable	$+_{(0)}$ $-_{(0)}$	$++_{(+)}$ $--_{(-)}$	$-_{(-)}$ $+_{(+)}$
„Scheinbare Nicht-Kausalität“	0 0	+ -	- +
Distorter Variable	+ -	- +	++ - -

Vorzeichenregel und/oder perfekte Maßzahl?

$$[xy] = \alpha[xy : z_1] + \beta[xy : z_2] + \gamma[xz][yz]$$

Bei Typen mit: $[xy : z_1] = [xy : z_2]$

$$[xy] = a\,[xy : z] + b\,[xz]\,[yz]$$
ges ber res

Man vergleicht den Gesamtzusammenhang und den bereinigten Zusammenhang:
Wenn a = 1, dann: res = ges – ber

Ob die Vorzeichenregel perfekt gilt, hängt von den Koeffizienten a und b ab, welche wiederum spezifisch sind für die gewählte Maßzahl bzw. „Quasi-Maßzahl“ [xy]. Falls a = 1, so gilt perfekte Vorzeichenregel. (b ist immer positiv und stört deshalb nicht, denn:
(res > 0) $\Leftrightarrow$ ([xz] [yz] > 0))

Γ

Z.B. für δ (δ ist „Quasi-Maßzahl“, denn δ ist nicht normiert.)

δ_{xy} =	$\delta_{xy:z_1} + \delta_{xy:z_2}$ +	$\gamma \cdot \delta_{xz}\delta_{yz}$
ges	direkter Kausaleffekt bzw. bereinigter Zusammenhang	res (Residuum)
(res > 0)	genau dann	(Beziehungen von x und y zu dem
(<)	wenn	Testfaktor z sind gleichlautend (gegenlautend) im Vorzeichen)

D.h.: Die Vorzeichenregel von Davis ist implizit in meinen Bilanzgleichungen enthalten.

⌊

Relativierung („Dilemma“ gemäß Davis): Die Vorzeichenregel für δ ist perfekt, für Φ eine Daumenregel, weil α und β i.a. bei Φ nicht einfach gleich 1 sind.
Andererseits ist Φ eine perfekte Maßzahl, während δ nicht normiert ist:

z.B.: $\begin{pmatrix} 10 & 0 \\ 0 & 10 \end{pmatrix}$ $\delta = \frac{10 \cdot 10}{10 + 10} = 5$

Die Vorzeichenregel auf Basis der Kovarianz

Eine Kompromisslösung in diesem „Dilemma“ ergibt sich nach meiner Auffassung für die Kovarianz s_{xy}:

s_{xy} ist im allgemeinen Fall nicht normiert, d.h. i.a. gilt nicht: $|s_{xy}| \leq 1$

Aber: Für den einfachen Fall der Vierfeldertafel ist die Kovarianz normiert.

$$s_{xy} = \frac{ad - bc}{n^2}$$

$$-1 \leq \frac{ad - bc}{n^2} \leq +1$$

s_{xy} ist also normiert, jedoch nicht so perfekt wie Φ_{xy}, die Maximalwerte $\pm$ 1 können in vielen Datenkonstellationen gar nicht erreicht werden. Dennoch ist die Kovarianz normiert und lässt sich auf eine Weise zerlegen, die ähnlich einfach ist wie bei δ, wobei δ andererseits nicht normiert ist.

Zerlegungsformel für die Kovarianz

$s_{xy} = s_{xy \cdot z} + s_{xz}\, s_{yz}$, wobei: $s_z^2 = \frac{n_1}{n}\frac{n_2}{n}$

Hierbei ist die partielle Kovarianz $s_{xy.z}$ gleich einem gewogenen arithmischen Mittel aus den beiden bedingten Kovarianzen für die beiden Teiltabellen:

$$s_{xy.z} = \frac{n_1}{n} s_{xy:z_1} + \frac{n_2}{n} s_{xy:z_2}$$

Aus dem Vergleich der Kovarianz mit der partiellen Kovarianz ergibt sich unmittelbar die Vorzeichenregel:

$\text{sign}\,(s_{xy} - s_{xy.z}) = \text{sign}\,(s_{xz} \cdot s_{yz})$

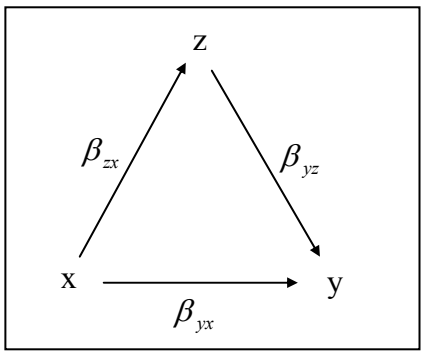
In diesem Sinne ist die **Kovarianz** nach meiner Auffassung ein Kompromiss in dem „Dilemma“ (Davis) zwischen Vorzeichenregel und perfekter Maßzahl.

Gesamtzusammenhang > bereinigter Zusammenhang/direkter Effekt? (<)

(Erster Mechanismus, der wirkt: Direkter Kausaleffekt)

1. Fall: Der zweite Mechanismus ist indirekt kausal.

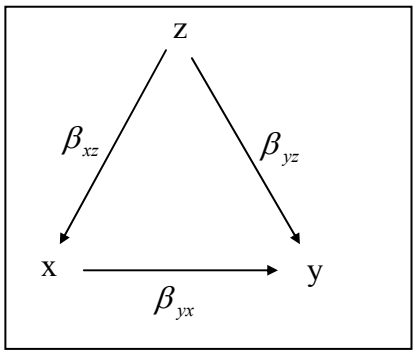
Abbildung 2-10: Indirekter Kausaleffekt



$$r_{xy} = \underbrace{\beta_{xy}}_{\text{direkter Kausaleffekt}} + \underbrace{\beta_{yz}\beta_{zx}}_{\text{indirekter Kausaleffekt}}$$

2. Fall: Der zweite Mechanismus beinhaltet eine Scheinkomponente.

Abbildung 2-11: Scheinkomponente



$r_{xy} = \underbrace{\beta_{yx}}_{\text{direkter Kausaleffekt}} + \underbrace{r_{yz}\beta_{xz}}_{\text{Scheinkomponente auf Grund von z}}$

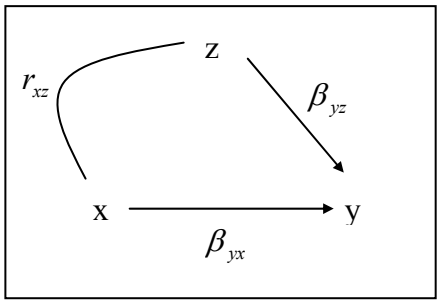
Vorzeichenregel für Fall 1 und 2:

Gesamtzusammenhang		Bereinigter Zusammenhang		Residuum
ges	=	ber	+	res

(res > 0) ⇔ (<) *(Die Effekte von x und y in Relation zu dem Drittfaktor z sind gleichlautend (gegenlautend) im Vorzeichen.)*

3. Fall: Der zweite Mechanismus ist ein korrelierter Effekt.

Abbildung 2-12: Korrelierter Effekt



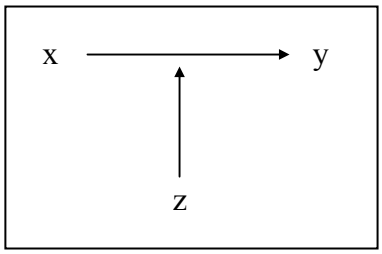
$r_{xy} = \underbrace{\beta_{yx}}_{\text{direkter Kausaleffekt}} + \underbrace{r_{xz}\beta_{yz}}_{\text{korrelierter Effekt}}$

Eine Vorzeichenregel ist nicht so sinnvoll, da es sich hier um unterschiedliche Konzepte handelt.

2.2.4.4 Typen mit: [xy : z] ≠ [xy ¬z] (Spezifikation)

Hier liegt eine **Spezifikation** (spezifizierende Variable, conditional relationship, qualifier variable oder Verfeinerung) vor. Die häufigste Beziehung zwischen drei Variablen besteht darin, dass nach der Einführung eines Testfaktors die Zusammenhänge in den Teiltabellen weder gleich dem ursprünglichen Zusammenhang sind (Bestätigung), noch verschwinden, sondern dass sich die Zusammenhänge in den beiden Teiltabellen unterscheiden. Die Beziehung zwischen x und y hängt also von der Ausprägung von z ab, wird durch z modifiziert.

Abbildung 2-13: Spezifikation



Statistische Beziehung:
[xy : z] ≠ [xy : ¬z]

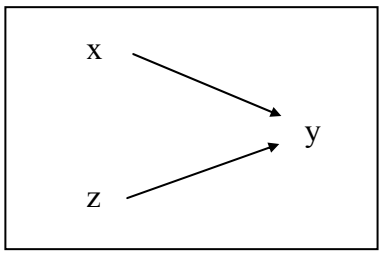
Auch die scheinbare Nicht-Kausalität kann durch die bedingten Beziehungen erzeugt werden. Die Zufriedenheit (x) mit dem Beruf korreliert nicht mit der Teilhabe (y) an Aktivitäten in der Wohngemeinde. Wird jedoch die Schicht (z) als Kontrollvariable eingeführt, so zeigt sich, dass bei den „white collar" – Berufen Unzufriedenheit mit Aktivität korreliert, während in der „working class" Zufriedenheit mit Aktivität korreliert. Diese beiden bedingten Zusammenhänge heben sich also insgesamt auf, sodass der Gesamteffekt 0 ist.

2.2.4.5 Conjoint influence

Hier wird nach Rosenberg (1962) das Merkmal z als eine weitere unabhängige Variable betrachtet.

a) Unabhängiger Effekt (vgl. Abbildung 2-14): Die beiden Einflussfaktoren x und z haben jeweils einen Effekt auf y, unabhängig davon, wie groß der Effekt des jeweils anderen Einflussfaktors ist.

Abbildung 2-14: Unabhängiger Effekt



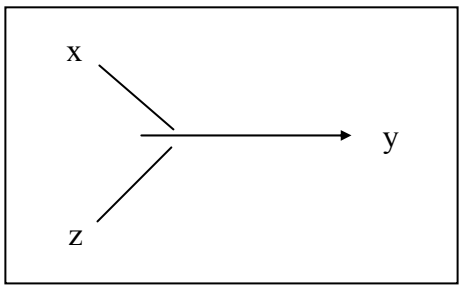
a1) Wenn sich die beiden unabhängigen Variablen (x und z) stark überschneiden, stellt sich die Frage, ob sie die abhängige Variable noch beeinflussen, wenn die jeweils andere unabhängige Variable kontrolliert wird. Ist dies der Fall, so handelt es sich um einen unabhängigen Effekt.
Statistische Beziehungen: [xy : z] ≠ 0,
[zy : x] ≠ 0

a2) Relativer Effekt: Falls zwei voneinander unabhängigen Erklärungsfaktoren (x und z) vorliegen, stellt sich die Frage, welcher Faktor den größeren Einfluss auf die abhängige Variable (y) hat.
[xy : z] größer oder kleiner als [zy : x]?

a3) Ein kumulativer Effekt tritt auf, falls der kombinierte Effekt der beiden unabhängigen Variablen größer ist als jeder der beiden Einzeleffekte.
Bei Überschneidungen ist es möglich, dass der kumulative Effekt kaum größer ist als die einzelnen Effekte.

b) Typologische Effekte: Ein Effekt in der abhängigen Variablen y tritt erst auf, wenn beide unabhängigen Variablen gemeinsam wirken. (Dies ist eine Interaktion, im Unterschied zu der Betrachtung additiver Effekte in (a1), (a2), (a3).)

Abbildung 2-15: Interaktionseffekt



Beispiel:
Nach Mayntz et al. (1971: 193) erzeugt ein Misserfolg (x) nur dann Unzufriedenheit (y), falls das Anspruchsniveau (z) hoch ist.

Diese Interaktion von Merkmalen entspricht „multiplikativem Wirken“ von Merkmalen.

$1_x :=$ { 1: Merkmal x liegt vor; 0: Merkmal x liegt nicht vor

1_z analog

$1_x \cdot 1z =$ { 1, falls: Merkmal x und Merkmal z liegen vor; 0, sonst

In dem Beispiel von Mayntz et al. gibt es nur einen Interaktionseffekt, keine „direkten Effekte“.

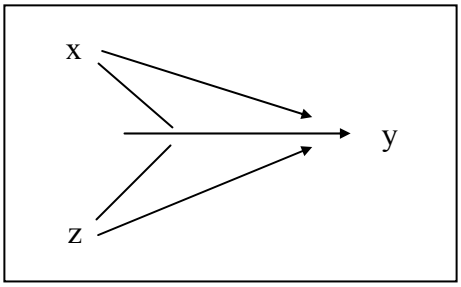
Ein Modell, das gleichzeitig „Haupt-Effekte“ (von x und z) sowie einen „Interaktionseffekt“ (von x · z) berücksichtigt, würde z.B. lauten (vgl. die Regressionsanalyse, Kap. 3.1):

$$\hat{y} = \alpha\, x + \beta\, z + \gamma\, x \cdot z$$

wobei α, β und γ Parameter für die Gewichtung sind.

Graphisch könnte man dies wie folgt veranschaulichen:

Abbildung 2-16: „Haupt-Effekte“ und Interaktionseffekt



2.2.4.6 Verschiedene typologische Effekte und Interaktionseffekte

Rosenberg (1968) unterscheidet vier typologische Effekte:

1. Distinctive type
   Beispiel: Eine spezielle Kombination – z.B. ein jüngerer Sohn mit überwiegend älteren Schwestern – weist die höchste Selbstachtung auf.

2. Modifying type
   Beispiel: In den USA haben Italo-Amerikaner, Katholiken, Schwarze und Juden eher eine Parteiaffinität zu den Demokraten, weiße Protestanten dagegen eher eine Parteiaffinität zu den Republikanern.
   Die Identifikation mit einer Ethnie hat allein keinen Effekt auf die Parteiaffinität, aber sie verstärkt den Effekt der Ethnien auf die Parteiaffinität.

3. The consistent-inconsistent type
   Beispiel: Statusinkonsistente weisen eher Stresssymptome auf als Statuskonsistente.

4. The relative type
   Beispiel (gemäß Lenski):
   Das Erziehungsziel Autonomie (versus Gehorsam) wird am stärksten von Mittelschichtangehörigen vertreten, die selbst aus der Arbeiterschicht aufgestiegen sind, und am wenigsten von Arbeiterschichtangehörigen, die selbst aus der Mittelschicht abgestiegen sind.

Tabelle 2-21: Prozentsatz derer, die das Erziehungsziel Autonomie (versus Gehorsam) befürworten

		Schicht des Sohnes	
		Mittelschicht	Arbeiterschicht
Schicht des Vaters	Mittelschicht	74 %	48 %
	Arbeiterschicht	77 %	55 %

Der Informationsgehalt dieser Typen besteht für Rosenberg gerade in den speziellen Kombinationen der Ausgangsvariablen, sodass Rosenberg die „Haupteffekte“ der eigenen Schicht und der Herkunftsschicht gar nicht berichtet, sondern nur die „Kombinationseffekte“.

Bei den Söhnen, die selbst in der Mittelschicht sind, zeigt sich, dass Aufsteiger in die Mittelschicht häufiger für Autonomie sind als dies bei Mittelschicht-„Stayer“ der Fall ist. Bei den Söhnen, die selbst in der Arbeiterschicht sind, zeigt sich, dass die Abgestiegenen den Wert Autonomie seltener aufweisen als die Arbeitsschicht-„Stayer“.

Insgesamt scheint also der Wert Autonomie am wenigsten charakteristisch für Absteiger und am stärksten charakteristisch für Aufsteiger.

2.2.4.6.1 Varianzanalytische Interpretation von Rosenbergs Mobilitäts-Beispiel

Zur Erinnerung an die Deskriptivstatistik:
Einfache Varianzanalyse

Abhängigkeit einer metrischen Variablen (y) von einer nominalen unabhängigen Variablen (x).

Beispiel: Einkommen erklären durch Stellung im Beruf.

Tabelle 2-22: Effekte in der Varianzanalyse

Allgemeiner Durchschnitt als Bezugspunkt: $\bar{y} = 1964$

	Schätzung mit Vorinformation über die Stellung im Beruf	Effekt der Stellung im Beruf auf das Einkommen
j = 1: Selbstständige	$\bar{y}_1 = 3225$	$\bar{y}_1 - \bar{y} = 1261$
j = 2: Beamte	$\bar{y}_2 = 2502$	$\bar{y}_2 - \bar{y} = 538$
j = 3: Angestellte	$\bar{y}_3 = 1846$	$\bar{y}_3 - \bar{y} = -118$
j = 4: Arbeiter	$\bar{y}_4 = 1718$	$\bar{y}_4 - \bar{y} = -246$

$\bar{y}_i$ = Mittelwert der i-ten Gruppe

Varianzanalyse: Haupteffekte und Interaktionseffekte

Tabelle 2-23: Einfache Varianzanalyse des Einkommens (y) durch die Stellung im Beruf (x)

	Prognose aufgrund des additiven Modells	Tatsächlicher Durchschnitt	Abweichung von Prognose
j = 1: Selbstständige	1964 + 1261	3225	0
j = 2: Beamte	1964 + 538	2502	0
j = 3: Angestellte	1964 - 118	1846	0
j = 4: Arbeiter	1964 - 246	1718	0

Wenn man nun als zweite unabhängige Variable Geschlecht betrachtet, so gibt es entsprechende Effekte des Geschlechts.

Die Effekte der Stellung im Beruf und des Geschlechts jeweils allein heißen **Haupteffekte**.

Die Prognose des additiven Modells der zweifachen Varianzanalyse mit Stellung im Beruf (A) und Geschlecht (B) lautet:

$$\bar{y} + \left(\bar{y}_{A_i} - \bar{y}\right) + \left(\bar{y}_{B_j} - \bar{y}\right)$$

Interaktionseffekte bzw. nicht-additive Effekte gibt es in dem Ausmaß, wie der tatsächliche Durchschnitt $\bar{y}_{ij}$ in den Kombinationen A_i, B_j von dieser Prognose auf Basis des additiven Modells der zweifachen Varianzanalyse abweicht.

Interaktionseffekte in der zweifachen Varianzanalyse in Rosenbergs Beispiel

Um Rosenbergs Beispiel genauer analysieren zu können, habe ich die absoluten Fallzahlen der Typen, die bei Rosenberg fehlen, aus dem Original von Lenski rekonstruiert (Gerhard Lenski: The religious factor. New York 1961. 1977[2]).

Befragt wurden 469 Personen, von denen
84 den Mobilitätsstatus Middle: non-mobile haben (davon sind 74 % pro Autonomie),
118 Middle: upwardly mobile (davon sind 77 % pro Autonomie),
40 Working: downwardly mobile (davon sind 48 % pro Autonomie) und
227 Working: non-mobile (davon sind 55 % pro Autonomie).

Das angemessene Analysemodell scheint mir zunächst die Varianzanalyse der abhängigen Variablen Autonomie vs. Gehorsam mit Hilfe der zwei Erklärungsfaktoren „Schicht des Sohnes“ und „Schicht des Vaters“ (deshalb: zweifache Varianzanalyse).

Die eigene soziale Lage ist bzgl. der Einstellung zur Autonomie deutlich trennschärfer als die Herkunftslage ($\left|\bar{y}_{z_1} - \bar{y}_{z_2}\right| = 21{,}8$ vs. $\left|\bar{y}_{x_1} - \bar{y}_{x_2}\right| = 2{,}7$). Wenn man von der eigenen Lage ausgeht, so geht die Differenzierung nach der Herkunftsschicht jeweils in die Richtung, dass Personen mit Arbeiterherkunft den Wert Autonomie etwas stärker aufweisen (77,1 % vs. 73,8 % (Differenz 3,3) bzw. 55,1 % vs. 47,5 % (Differenz 7,6)).

Dies ist erstaunlich, da bei Arbeiterherkunft insgesamt der Wert eher weniger anzutreffen ist (nämlich nur 62,6 % vs. 65,3 % bei Mittelschicht-Herkunft, Differenz = -2,7).

Da es sich bei der Randverteilung um gewogene arithmetische Mittel handelt, liegt dies daran, dass die etwas geringere Ausprägung pro Autonomie bei Mittelschicht-„Stayer“ mit 73,8 % vs. der Aufsteiger mit 77,1 % dennoch stärker zu Buche schlägt, weil die „soziale Vererbung“ bei der Mittelschicht mit 67,7 % stark ausgeprägt ist (bei der Arbeiterschicht ähnlich mit 65,8 %, aber die Arbeiterschicht-„Stayer“ sind mit 55,1 % deutlich weniger pro Autonomie).

Tabelle 2-24: Anteil der Personen, die persönliche Autonomie hoch bewerten, in Abhängigkeit von der eigenen bzw. der Herkunftsschicht bzw. vom Mobilitätsstatus (Aufstieg, Abstieg etc.) $n_{y_2}/(n_{y_1}+n_{y_2}) = f(x,z)$

		Schicht des Sohnes (z): Mittelschicht	Schicht des Sohnes (z): Arbeiterschicht	
Schicht des Vaters (x)	Mittelschicht	$\frac{62}{84} = 73{,}8\%$ Prognose: 63,3+12,4+2,0=77,7 Abweichung von Prognose: 73,8-77,7=-3,9	$\frac{19}{40} = 47{,}5\%$ Prognose: 63,3-9,4+2,0=55,9 Abweichung von Prognose: 47,5-55,9=-8,4	$\frac{81}{124} = 65{,}3\%$ $\bar{y}_{x_1} - \bar{y} =$ 65,3 – 63,3 = +2,0
	Arbeiterschicht	$\frac{91}{118} = 77{,}1\%$ Prognose: 63,3 + 12,4-0,7 = 75,0 Abweichung von Prognose: 77,1-75,0=+2,1	$\frac{125}{227} = 55{,}1\%$ Prognose: 63,3-9,4-0,7 = 53,2 Abweichung von Prognose: 55,1-53,2=+1,9	$\frac{216}{345} = 62{,}6\%$ $\bar{y}_{x_2} - \bar{y} =$ 62,6-63,3=-0,7
		$\frac{153}{202} = 75{,}7\%$ $\bar{y}_{z_1} - \bar{y} = 75{,}7 - 63{,}3 = +12{,}4$	$\frac{144}{267} = 53{,}9\%$ $\bar{y}_{z_2} - \bar{y} = 53{,}9 - 63{,}3 = -9{,}4$	$\frac{297}{469} = 63{,}3\%$

Insgesamt:

$$\frac{84}{124}\cdot\frac{62}{84}\cdot\frac{40}{124}\cdot\frac{19}{40}=\frac{81}{124}$$

$$\begin{array}{ccccc} & (50,0) & & (15,3) & \\ 67{,}7 & \cdot & 73{,}8 & + & 32{,}3 \cdot 47{,}5 = 65{,}3 \\ \wedge & & \wedge & & \vee \\ 34{,}2 & \cdot & 77{,}1 & + & 65{,}8 \cdot 55{,}1 = 62{,}6 \\ & (26{,}4) & & (36{,}2) & \end{array}$$

$$\frac{118}{345}\cdot\frac{91}{118}+\frac{227}{345}\cdot\frac{125}{227}=\frac{216}{345}$$

Das Beispiel ist also viel komplexer, als es bei Lenski und Rosenberg dargestellt wird. Der angemessene Bezugspunkt scheint mir zunächst das additive Modell der zweifachen Varianzanalyse, das hier zur Prognose verwendet wird: $\hat{y}_{ij} = \bar{y} + (\bar{y}_{i+} - \bar{y}) + (\bar{y}_{+j} - \bar{y})$

Die vier Interaktionseffekte bestehen dann in den vier Abweichungen der tatsächlichen Durchschnitte ($\bar{y}_{ij}$) in den Kombinationen von den beim additiven Modell zu erwartenden Werten $(\hat{y}_{ij})$: $\quad \bar{y}_{ij} - \hat{y}_{ij}$

Gegenüber dem allgemeinen Durchschnitt und den Haupteffekten der Herkunftslage und der eigenen Lage weisen Aufsteiger eine um 2,1 % stärker ausgeprägte Präferenz für Autonomie (selbstständiges Denken) auf. Absteiger dagegen weisen diese Präferenz unterproportional auf (-8,4 %). (Mittelschicht-„Stayer“ bleiben mit –3,9 % unterproportional und Arbeiter-„Stayer“ mit +1,9 % überproportional, was noch zu erklären wäre.)

Vieles spricht dafür, dass Autonomie, d.h. der Wert, selbstständig zu denken, eher eine Ursache des Aufstiegs ist als eine Folge. Dies sah auch Lenski so, da eine Überanpassung von Aufsteigern noch plausibel sein könne, nicht aber eine Überanpassung von Absteigern. Deshalb soll diese Kausalrichtung nach einer kurzen Diskussion der „Effekt-Codierung“ weiter verfolgt werden.

Nicht gewogene Daten bzw. **„Effekt-Codierung“** führen zu falschen Schlüssen.

Wenn man die Lenski-Daten ohne die Fallzahlen analysieren würde, die bei Rosenberg fehlen, wäre man zur Analyse mit „nicht gewogenen“ Daten gezwungen.
An diesem Beispiel soll gezeigt werden, dass dies zu – im Sinne Lenskis – eher unsinnigen Ergebnissen führt, dann wären nämlich Aufsteiger und Absteiger ähnlich in ihrer Einstellung, während Lenski das selbstständige Denken (Autonomie) gerade für ein Charakteristikum von Aufsteigern hält.

Unter „Effekt-Codierung“ versteht man eine Codierung T_i , bei der die Gewichte b_i gerade die Effekte bzgl. des nicht gewogenen Mittelwerts der Mittelwerte sind:

$\hat{y} = b_0 + b_1 T_1 + \ldots + b_{k-1} T_{k-1}$

$b_0 = \sum_{j=1}^{k} \bar{y}_j / k =: \bar{\bar{y}}$

$b_i = \bar{y}_i - \bar{\bar{y}}$ (= „Effekt“ der i-ten Ausprägung)

Die bessere Art, Effekte zu messen, läge in der Wahl des tatsächlichen (= gewogenen) Mittelwerts $\bar{y}$ *als Bezugspunkt:* $\bar{y}_i - \bar{y}$

Tabelle 2-25: Nicht gewogene Daten

		Schicht des Sohnes (z)		
		Mittelschicht	Arbeiterschicht	
Schicht des Vaters (x)	Mittel-schicht	74 Prognose: 63,5 + 12 - 2,5 =73 Abweichung von Prognose: +1	48 Prognose: 63,5 – 12 – 2,5 = 49 Abweichung von Prognose: -1	61 61 – 63,5 = -2,5
	Arbeiter-schicht	77 Prognose: 63,5 + 12 + 2,5 =78 Abweichung von Prognose: -1	55 Prognose: 63,5 – 12 + 2,5 = 54 Abweichung von Prognose: +1	66 66 – 63,5 = +2,5
		75,5 75,5 - 63,5 = +12	51,5 51,5 – 63,5 = -12	63,5

Der Effekt von Aufstieg und Abstieg wäre jeweils „-1 %“ weg vom Wert Autonomie, während Lenski und Rosenberg gerade den Unterschied von Aufstieg und Abstieg herausstellen wollten.

Hypothese: Der Wert Autonomie strukturiert das Mobilitätsverhalten

Zu erklären seien die Chancen, zu einem der vier Mobilitätstypen zu gehören.

Tabelle 2-26:

		Schicht des Sohnes
		Mittelschicht	Arbeiterschicht
Schicht des Vaters	Mittelschicht	$\frac{84}{469} = 17{,}9\%$	$\frac{40}{469} = 8{,}5\%$
	Arbeiterschicht	$\frac{118}{469} = 25{,}2\%$	$\frac{227}{469} = 48{,}4\%$

Relative Häufigkeit der Mobilitätstypen als Bezugspunkt

Tabelle 2-27: Effekte, wenn man Autonomie als Wert präferiert.

		Schicht des Sohnes
		Mittelschicht	Arbeiterschicht
Schicht des Vaters	Mittelschicht	$\frac{62}{297} = 20{,}9\%$ 20,9 % - 17,9 % = +3,0 %	$\frac{19}{297} = 6{,}4\%$ 6,4 % - 8,5 % = -2,1 %
	Arbeiterschicht	$\frac{91}{297} = 30{,}6\%$ 30,6 % - 25,2 % = +5,4 %	$\frac{125}{297} = 42{,}1\%$ 42,1 % - 48,4 % = -6,3 %

Der Wert Autonomie begünstigt also das Auftreten des Typs „Aufstieg“ derart, dass man ihn zu +5,4 % überproportional vorfindet. Entsprechend findet man Abstieg mit –2,1 % unterproportional.

Die Effekte sind noch deutlicher, wenn man den Zusammenhang aus der Perspektive betrachtet, dass man den Wert Autonomie nicht präferiert.

Tabelle 2-28: Effekte, wenn man Autonomie als Wert nicht präferiert

Schicht des Vaters	Schicht des Sohnes: Mittelschicht	Schicht des Sohnes: Arbeiterschicht
Mittelschicht	$\frac{22}{172} = 12{,}8\%$ 12,8 % - 17,9 % = -5,1 %	$\frac{21}{172} = 12{,}2\%$ 12,2 % - 8,5 % = +3,7 %
Arbeiterschicht	$\frac{27}{172} = 15{,}7\%$ 15,7 % - 25,2 % = -9,5 %	$\frac{102}{172} = 59{,}3\%$ 59,3 % - 48,4 % = +10,9 %

Wenn man Autonomie nicht als Wert präferiert, ist man beim Typ Aufstieg mit –9,5 % stark unterproportional vertreten und beim Typ Abstieg mit +3,7 % überproportional. Besonders groß ist der Effekt auf den Typ Arbeiter-„Stayer“.

Berücksichtigung der „Randeffekte“, dass man von der Spitze nicht aufsteigen kann und von unten nicht absteigen kann.

In dem vorliegenden Beispiel gibt es nur zwei soziale Lagen in der Hierarchie, sodass sich die „Randeffekte“ hier darin zeigen, dass man bei Mittelschicht-Herkunft nicht aufsteigen und bei Arbeiterschicht-Herkunft nicht absteigen kann.
Um dies angemessen zu berücksichtigen, könnte man als Bezugspunkte jeweils die Mittelschicht-Herkunft bzw. die Arbeiterschicht-Herkunft gesondert betrachten.

Personen mit Mittelschicht-Herkunft

Schicht des Vaters: Mittelschicht	Schicht des Sohnes: Mittelschicht	Schicht des Sohnes: Arbeiterschicht
	$\frac{84}{124} = 67{,}7\%$	$\frac{40}{124} = 32{,}3\%$

Personen mit Präferenz für den Wert Autonomie

	Schicht des Sohnes: Mittelschicht	Schicht des Sohnes: Arbeiterschicht
Schicht des Vaters: Mittelschicht	$\frac{62}{81} = 76{,}5\%$ 76,5 % - 67,7 % = +8,8 %	$\frac{19}{81} = 23{,}5\%$ 23,5 % - 32,3 % = -8,8 %

Der Wert Autonomie fördert das Verbleiben in der Mittelschicht mit +8,8 %, andererseits sind diese Personen bei den Absteigern mit den spiegelbildlichen –8,8 % unterproportional vertreten.

Personen mit Arbeiterschicht-Herkunft

	Schicht des Sohnes: Mittelschicht	Schicht des Sohnes: Arbeiterschicht
Schicht des Vaters: Arbeiterschicht	$\frac{118}{345} = 34{,}2\%$	$\frac{227}{345} = 65{,}8\%$

Personen mit Präferenz für den Wert Autonomie

	Schicht des Sohnes: Mittelschicht	Schicht des Sohnes: Arbeiterschicht
Schicht des Vaters: Arbeiterschicht	$\frac{91}{216} = 42{,}1\%$ 42,1 % - 34,2 % = 7,9 %	$\frac{125}{216} = 57{,}9\%$ 57,9 % - 65,8 % = -7,9 %

Bei Arbeiterschicht-Herkunft steigt man bei der Präferenz für Autonomie zu +7,9 % überproportional auf und verbleibt spiegelbildlich zu –7,9 % unterproportional in der Arbeiterschicht.

Effekte, wenn man Autonomie als Wert *nicht* präferiert.

Personen mit Mittelschicht-Herkunft

	Schicht des Sohnes: Mittelschicht	Schicht des Sohnes: Arbeiterschicht
Schicht des Vaters: Mittelschicht	$\frac{22}{43} = 51{,}2\%$ 51,2 % - 67,7 % = -16,5 %	$\frac{21}{43} = 48{,}8\%$ 48,8 % - 32,3 % = +16,5 %

Bei Mittelschicht-Herkunft verbleibt man ohne den Wert Autonomie mit –16,5 % deutlich unterproportional in der Mittelschicht und steigt spiegelbildlich mit +16,5 % deutlich überproportional ab.

Personen mit Arbeiterschicht-Herkunft

	Schicht des Sohnes: Mittelschicht	Schicht des Sohnes: Arbeiterschicht
Schicht des Vaters: Arbeiterschicht	$\frac{27}{129} = 20{,}9\%$ 20,9 % - 34,2 % = -13,3 %	$\frac{102}{129} = 79{,}1\%$ 79,1 % - 65,8 % = +13,3 %

Ohne den Wert Autonomie steigt man bei Arbeiterschicht-Herkunft mit –13,3 % unterproportional häufig auf und verbleibt spiegelbildlich mit +13,3 % überproportional häufig in der Arbeiterschicht.

2.2.4.7 Interaktion, Spezifikation und typologische Effekte aus Sicht der Varianzanalyse

Die fruchtbarste Art, Interaktionseffekte in der Tabellenanalyse für 3 Variablen zu behandeln, wenn es eine zu erklärende Variable y gibt (asymmetrische Fragestellung), ist aus meiner Sicht die Orientierung an der zweifachen Varianzanalyse:

$$y = f(x, z)$$

In der Varianzanalyse wird versucht, die Variation in y dadurch zu erklären, dass sie auf Muster zurückgeführt wird, die sich aus Mechanismen und Effekten „möglichst niedriger Ordnung“ ergeben ($\bar{y}_{x_i}, \bar{y}_{z_j}$ wären 1. Ordnung, $\bar{y}_{x_i z_j}$ wäre 2. Ordnung).

(Auch die log-lineare Modellierung wurde genau aus dieser Analogie heraus entwickelt, allerdings für die symmetrische Fragestellung, aufgrund welcher Mechanismen es zu den Beobachtungen (x_i, z_j, y_u) kommt.)

Unter Interaktionen in der zweifachen Varianzanalyse versteht man die Abweichungen von dem Muster, das sich aufgrund des allgemeinen Bezugspunkts $\bar{y}$ und der Haupteffekte ($\bar{y}_{x_i} - \bar{y}$ und $\bar{y}_{z_j} - \bar{y}$) prognostizieren lässt:

$$\bar{y}_{x_i z_j} - \left(\bar{y} + \left(\bar{y}_{x_i} - \bar{y}\right) + \left(\bar{y}_{z_j} - \bar{y}\right)\right)$$

Deshalb ist es sinnvoll, diese Effekte als „nicht additive" Effekte zu interpretieren. „Multiplikativ" sollte man dann nur Effekte in y nennen, die genau und nur bei dem gemeinsamen Vorliegen zweier Bedingungen x_i und z_j auftreten.
Im allgemeinen Fall gibt es also vier Interaktionseffekte bzw. nicht additive Effekte (nämlich für die vier Kombinationen (x_i, z_j)).

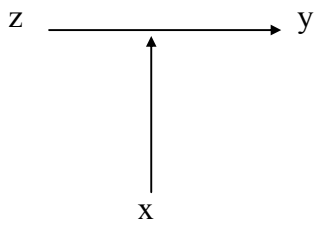
Die typologischen Effekte, die Rosenberg und wir bisher unterschieden haben, sind also alles spezielle Interaktionseffekte.

Wenn es einen Interaktionseffekt gibt wie
z.B.

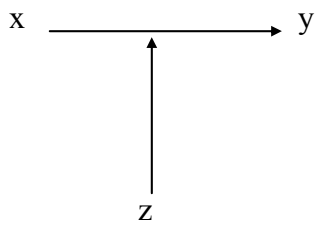
(xz) ⟶ y

, dann lässt sich dies asymmetrisch so formulieren, dass x die Beziehung (zy) spezifiziert und z die Beziehung (xy). (Wenn y nicht als abhängige Variable betrachtet wird, gilt auch : y spezifiziert die Beziehung (xz).)

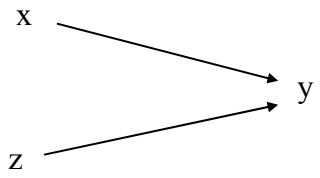
Anschaulich:



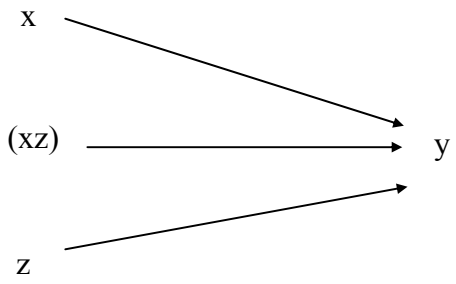
Oder auch:



Dies soll im folgenden Beispiel an der Analyse der Lebenszufriedenheit in Abhängigkeit von den Zufriedenheiten mit dem Beruf und mit Primärbeziehungen (Familie, Freundeskreis) gezeigt werden. Die Zahlen sind von Mayntz et al. (1971: 204f.), die dies als Beispiel für die multikausale Abhängigkeit einer Variablen y von den beiden Ursachen x und z angeben, wobei die Ursachen nicht korrelieren:



Was Mayntz et al. übersehen haben, ist, dass man in diesem Modell die starke Interaktion mit berücksichtigen sollte, was ich im Folgenden ausführe:



Erläuterung der folgenden varianzanalytischen Tabelle

So wie man bei statistischer Unabhängigkeit in einer Kontingenztabelle aus der Randverteilung die Besetzungszahlen in den einzelnen Kombinationszellen vorhersagen kann, lässt sich in der folgenden varianzanalytischen Tabelle eine Prognose für die Kombinationszellen aufgrund der Haupteffekte bzw. der Randverteilungen vornehmen.
Beispiel: Personen mit großer Berufszufriedenheit und kleiner Zufriedenheit mit den Primärbeziehungen: Ausgegangen wird von der durchschnittlichen Lebenszufriedenheit von 57,89 %. Bei großer Berufszufriedenheit liegt man um 18,11 % über der durchschnittlichen Lebenszufriedenheit. Bei kleiner Zufriedenheit mit den Primärbeziehungen liegt man um –7,89 % unter dem Durchschnitt. Die Prognose aufgrund des additiven Modells beträgt also 57,89 – 7,89 + 18,11 = 68,11 %. Tatsächlich wird aber eine durchschnittliche Lebenszufriedenheit von 85,71 % beobachtet. Personen mit dieser Kombination weisen also eine um 85,71 – 68,11 = + 17,60 % höhere Lebenszufriedenheit auf, als bei dem additiven Modell (aufgrund der Haupteffekte) zu erwarten. Die Berufszufriedenheit erweist sich in dieser Kombination als relativ wichtig für die allgemeine Lebenszufriedenheit.

Tabelle 2-29: Anteil der Personen mit großer Lebenszufriedenheit in Abhängigkeit von den Zufriedenheiten mit dem Beruf und den Primärbeziehungen $n_{y_2} / \left(n_{y_1} + n_{y_2}\right) = f\left(x, z\right)$

Zufriedenheit mit dem Beruf (z)	Zufriedenheit mit Primärbeziehungen (x): klein	Zufriedenheit mit Primärbeziehungen (x): groß	
klein	$\frac{30}{30+130} = 18{,}75\%$ Prognose: 57,89 - 7,89 - 20,11 = 29,89 Abweichung von Prognose: 18,75-29,89 = -11,14% Anteil des Typs: 160/950 = 16,84 %	$\frac{140}{140+150} = 48{,}28\%$ Prognose: 57,89+3,65-20,11=41,43 % Abweichung von Prognose: 48,28-41,43= +6,85 Anteil des Typs: 290/950 = 30,53 %	$\frac{170}{170+280} = 37{,}78\%$ $\bar{y}_{z_1} - \bar{y} =$ 37,78 – 57,89 = -20,11 % n_{z_1} = 450 ; 450/950 = 47,37 %
groß	$\frac{120}{120+20} = 85{,}71\%$ Prognose: 57,89 - 7,89 + 18,11 = 68,11 % Abweichung von Prognose: 85,71-68,11= +17,60% Anteil des Typs: 140/950 = 14,74 %	$\frac{260}{260+100} = 72{,}22\%$ Prognose: 57,89 + 3,65 + 18,11 = 79,65 % Abweichung von Prognose: 72,22-79,65= -7,45% Anteil des Typs: 360/950 = 37,89 %	$\frac{380}{380+120} = 76{,}00\%$ $\bar{y}_{z_2} - \bar{y} =$ 76,00 – 57,89 = +18,11 % n_{z_2} = 500 ; 500/950 = 52,63 %
	$\frac{150}{150+150} = 50{,}00\%$ $\bar{y}_{x_1} - \bar{y}$ = 50,00 – 57,89 = -7,89 % n_{x_1} = 300 ; 300/950 = 31,58 %	$\frac{400}{400+250} = 61{,}54\%$ $\bar{y}_{x_2} - \bar{y}$ = 61,54 – 57,89 = +3,65 % n_{x_2} = 650 ; 650/950 = 68,42 %	n = 950 $\frac{550}{550+400} = 57{,}89\%$

Das „saturierte Modell“ (=vollständige Modell mit allen möglichen Effekten) lässt sich auf die beiden folgenden äquivalenten Arten darstellen:

1) Deskription der 4 Kombinations-Typen

4 Typen:

$\bar{y}_{x=0,z=0} = 18{,}75\,\%$

$\bar{y}_{x=1,z=0} = 48{,}28\,\%$

$\bar{y}_{x=0,z=1} = 85{,}71\,\%$

} Interpretation: Die Zufriedenheit mit dem Beruf (z) geht einher mit einem stärkerem Zuwachs in der allgemeinen Lebenszufriedenheit als die Zufriedenheit mit den Beziehungen (x).

$\bar{y}_{x=1,z=1} = 72{,}22\,\%$ } Interpretation: Der Kombinations-Typ weist entgegen der Erwartung nicht die höchste allgemeine Lebenszufriedenheit auf, denn:

$$\bar{y}_{x=0,z=1} = 85{,}71\% > 72{,}22\% = \bar{y}_{x=1,z=1}$$

2) Strukturerklärung durch Haupteffekte und Interaktionseffekte

Beispiel für Typ :

$$\begin{aligned} 72{,}22\% &= \bar{y}_{x_2,z_2} \\ &= \bar{y} + (\bar{y}_{x_2} - \bar{y}) + (\bar{y}_{z_2} - \bar{y}) \\ &+ (\bar{y}_{x_2,z_2} - (\bar{y} + (\bar{y}_{x_2} - \bar{y}) + (\bar{y}_{z_2} = \bar{y}))) \\ &= 57{,}89 + (+3{,}65) + (+18{,}11) \\ &+ (72{,}22 - (57{,}89 + (+3{,}65) + (+18{,}11))) \end{aligned}$$

Interaktionseffekt (x_2,z_2) = -7,45%

Interpretation: Der (Haupt-) Effekt des Berufs (z) ist mit + 18,11 größer als der (Haupt-) Effekt der Beziehungen (x). Der Interaktionseffekt für (x_2,z_2) zeigt, dass die allgemeine Lebenszufriedenheit unter diesen Bedingungen um 7,45% weniger groß ist, als nach dem additiven Modell aufgrund der Haupteffekte zu erwarten.

Tabelle 2-30: Anteil der Personen mit kleiner Lebenszufriedenheit in Abhängigkeit von den Zufriedenheiten mit dem Beruf und den Primärbeziehungen $n_{y_1} / \left(n_{y_1} + n_{y_2}\right) = f(x, z)$

Zufriedenheit mit dem Beruf (z)	Zufriedenheit mit Primärbeziehungen (x): klein	Zufriedenheit mit Primärbeziehungen (x): groß	
klein	$\frac{130}{130+30} = 81{,}25\%$ Prognose: 42,11+7,89+20,11=70,11 Abweichung von Prognose: 81,25-70,11= +11,14 Anteil des Typs: 160/950 = 16,84 %	$\frac{150}{150+140} = 51{,}72\%$ Prognose: 42,11-3,65+20,11= +58,57 Abweichung von Prognose: 51,72-58,57=-6,85 Anteil des Typs: 290/950 = 30,53 %	$\frac{280}{280+170} = 62{,}22\%$ $\bar{y}_{z_1} - \bar{y} =$ 62,22 – 42,11 = +20,11 n_{z_1} = 450 ; 450/950 = 47,37 %
groß	$\frac{20}{20+120} = 14{,}29\%$ Prognose: 42,11 + 7,89-18,11 = 31,89 Abweichung von Prognose: 14,29-31,89 = -17,60 Anteil des Typs: 140/950 = 14,74 %	$\frac{100}{100+260} = 27{,}78\%$ Prognose: 42,11-3,65-18,11 = 20,35 Abweichung von Prognose: 27,78-20,35= +7,43 Anteil des Typs: 360/950 = 37,89 %	$\frac{120}{120+380} = 24{,}00\%$ $\bar{y}_{z_2} - \bar{y} =$ 24,00 – 42,11 = -18,11 % n_{z_2} = 500 ; 500/950 = 52,63 %
	$\frac{150}{150+150} = 50{,}00\%$ $\bar{y}_{x_1} - \bar{y}$ = 50,00 – 42,11 = +7,89 % n_{x_1} = 300 ; 300/950 = 31,58 %	$\frac{250}{250+400} = 38{,}46\%$ $\bar{y}_{x_2} - \bar{y}$ = 38,46 – 42,11 = -3,65 n_{x_2} = 650 ; 650/950 = 68,42 %	$\frac{400}{950} = 42{,}11\%$

Tabelle 2-31: Lebenszufriedenheit und Zufriedenheit mit den Primärbeziehungen (yx)

	Zufriedenheit mit den Primärbeziehungen (x)		
Lebenszufriedenheit (y)	150	250	400
	150	400	550
	300	650	950

$[xy] = 22500$
$\delta_{xy} = [xy]/n = 23{,}684$
$s_{xy} = [xy]/n^2 = 0{,}025$
$\Phi_{xy} = [xy]/207123 = 0{,}109$

Tabelle 2-32: Lebenszufriedenheit und Berufszufriedenheit (yz)

	Berufszufriedenheit (z)		
Lebenszufriedenheit (y)	280	120	400
	170	380	550
	450	500	950

$[yz] = 86000$
$\delta_{yz} = [yz]/n = 90{,}526$
$s_{yz} = [yz]/n^2 = 0{,}095$
$\Phi_{yz} = [yz]/222486 = 0{,}387$

Tabelle 2-33: Berufszufriedenheit und Zufriedenheit mit den Primärbeziehungen (zx)

	Zufriedenheit mit den Primärbeziehungen (x)		
Berufszufriedenheit (z)	160	290	450
	140	360	500
	300	650	950

$[zx] = 17000$

$\delta_{zx} = [zx] / n == 17{,}895$

$s_{zx} = [zx] / n^2 = 0{,}019$

$\Phi_{zx} = [zx] / 209464 = 0{,}081$

Tabelle 2-34: Kombinationen (Typen) von Lebenszufriedenheit, Zufriedenheit mit dem Beruf und Zufriedenheit mit den Primärbeziehungen (y, x, z)

Berufszufriedenheit klein ($z = z_1$)

Lebenszufriedenheit (y)	Zufriedenheit mit den Primärbeziehungen (x): klein	Zufriedenheit mit den Primärbeziehungen (x): groß	
klein	130 (/160 = 81,25 %) Anteil des Typs: 130/950=13,68 %	150 (/290 = 51,72 %) Anteil des Typs: 150/950=15,79 %	280
groß	30 (/160 = 18,75 %) Anteil des Typs: 30/950=3,16 %	140 (/290 = 48,28 %) Anteil des Typs: 140/950=14,74 %	170
	160	290	450(/950 = 47,37 %)

$[yx : z_1] = 13700$

$\delta_{yx} = [yx] / n_{z_1} == 30{,}444$

$s_{yx} = [yx] / n_{z_1}^2 = 0{,}068$

$\Phi_{yx} = [yx] / 46996 = 0{,}292$

Berufszufriedenheit groß ($z = z_2$)

Lebenszufriedenheit (y)	Zufriedenheit mit den Primärbeziehungen (x): klein	Zufriedenheit mit den Primärbeziehungen (x): groß	
klein	20 (/140 = 14,29 %) Anteil des Typs: 20/950=2,11%	100 (/360 = 27,78 %) Anteil des Typs: 100/950=10,53%	120
groß	120 (/140 =85,71 %) Anteil des Typs: 120/950=12,63%	260 (/360 = 72,22 %) Anteil des Typs: 260/950=27,37%	380
	140	360	500 (/950=52,63%)

$\lfloor yx : n_{z_2} \rfloor = -6800$

$\delta_{yx} = [yx] / n_{z_2} == -13{,}600$

$s_{yx} = [yx] / n_{z_2}^2 = -0{,}027$

$\Phi_{yx} = [yx] / 47940 = -0{,}142$

n = 950

Deskriptiv sind diese Typen besonders informativ: Die höchste allgemeine Lebenszufriedenheit findet sich mit 85,71 % bei Personen mit großer Berufszufriedenheit und kleiner Zufriedenheit mit den Primärbeziehungen, was auf eine reine Berufsorientierung dieser Personen hindeutet. Die geringste Lebenszufriedenheit findet sich mit 18,75 % bei den Personen mit geringer Zufriedenheit sowohl mit dem Beruf als auch mit den Primärbeziehungen.

Kovarianzzerlegung von yx nach z
(z spezifiziert yx)

$$s_{yx} = \frac{n_{z_1}}{n} s_{yx:z_1} + \frac{n_{z_2}}{n} s_{yx:z_2} + \frac{s_{xz} \quad s_{yz}}{\frac{n_{z_1}}{n} \quad \frac{n_{z_2}}{n}}$$

$$0{,}025 = 0{,}474 \cdot 0{,}068 + 0{,}526 \cdot (-0{,}027) + \frac{0{,}019 \cdot 0{,}095}{0{,}474 \cdot 0{,}526}$$

$$= 0{,}032 \qquad -0{,}014 \qquad +0{,}007$$

Der geringe Zusammenhang zwischen der Lebenszufriedenheit und der Zufriedenheit mit den Primärbeziehungen (s_{yx} = 0,025) setzt sich daraus zusammen, dass in den beiden Teilgruppen Gegenläufiges auftritt: Bei geringer Berufszufriedenheit hängen die Lebenszufriedenheit und die Zufriedenheit mit den Primärbeziehungen schwach positiv zusammen, während bei großer Berufszufriedenheit die Lebenszufriedenheit und die Zufriedenheit mit den Primärbeziehungen schwach negativ zusammenhängen. Das Produkt der Beziehungen mit dem Drittfaktor (z) fällt nicht ins Gewicht.

Tabelle 2-35: Kombinationen (Typen) von Lebenszufriedenheit, Zufriedenheit mit den Primärbeziehungen und Zufriedenheit mit dem Beruf (y, z, x)

Primärbeziehungen klein ($x = x_1$)

Lebenszufriedenheit (y)	Zufriedenheit mit dem Beruf (z): klein	Zufriedenheit mit dem Beruf (z): groß	
klein	130 (/160=81,25%) Anteil des Typs: 130/950=13,68 %	20 (/140=19,29%) Anteil des Typs: 20/950=2,11%	150
groß	30 (/160=18,75%) Anteil des Typs: 30/950=3,16 %	120 (/140=85,71%) Anteil des Typs: 120/950=12,63 %	150
	160	140	$\frac{300}{950} = 31{,}58\%$

$[yz : x_1] = 15000$

$\delta_{yz} = [yz] / n_{x_1} = 50$

$s_{yz} = [yz] / n_{x_1}^2 = 0{,}167$

$\Phi_{yz} = [yz] / 22450 = 0{,}668$

Primärbeziehungen groß ($x = x_2$)

Lebenszufriedenheit (y)	Zufriedenheit mit dem Beruf (z): klein	Zufriedenheit mit dem Beruf (z): groß	
klein	150 (/290=51,72%) Anteil des Typs: 150/950=15,79%	100 (/360=27,78%) Anteil des Typs: 100/950=10,53%	250
groß	140 (/290=48,28%) Anteil des Typs: 140/950=14,74%	260 (/360=72,22%) Anteil des Typs: 260/950=27,37%	400
	290	360	$\frac{650}{950} = 68{,}42\%$

$\lfloor yz : x_{z_2} \rfloor = 25000$

$\delta_{yz} = [yz] / n_{x_2} = 38{,}462$

$s_{yz} = [yz] / n_{x_2}^2 = 0{,}059$

$\Phi_{yz} = [yz] / 202176 = 0{,}245$

n = 950

Kovarianzzerlegung von yz nach x
(x spezifiziert yz)

$$s_{yz} = \frac{n_{x_1}}{n} s_{yz:x_1} + \frac{n_{x_2}}{n} s_{yt:x_2} + \frac{s_{yx} \quad s_{zx}}{\frac{n_{x_1}}{n} \quad \frac{n_{x_2}}{n}}$$

$$0{,}095 = 0{,}316 \cdot 0{,}167 + 0{,}684 \cdot 0{,}059 + \frac{0{,}025 \cdot 0{,}019}{0{,}316 \cdot 0{,}684}$$

$$= 0{,}053 \qquad = 0{,}040 \qquad + 0{,}002$$

Der mittelgroße Zusammenhang zwischen der Lebenszufriedenheit und der Berufszufriedenheit (s_{yz} = 0,095) setzt sich daraus zusammen, dass in der schwächer besetzten Gruppe mit kleiner Zufriedenheit mit den Primärbeziehungen (31,58 %) ein relativ starker Zusammenhang zwischen Lebenszufriedenheit und Berufszufriedenheit besteht (s_{yz} = 0,167) und in der stärker besetzten Gruppe mit großer Zufriedenheit mit den Primärbeziehungen (68,4 %) ein ebenfalls positiver, aber kleinerer Zusammenhang zwischen Lebenszufriedenheit und Berufszufriedenheit besteht. Das Produkt der Beziehungen zu dem Drittfaktor (x) fällt nicht ins Gewicht.

2.2.4.8 Kausale Interpretation von Zusammenhängen

Nachdem man zwischen je zwei Variablen Zusammenhänge gefunden hat oder auch nicht, stellt sich also die Frage, ob diese Beziehungen echt, d.h. kausal interpretierbar sind. Einige der Möglichkeiten, insbesondere zu Fehlschlüssen, seien noch einmal kurz gegenübergestellt (vgl. Tabelle 2-36).

Tabelle 2-36: „Echtheit“ von Korrelation im Hinblick auf die kausale Interpretation

Zusammenhang	Kausalität
Scheinkorrelation $[xy] \neq 0$ $[xy:z] = [xy:\neg z] = 0$	kein Zusammenhang nach Einbezug von Kontrollvariable(n)
Suppressor Variable im Fall $[xy] = 0$ $[xy] = 0, [xy:z] = [xy:\neg z] \neq 0$	Zusammenhang nach Einbezug von Kontrollvariable(n)
Spezifikation im Fall [xy] = 0. $[xy] = 0,\ sign\ [xy:z] = -sign\ [xy:\neg z]$	Zusammenhänge entgegengesetzt und heben sich auf
Bestätigung $[xy] = [xy:z] = [xy:\neg z]$	nicht falsifiziert; vorläufig akzeptiert

Résumé: Die Lösung des „Kausalitätsproblems“ liegt nach meiner Auffassung nicht in einer perfektionierten Definition der Kausalität, denn einerseits beginnt man mit der Forderung, es solle ein statistischer Zusammenhang vorliegen, andererseits zeigt die Diskussion der Kausaltypen, dass dies im Fall der scheinbaren Nicht-Kausalität gerade nicht gegeben ist. Die Lösung liegt nach meiner Auffassung in der Interaktion von theoretischen Vorstellungen, die in Modellen verdichtet werden, und Erfahrungswissen, das als Empirie aufbereitet wird. Sind die tatsächlichen

Beobachtungen verträglich mit Beobachtungen, die aufgrund eines theoretischen Modells zu erwarten wären?

Diese Frage wird i.a. nicht in einem einzigen Schritt beantwortet, sondern in Zyklen von Modellgenerierung und Modell-Modifikation aufgrund von Konsistenzüberlegungen und Erfahrungswissen.

Falls [xy] als Zusammenhang in der Gesamttabelle ermittelt ist, sollte man Kausalmodelle formulieren mit den Mechanismen, die den Zusammenhang produzieren:

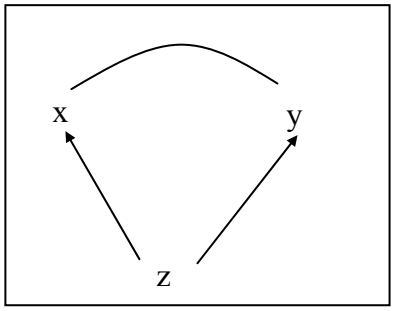
- direkter Kausaleffekt?
- indirekte/r Kausaleffekt/e?
- spurious bzw. Scheinkomponente/n des Zusammenhangs?

An diese Denkfigur knüpft die Pfadanalyse an.

2.3 Partielle Korrelation

Für metrisches Messniveau leistet der partielle Korrelationskoeffizient das Gleiche wie die bedingten Zusammenhänge in Teilgruppen auf nominalem Messniveau. Eine partielle Korrelation zweier Variablen misst den Zusammenhang zweier Variablen unter der Bedingung, dass der Einfluss einer oder mehrerer anderer Variablen kontrolliert wird. Man sagt auch, dieser Einfluss wird herauspartialisiert.

Abbildung 2-17: Einfluss von z auf x und y



Da Korrelationskoeffizienten berechnet werden sollen, kann man ohne Beschränkung der Allgemeinheit mit standardisierten Variablen arbeiten. Standardisierung bedeutet die Normierung von x und y derart, dass $\bar{x} = \bar{y} = 0$ und $s_x = s_y = 1$. Damit ist $r_{xy} = s_{xy}$.
Im Folgenden wird an die lineare Regression angeknüpft, weshalb wir uns deren Prinzip nochmals vergegenwärtigen.

Ausgehend von der Funktion y = a + bx soll die Summe der quadrierten Abstände der Schätzwerte von den y-Werten, d.h. $\sum_{i=1}^{n}(y_i - \hat{y})$, minimiert werden. Dabei errechnet sich der Anstieg der Regressionsgeraden (Regressionskoeffizient) als $b = \frac{s_{xy}}{s_x^2}$ und der Abstand zur x-Achse als $a = \bar{y} - b\bar{x}$.

Bei der Partialkorrelation verfährt man wie folgt: Zunächst wird eine Regressionsgleichung $\hat{x}$ aufgestellt, mit deren Hilfe der Einfluss von Z auf X vorhergesagt werden kann. Ziel ist es dann, durch Subtraktion die Residualwerte x - $\hat{x}$ zu erhalten, die nicht von Z beeinflusst werden. Entsprechend geht man mit y vor. Werden die derart „bereinigten“ x- und y-Werte miteinander korreliert, erhält man die Partialkorrelation von x und y. Die Partialkorrelation drückt also den linearen Zusammenhang zweier Variablen aus, bei denen der lineare Einfluss der Drittvariablen „herauspartialisiert” wurde.

Zur Erinnerung: Die **Kovarianz** zwischen x und y berechnet sich nach

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y}).$$

Bei zentrierten Variablen ist $\bar{x} = \bar{y} = 0$ und somit $s_{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i \; y_i$.

Die Regressionsgleichung von x auf z gibt den Einfluss von z auf x an und hat bei standardisierten Variablen die Form:
$\hat{x} = r_{xz} z$

Entsprechend:
$\hat{y} = r_{yz} z$

Mit Hilfe der Vektorrechnung kann gezeigt werden, dass in der Variablen x - $\hat{x}$ der Einfluss von Z eliminiert ist. Durch das skalare Produkt kann die Länge eines Vektors, der Winkel zwischen zwei Vektoren u.a. berechnet werden. Das **innere Produkt** oder **Skalarprodukt** ‹ x, z › ist definiert als ‹ x, z › := n s_{xz}. Für x = (x_1, ..., x_n) und y = (y_1, ..., y_n) beträgt das Skalaprodukt $< x, y >= \sum_{i=1}^{n} x_i \; y_i$.
Zwei Vektoren stehen senkrecht aufeinander (sind orthogonal) genau dann, wenn gilt ‹ x, y › = 0.
Hiernach kann der Zähler von $r_{x-\hat{x},z}$ bestimmt werden: ‹x - r_{xz}z, z › = ‹ x, z › - r_{xz} ‹ z, z › = 0, wobei ‹ z, z › = n ist. Somit ist $r_{x-\hat{x},z} = 0$.

Entsprechendes gilt für y - $\hat{y}$ mit $\hat{y} = r_{yz}$ z.

Der **partielle Korrelationskoeffizient** zwischen x und y unter Kontrolle des Einflusses der Variablen z (Bezeichnung: $r_{xy \cdot z}$) ist definiert als der einfache Korrelationskoeffizient zwischen x - $\hat{x}$ und y - $\hat{y}$.
In dem einfachsten Fall, nämlich bei *einer* Kontrollvariablen z, erhält man eine einfache Formel:

$$r_{xy \cdot z} = \frac{r_{xy} - r_{xz}\, r_{yz}}{\sqrt{1 - r_{xz}^2}\sqrt{1 - r_{yz}^2}}$$

Beweis:

$$< x - \hat{x}, y - \hat{y} >=< x, y > - r_{yz} \underbrace{< x, z >}_{n \cdot r_{xz}} - r_{xz} \underbrace{< z, y >}_{n \cdot r_{zy}} + r_{xz} r_{yz} \underbrace{< z, z >}_{n} =< x, y > - n \cdot r_{xz} r_{yz}$$

Also: $\frac{1}{n} < x - \hat{x}, y - \hat{y} >= r_{xy} - r_{xz}\, r_{yz}$

Für die erste Varianz kann man berechnen:

$$< x - \hat{x}, x - \hat{x} >=< \underbrace{x, x}_{n} > - r_{xz} \underbrace{< x, z >}_{n \cdot r_{xz}} - r_{xz} \underbrace{< z, x >}_{n \cdot r_{xz}} + r_{xz}^2 \underbrace{< z, z >}_{n} = n - n\, r_{xz}^2$$

Also ist die Varianz: $\frac{< x - \hat{x}, x - \hat{x} >}{n} = 1 - r_{xz}^2$

Entsprechend die zweite Varianz: $1 - r_{xz}^2$

Beispiel: Erklärung der durchschnittlichen Ausbildungsvergütung in einem Betrieb
x: In dem Betrieb gibt es einen Betriebsrat.
y: Höhere Ausbildungsvergütung.

$r_{yx} = 0{,}41$ S = .03

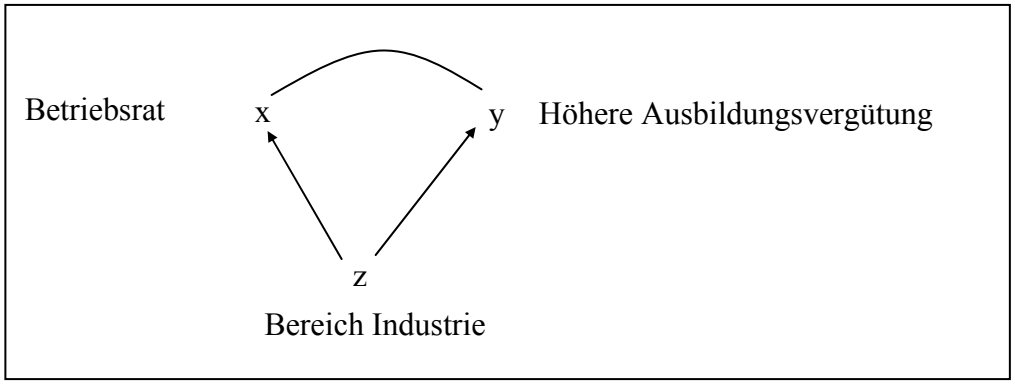
Liegt es an Beschäftigtenzahl z?

$r_{yx.z} = 0{,}39$ S = .05

Aber: Für z = Bereich Industrie

$r_{yx.z} = 0{,}04$ S = .40

Abbildung 2-18: Partielle Korrelation im Fall der „Scheinkorrelation“



Untersucht man den Effekt eines Testfaktors z auf die Beziehung zwischen zwei Variablen x und y, so kann man diese Formel umformen zu:

Zerlegungsformel in Analogie zur Tabellenanalyse:

$$\underbrace{r_{xy}}_{\text{Gesamtzu-sammenhang}} = \underbrace{\sqrt{1-r_{xz}^{\,2}}\sqrt{1-r_{yz}^{\,2}}}_{\text{Koeffizient i.a. } \neq 1} \underbrace{r_{xy.z}}_{\text{Bereinigter Zusammenhang}} + \underbrace{r_{xz}\,r_{yz}}_{\text{Zusammenhänge mit dem Drittfaktor}}$$

$[xy] = a\,[xy : z] + b\,[xz]\,[yz]$

Für z = Dichotomie:
$[xy] = \alpha\,[xy : z_1] + \beta\,[xy : z_2] + \gamma\,[xz]\,[yz]$

„Dilemma“ (Davis) zwischen perfekter Vorzeichenregel und Aussagen über „perfekte“ Maßzahlen:

Tabellenanalyse:

$$\delta_{xy} = \delta_{xy:z_1} + \delta_{xy:z_2} + \gamma \ \delta_{xz}\delta_{yz}$$

Für δ : perfekte Vorzeichenregel,
aber: δ nicht normiert.
Für Φ : Vorzeichenregel nur „Daumenregel“,
aber: Φ ist normierte Maßzahl.

Partielle Korrelation:
Für r: Vorzeichenregel nur „Daumenregel“,
aber: r ist normierte Maßzahl.

Für die Kovarianz: $s_{xy} = s_{xy.z} + \frac{1}{s_z^2} s_{xz} s_{yz}$

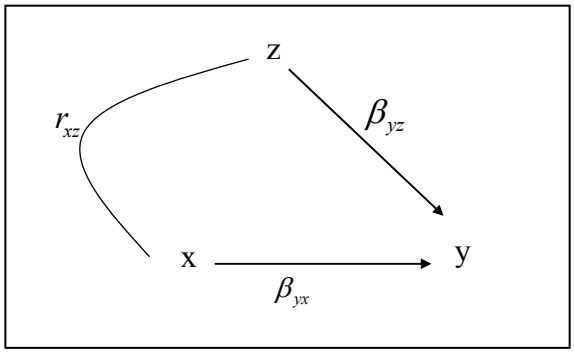
Perfekte Vorzeichenregel, aber: s_{xy} ist keine „perfekte“ Maßzahl, da die Maximalwerte +1/-1 bei vielen Datenkonstellationen gar nicht angenommen werden können.

Die Kovarianz ist im Allgemeinen nicht normiert, im Fall der Vierfeldertafel allerdings erfüllt sie die Normierungsbedingung, denn:

$$-1 \le \frac{ad - bc}{n^2} \le +1$$

In dem Sinne, dass sie diese notwendige Bedingung für eine Maßzahl erfüllt, löst die Kovarianz das „Dilemma“ von Davis zwischen Vorzeichenregel und Maßzahl.

Abbildung 2-19: Pfadanalyse



$$r_{xy} = \beta_{yx} + r_{xz}\beta_{yz}$$

Gesamtzusammenhang gleich direkter (bereinigter) Effekt plus:

- indirekter Effekt über z $(\beta_{yz}\beta_{zx})$
- bzw. Scheinkomponente aufgrund von z $(\beta_{yz}\beta_{xz})$
- bzw. korrelierter Effekt $(r_{xz}\beta_{yz})$

In den ersten beiden Fällen gilt die Vorzeichenregel für den Vergleich des Gesamtzusammenhangs mit den jeweiligen Effekten (β). Im dritten Fall ist eine Vorzeichenregel nicht so sinnvoll, da es sich um unterschiedliche Konzepte handelt.

Allgemein gilt:

Cov (x, y) = E [(x - μ_x) (y - μ_y)]
= E (xy) - μ_x μ_y
Var (x) = $\mathrm{Cov}(x,x) = E(x^2) - \mu_x^2$

Speziell für Dichotomien geht man von den Indikatorenfunktionen x = 1_A, y = 1_B aus. Sie können nur die Werte 0 und 1 annehmen und zwar den Wert 1 genau für die Wertemenge A bzw. B. Man erhält dann als Differenz des Kreuzproduktes von x, y:
Cov (x, y) = E (1_A 1_B) - p_A p_B

Diese Kovarianz entspricht nach der Zerlegungsformel im Abschnitt 2.2.3 der Differenz
p_{ij} - $p_i \cdot p_j$.

Für z = 1_C:

Var (z) = E (1_C^2) - p_C^2 = p_C - p_C^2 = p_C (1 - p_C)

Für die **Kovarianz** erhält man im Fall von Dichotomien die Zerlegung:

$$s_{ij} = p_k s_{k(i,j)} + p_{\bar{k}} s_{\bar{k}}(i,j) + \frac{s_{ik}\, s_{jk}}{p_k \cdot p_{\bar{k}}}$$

Die Lazarsfeld'sche Zerlegung mit dem Drittfaktor $(k, \bar{k})$ lautet dagegen:

$$|ij| = \frac{|ij:k|}{p_k} + \frac{|ij:\bar{k}|}{p_{\bar{k}}} + \frac{|ik\|jk|}{p_k\, p_{\bar{k}}}$$

Der Zusammenhang besteht darin, dass |ij| = s_{ij}, aber $\frac{|ij:k|}{p_k} = p_k \cdot s_{k(i,j)}$.
(Vgl. hierzu ausführlicher die allgemeine Kovarianzzerlegung in Kap. 4, Punkt 4.8.1).

Falls z weder mit x noch mit y korreliert (r_{xz} = r_{yz} = 0), so folgt aus der Formel für die partielle Korrelation, dass die partielle Korrelation in diesem Fall gleich der einfachen Korrelation ist: $r_{xy \cdot z}$ = r_{xy}

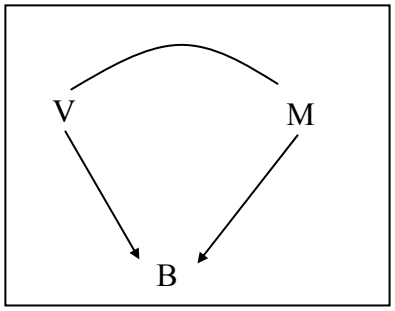
Mit der obigen Definition kann man jeweils rekursiv definieren, wie die partielle Korrelation unter der Kontrolle von mehreren Variablen zu berechnen ist.
Zum Beispiel gilt bei der Kontrolle von zwei Variablen z und w:

$$r_{xy\cdot zw} = \frac{r_{xy\cdot z} - r_{xw\cdot z}\ r_{yw\cdot z}}{\sqrt{1-r_{xw\cdot z}^2}\sqrt{1-r_{yw\cdot z}^2}}$$

Damit ist der Einfluss von z jeweils schon ausgeschaltet. Das Ergebnis ist unabhängig von der Reihenfolge, in der man die Kontrollvariablen berücksichtigt.

Inhaltliches Beispiel:
Politisches Interesse (Vater, Mutter, Befragte)

Abbildung 2-20: Beispiel für partielle Korrelation



r_{VM} = 0,37, S = .001
r_{VB} = 0,27, S = .001
r_{MB} = 0,23, S = .001

$r_{VB.M}$ = 0,21, S = .001
$r_{MB.V}$ = 0,15, S = .006

Das politische Interesse des Befragten korreliert sowohl mit dem politischen Interesse des Vaters als auch der Mutter, deren politisches Interesse ebenfalls korreliert.
Kontrolliert man den Einfluss jeweils eines Elternteils, so geht zwar der bereinigte Zusammenhang zwischen dem anderen Elternteil und dem Befragten zurück, die Korrelation bleibt aber signifikant.

Literaturverzeichnis

Andersen, E.B., 1997: *Introduction to the statistical analysis of categorial data.* New York: Springer.

Backhaus, K., Erichson, B., Plinke, W., Weiber, R., 2008[12]: *Multivariate Analysemethoden.* Berlin: Springer.

Davis, J.A., 1971: *Elementary survey analysis*. Englewood Cliffs: Prentice Hall.

Durkheim, E., 1897: *Le suicide: étude de sociologie*. 9. éd. Paris: Presses Univ. de France (1997).

Goodman, L.A., 1970: *The multivariate analysis of qualitative data: Interactions among multiple classifications.* In: Journal of the American Statistical Association 65: 226-256.

Goodman, L.A., 1972: *A general model for the analysis of surveys*. In: American Journal of Sociology 77: 1035-1086.

Harder, Th., 1974: *Werkzeug der Sozialforschung*. München: UTB.

Hirschi, T., Selvin, H.C., 1967: *Delinquency research. An appraisal of analytic methods.* New Brunswick: Transaction Publications, 1996.

Hummell, H.-J., Ziegler, R., 1982: *Zur Verwendung linearer Modelle bei der Kausalanalyse nicht-experimentieller Daten.* In: Dieselben (Hg.): Korrelation und Kausalität. Band 1. Stuttgart: Enke.

Hyman, H., 1968: *Survey design and analysis: principles, cases and procedures*. With a Foreword by Paul F. Lazarsfeld. 10. printing. New York: The Free Press.

Lazarsfeld, P.F., 1955: *Interpretation of statistical relations as a research operation*. In: Lazarsfeld, P.F., Rosenberg, M. (Hg.): The language of social research. 7. printing. New York: The Free Press (1967), 115-125.

Lazarsfeld, P.F., 1961: *The Algebra of Dichotomous Systems*. In: Solomon, H. (Hg.): *Item Analysis and Prediction*. Stanford: Stanford University Press, 111-157.

Lazarsfeld, P. F., 1966: *L' algèbre des systèmes dichotomiques*. In: Boudon, R., Lazarsfeld, P.F.: L'analyse empirique de la causalité. 2me éd. Paris: Mouton (1969), 255-275.

Lazarsfeld, P. F., 1968: *The analysis of attitude data*. In: International Encyclopaedia of the Social Sciences, Vol. 15, 419-429.

Mayntz, R., Holm, K., Hübner, P., 1978[5]: *Einführung in die Methoden der empirischen Soziologie.* Opladen: Westdeutscher Verlag.

Nie, N.H. et al., 1975[2]: *Statistical package for the social sciences* (SPSS). New York: McGraw-Hill.

Rosenberg, M., 1968: *The logic of survey analysis*. New York: Basic Books.
Schmierer, Chr., 1975: *Tabellenanalyse*. In: Holt, K. (Hg.): Die Befragung. Band 2. München: UTB.

Schnell, R., Hill, P.B., Esser, E., 2008[8]: *Methoden der empirischen Sozialforschung*. München, Wien: Oldenbourg.

Van de Geer, J.P., 1971: *Introduction to multivariate analysis for the social sciences*. San Francisco: Freeman.

Zeisel, H., 1970: *Die Sprache der Zahlen*. Köln: Kiepenheuer & Witsch.

3. Multiple Regressionsanalyse und Pfadanalyse

Im Folgenden werden die beiden wichtigsten metrischen Ansätze zur sozialwissenschaftlichen Datenanalyse behandelt: In der multiplen Regression versucht man, die Variation eines interessierenden Phänomens (abhängige Variable) auf die Variation einer Reihe von Erklärungsfaktoren (unabhängige Variablen) zurückzuführen. Die wichtigsten Interpretationshilfen dabei sind der Anteil erklärter Varianz und die Effekte (= Wirkungen) der unabhängigen Variablen auf die abhängige Variable (bei quasi-experimenteller Betrachtung). In der Pfadanalyse werden alle theoretisch möglichen Mechanismen herausgearbeitet, die durch ihr additives Zusammenwirken die Höhe jedes statistischen Zusammenhangs bestimmen: Direkte und indirekte Kausaleffekte, scheinkausale Komponenten und Assoziationseffekte.

3.1 Multiple Regressionsanalyse

Die Nützlichkeit der multiplen Regression besteht insbesondere in folgenden Punkten:

1) Es kann angegeben werden, in welchem Ausmaß eine abhängige Variable durch mehrere unabhängige Variablen erklärt wird. Anstelle von „abhängiger Variable“ wird auch von Kriteriumsvariable gesprochen, die unabhängigen Variablen werden Prädiktoren bzw. Prädiktorvariablen genannt. Prädiktorvariablen und Kriteriumvariablen setzen mindestens Intervallskalenniveau (metrisches Messniveau) voraus.
2) Der Einfluss einer unabhängigen Variablen (Prädiktor) auf die abhängige Variable (Kriterium) kann untersucht werden.
3) Die Struktur eines solchen komplexen Abhängigkeitsverhältnisses lässt sich – ähnlich wie bei der partiellen Korrelation – untersuchen.
4) Mit Hilfe der Regressionskoeffizienten und der Regressionsgleichung kann die abhängige Variable geschätzt werden (Interpolation: zwischen den beobachteten Werten; Extrapolation: über den Bereich der beobachteten Werte hinaus). Die Genauigkeit der Schätzung lässt sich bestimmen.
5) Der Einsatz der multiplen Regression sollte theoriegeleitet erfolgen. Der Forscher muss sich genau überlegen, welche Variablen in welcher Beziehung zueinander stehen könnten. Hierbei spielt sein Vorwissen und Informationen aus anderen Forschungen eine wichtige Rolle.
6) Die Komplikation besteht darin, dass in nahezu allen Anwendungen die Erklärungsfaktoren untereinander korrelieren, sodass sich die Frage stellt, welchen Beitrag die einzelnen Erklärungsfaktoren zur Gesamterklärung leisten. Dabei muss man mit den Phänomenen rechnen, die in der Tabellenanalyse eingeführt wurden: Überschneidungen, Suppressor, Distorter, Scheinkomponenten der Erklärung, Interaktionen etc.

Beispiel:
Zu erklären ist die Affinität zu den Jusos (versus „Basisgruppen“). Die Streuung dieser Variablen lässt sich zu 85,4 % auf die Variation der 8 ausgewählten Prädiktoren zurückführen (vgl. Tabelle 3-1). Betrachtet man die Erklärungskraft jedes Prädiktors isoliert und kumuliert die Erklärungsanteile, so erhielte man über 100 %, ein nicht sinnvolles Ergebnis. D.h. dass es Überschneidungen in der Erklärung gibt, die sowohl bei den Erklärungsanteilen als auch bei den Effekten berücksichtigt werden müssen. Dazu gibt es eine Reihe von Koeffizienten, die in der Tabelle 3-1 zusammengefasst sind und die nun im Folgenden erläutert werden. Hervorgehoben werden das Suppressor-Phänomen, dass ein Effekt nach Herausrechnen der übrigen Prädiktoren größer ist als vorher (Erklärungsfaktor x_3), und das Distorter-Phänomen, dass der eigentliche Kausaleffekt ein anderes Vorzeichen hat, als „an der Oberfläche der Korrelationen“ zu vermuten (Erklärungsfaktor x_4).

Tabelle 3-1: Zu erklärende Variable y: Affinität zu den Jusos (versus „Basisgruppen)

Prädiktoren	Gesamt-zusammen-hang r_{y,x_i}	Erklärungs-kraft bei isolierter Betrachtung r^2_{y,x_i}	Direkter Effekt oder „bereinigter“ Effekt β_{y,x_i}	„Bereinigter“ Zusammen-hang $r_{y,x_i-\hat{x}_i}$	„Bereinigte“ Erklärungskraft $r^2_{y,x_i-\hat{x}_i}$ $\left(= F \cdot \frac{1-R^2}{n-k-1}\right)$	Testwert F (Signifi-kanz)	Multiple R^2 bei schrittweiser Regression	ΔR^2 bei schrittweiser Regression
x_1: Affinität zu „grüner“ Kommunalprartei	- 0,56	31 %	- 0,38	- 0,35	12 %	49,3	31 %	31 %
x_2: Affinität zur SPD	0,48	23 %	0,31	0,30	9 %	37,4	47 %	16 %
x_3: Pro Frauen-erwerbstätigkeit	0,24	6 %	↑ 0,37	↑ 0,36	13 %	53,3	57 %	10 %
x_4: Pro Wohnge-meinschaft als Lebensform	0,39	15 %	**- 0,21**	**- 0,20**	4 %	15,1	65 %	8 %
x_5: Verdienst wichtig bei Berufsentscheidung	0,29	8 %	0,27	0,26	7 %	28,6	71 %	6 %
x_6: Vater gewerkschaftsnah	0,28	8 %	0,23	0,22	5 %	22,3	77 %	6 %
x_7: Kommunistische Partei positiv	- 0,46	21 %	- 0,22	- 0,20	4 %	17,2	82 %	5 %
x_8: Eigene Berufs-aussichten positiv	0,40	16 %	0,20	0,20	4 %	14,7	85 %	3 %
Multiple R^2 = 85,4 %		($\sum$ > 100 %)			($\sum$ = 58 %)			

3.1.1 Das Grundprinzip der einfachen Regression, geometrische Interpretation und Matrixschreibweise

(1) Grundprinzip

In der **einfachen Regression** (vgl. meinen Band zur Deskriptiv- und Inferenzstatistik) wird eine abhängige Variable y mit Hilfe der Information über eine unabhängige Variable x aufgrund eines linearen Ansatzes geschätzt: $\hat{y} = a + bx$.

Falls für n Untersuchungseinheiten bzw. Personen die Wertepaare (x_1, y_1), ..., (x_n, y_n) gegeben sind, lassen sich aufgrund der Forderung, dass die Schätzung einen minimalen Fehler $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - (a + bx_i))^2$ (das ist die Summe aller quadrierten Differenzen zwischen den beobachteten Werten y_i und den durch die Gleichung $\hat{y}_i = a + bx_i$ geschätzten Werten y_i über alle Untersuchungseinheiten hinweg) haben soll, die Koeffizienten a und b nach der *Methode der kleinsten Quadrate* berechnen als $b = \frac{s_{xy}}{s_x^2}$ und $a = \bar{y} - b\bar{x}$, wobei $s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ die Kovarianz von x und y und $s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ die Varianz von x ist.

Die Streuung der abhängigen Variablen (der Einfachheit halber nehmen wir die „Sum of Squares“ $SS_y := n \cdot s_y^2$) lässt sich zerlegen in zwei Komponenten:

die durch die Regression *nicht erklärte* residuale Streuung $\left(SS_{res} = \sum(y_i - \hat{y}_i)^2\right)$,
die durch die Regression *erklärte* Streuung $\left(SS_{reg} = SS_y - SS_{res} = \sum(\hat{y}_i - \bar{y})^2\right)$:

$$\boxed{SS_y = SS_{reg} + SS_{res}}$$

Als Anteil der durch die Regression auf x erklärten Varianz von y erhält man:

$$\frac{SS_{reg}}{SS_y} = b^2\frac{s_x^2}{s_y^2} = \frac{s_{xy}^2}{s_x^2 s_y^2} =: r_{xy}^2$$

Das Bestimmtheitsmaß (Determinationskoeffizient) r_{xy}^2 ist das Quadrat des Pearson-Bravais' schen Produkt-Moment-Korrelationskoeffizienten r_{xy}. Der Anteil der durch die Regression erklärten Varianz gibt an, in welchem Ausmaß die abhängige Variable y durch die unabhängige Variable x erklärt wird.

$b = r \Leftrightarrow s_x = s_y$

Für standardisierte Variablen gilt also (wegen $s_x = s_y = 1$): $b = r$

Der Korrelationskoeffizient r hat die Eigenschaft, dass für Konstanten c, d (mit $c > 0$) gilt:

$r_{y,\, cx + d} = r_{y,\, x}$. Also ist r gegenüber linearen Transformationen invariant. Die Korrelation zwischen der linear transformierten x-Variablen und der y-Variablen ist mit der ursprünglichen Korrelation identisch.

Für die Schätzung von y nach $\hat{y} = a + bx$ erhält man also: $\left|r_{y,\hat{y}}\right| = \left|r_{y,x}\right|$

Der einfache Korrelationskoeffizient zwischen x und y ist dem absoluten Betrag nach gleich dem Korrelationskoeffizienten zwischen den beobachteten und den geschätzten Werten für die abhängige Va
riable.

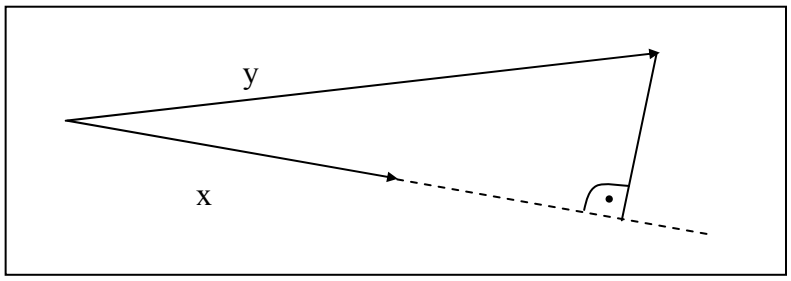
Für den multiplen Regressionsansatz $\hat{y} = a + b_1 x_1 + ... + b_k x_k$ (also mit der Einführung von k Prädiktoren) wird in entsprechender Verallgemeinerung definiert werden:
$R_{y;x_1,...,x_k} := r_{y,\hat{y}}$

R wird als multipler Korrelationskoeffizient bezeichnet. Er drückt die Korrelation zwischen den vorausgesagten Kriteriumswerten und den tatsächlichen Kriteriumswerten aus.

(2) Geometrische Interpretation der Regressionsanalyse

Die Gleichung $\hat{y} = a + bx$ beschreibt eine Gerade ($\hat{y} = a + bx_1 + cx_2$ eine Ebene). Die Minimierung von $\sum (y_i - \hat{y}_i)^2$ ist äquivalent zur Minimierung des Abstandes $|| y - \hat{y} ||$ des Vektors der beobachteten Werte von dem geschätzten Unterraum (Gerade, Ebene etc.). Diese Eigenschaft des minimalen Abstandes kommt genau der Projektion des Beobachtungsvektors in diesen Unterraum (Gerade, Ebene etc.) zu (vgl. Abbildung 3-1).

Abbildung 3-1: Bestimmung der Regressionsschätzung als orthogonale Projektion



Insbesondere gilt also, dass das Residuum $y - \hat{y}$ orthogonal ist zu (senkrecht steht auf) x_i, i = 1, ..., k. (Zum algebraischen Beweis s.u.)

(3) Matrixschreibweise

Eine wesentliche Vereinfachung der Darstellung der multiplen Regression erhält man, wenn man zur Matrix-Schreibweise übergeht.
Die Matrixschreibweise für den Regressionsansatz soll dadurch eingeführt werden, dass der Fall der einfachen linearen Regression noch einmal mit ihr formuliert wird.

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}_{(n,1)} \qquad y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}_{(n,1)} \qquad u := \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}_{(n,1)}$$

Aus der Geradengleichung $\hat{y} = a + bx$ wird in Matrixschreibweise: $\hat{y} = au + bx$.

Minimiert werden soll:

$$\|y-\hat{y}\|^2 = \prec y-\hat{y}, y-\hat{y} \succ = \underbrace{\underbrace{(y-\hat{y})'}_{(1,\,n)} \cdot \underbrace{(y-\hat{y})}_{(n,\,1)}}_{(1,\,1)\text{ nach der Matrixmultiplikation}}$$

$$= \quad (y-bx-au)' \quad (y-bx-au)$$

$$= \quad y'y + b^2 x'x + a^2 \underbrace{u'u}_{n} - 2\,b\,x'y - 2\,a\,u'y + 2\,a\,b\,u'x$$

Die transponierte Matrix A' zu einer Matrix A ist die an der Diagonale gespiegelte Matrix, sodass aus einer (m, n)-Matrix A eine (n, m)-Matrix A' wird.
Die partiellen Ableitungen nach a und b werden als notwendige Bedingung gleich Null gesetzt:

$$\frac{\partial f}{\partial a} = 2an - 2u'y + 2bu'x = 0$$

$$\frac{\partial f}{\partial b} = 2bx'x - 2x'y + 2au'x = 0$$

$$\Rightarrow \quad a = \frac{u'y - bu'x}{n}$$

$$b = \frac{nx'y - (u'x)\,(u'y)}{nx'x - (u'x)^2}$$

Für normierte Variablen (a = 0) also: x' x b = x' y

$$\left(\frac{x'x}{n}\right) b = \left(\frac{x'x}{n}\right) = r_{xy}$$

$$b = \left(\frac{x'x}{n}\right)^{-1} r_{xy}$$

Genau dieser Weg wird in der multiplen Regressionsanalyse ebenfalls beschritten werden.

3.1.2 Die Regressionskoeffizienten b und beta

Wenn man mehrere unabhängige Variablen x_i $(i = 1,\ldots,k)$ berücksichtigt, kann man die Koeffizienten b_i benutzen, um die abhängige Variable y für bestimmte Werte der unabhängigen Variablen x_i zu schätzen: $\hat{y} = a + b_1 x_1 + \ldots + b_k x_k$. Alle Werte werden dabei in den ursprünglichen Maßeinheiten verwendet. Wegen der in der Regel verschiedenen Maßeinheiten der unabhängigen Variablen geben die Koeffizienten b_i keinen Aufschluss über die relative Bedeutung der jeweiligen Prädiktoren für die abhängige Variable. Ist man an letzterem interessiert, so muss man die verschiedenen Größenordnungen der Maßeinheiten herausrechnen, d.h. mit **standardisierten Variablen** arbeiten. Die z-Transformation z.B. für die Variable x lautet: $z_x = \frac{x - \bar{x}}{s_x}$

(Der Regressionskoeffizient b wird genauer als b_{yx} bezeichnet, womit der Einfluss von x auf y gekennzeichnet wird.)

Bei *nicht standardisierten* Variablen ergab sich der **Regressionskoeffizient b** als:

$$b_{yx} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{s_{xy}}{s_x^2}$$

Für die *standardisierten Variablen* $z_{x_i} = \frac{x_i - \bar{x}}{s_x}$, $z_{y_i} = \frac{y_i - \bar{y}}{s_y}$ erhält man den **Regressionskoeffizienten beta**:

$$\beta_{yx} = b_{z_y, z_x} = \frac{\sum \left(\frac{x_i - \bar{x}}{s_x} \right) \cdot \left(\frac{y_i - \bar{y}}{s_y} \right)}{\sum \left(\frac{x_i - \bar{x}}{s_x} \right)^2}$$

Dabei ist der Nenner gleich:

$$\frac{\sum (x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2 / n} = n$$

Deshalb gilt:

$$b_{z_y, z_x} = \frac{\frac{1}{n} \sum (x_i - \bar{x}) \cdot (y_i - \bar{y})}{s_x \cdot s_y} = \frac{s_{xy}}{s_x \cdot s_y} = \frac{s_{xy}}{s_x^2} \cdot \frac{s_x}{s_y} = b_{y,x} \cdot \frac{s_x}{s_y}$$

Für den **Korrelationskoeffizienten** galt bei nicht standardisierten Variablen die Berechnungsformel:

$r_{xy} = \frac{\sum (x_i - \bar{x}) \quad (y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \cdot \sqrt{(y_i - \bar{y})^2}}$, für standardisierte Werte entsprechend:

$$r_{z_y z_x} = \frac{\sum \left(\frac{x_i - \bar{x}}{s_x} \right) \cdot \left(\frac{y_i - \bar{y}}{s_y} \right)}{\sqrt{\sum \left(\frac{x_i - \bar{x}}{s_x} \right)^2} \cdot \sqrt{\sum \left(\frac{y_i - \bar{y}}{s_y} \right)^2}}$$

Dabei ist der Nennerausdruck (wie oben) gleich n.

Also folgt:

$$r_{z_x z_y} = \frac{\frac{1}{n} \sum (x_i - \bar{x}) \cdot (y_i - \bar{y})}{s_x \cdot s_y} = \frac{s_{xy}}{s_x \cdot s_y} = r_{xy}$$

Für die Kovarianz erhält man:

$$s_{z_x z_y} = \frac{1}{n} \sum \left(\frac{x_i - \bar{x}}{s_x} \right) \cdot \left(\frac{y_i - \bar{y}}{s_y} \right) = \frac{s_{xy}}{s_x \cdot s_y}$$

Also gilt:

$$r_{xy} = r_{z_x, z_y} = s_{z_x, z_y} = b_{z_y, z_x} = \frac{s_{xy}}{s_x \cdot s_y} = b_{yx} \cdot \frac{s_x}{s_y}$$

Das heißt:

1) Bei standardisierten Variablen stimmen Regressionskoeffizient, Kovarianz und Korrelationskoeffizient überein.
2) Für alle Variablen gilt: Korrelationskoeffizienten sind Regressionskoeffizienten der zugehörigen standardisierten Variablen.
3) Die Regressionskoeffizienten bei standardisierten Variablen und bei nicht standardisierten Variablen stehen in folgendem Zusammenhang:

$$b_{z_y, z_x} = b_{yx} \cdot \frac{s_x}{s_y}$$

Zwar haben standardisierte Regressionskoeffizienten den Nachteil, dass y durch die Regressionsgleichung nicht in den ursprünglichen Einheiten geschätzt werden kann. Aber die Vorteile überwiegen, weil die Konstante a verschwindet: $a = \bar{z}_y - b\bar{z}_x = 0,$ weil $\bar{z}_y = \bar{z}_x = 0.$ Da $a = 0,$ so erhält man: $z_y = \beta_{yx}\, z_x$ oder $z_y = r_{xy}\, z_x.$

Die Stärke der Effekte der verschiedenen unabhängigen Variablen (Prädiktoren) lässt sich vergleichen, wenn die Variablen auf standardisierten Skalen gemessen werden.

3.1.3 Gleichungsansatz der multiplen Regression und Matrixschreibweise

In der multiplen Regressionsanalyse werden mehrere unabhängige Variablen x_i mit $i = 1, \ldots, k$ ($k \geq 2$) berücksichtigt. Der allgemeine multiple Regressionsansatz lautet:

$$\hat{y} = a + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k$$

Nach der Methode der kleinsten Quadrate wird der Fehler $f(a, b_1, \ldots, b_k) = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2$ als Funktion von $a, b_1, \ldots, b_k$ minimiert, um die beobachteten Werte y_i durch die Schätzwerte $\hat{y}_i$ möglichst gut zu approximieren. Dazu wird die Funktion $f(a, b_1, \ldots, b_k)$ jeweils nach den Parametern $a, b_1, \ldots, b_k$ abgeleitet, und die partiellen Ableitungen werden jeweils gleich Null gesetzt. Diese Methode liefert für die Koeffizienten $a; b_1, \ldots, b_k$ jeweils ein Gleichungssystem der folgenden Art, wobei k die Anzahl der Prädiktoren ist:

k = 1:
$$a + b_1 \bar{x}_1 = \bar{y}$$
$$b_1 s_{x_1}^2 = s_{y, x_1}$$

k = 2:
$$a + b_1 \bar{x}_1 + b_2 \bar{x}_2 = \bar{y}$$
$$b_1 s_{x_1}^2 + b_2 s_{x_1, x_2} = s_{y, x_1}$$
$$b_1 s_{x_2, x_1} + b_2 s_{x_2}^2 = s_{y, x_2}$$

k = 3:
$$a + b_1 \bar{x}_1 + b_2 \bar{x}_2 + b_3 \bar{x}_3 = \bar{y}$$
$$b_1 s_{x_1}^2 + b_2 s_{x_1, x_2} + b_3 s_{x_1, x_3} = s_{y, x_1}$$
$$b_1 s_{x_2, x_1} + b_2 s_{x_2}^2 + b_3 s_{x_2, x_3} = s_{y, x_2}$$
$$b_1 s_{x_3, x_1} + b_2 s_{x_3, x_2} + b_3 s_{x_3}^2 = s_{y, x_3}$$

Für den Fall k = n soll sowohl der Gleichungsansatz als auch die Matrixschreibweise ausgearbeitet werden:

$$\text{an} + b_1 \sum_{i=1}^{n} x_{1i} + b_2 \sum_{i=l}^{n} x_{2i} + \ldots + b_k \sum_{i=1}^{n} x_{ki} = \sum_{i=1}^{n} y_i$$

$$a\sum_{i=1}^{n} x_{1i} + b_1 \sum x_{1i}^2 + b_2 \sum_{i=1}^{n} x_{2i}x_{1i} + \ldots + b_k \sum_{i=1}^{n} x_{1i} \cdot x_k = \sum_{i=1}^{n} x_{1i}y_i$$

$$a\sum_{i=1}^{n} x_{2i} + b_1 \sum_{i=1}^{n} x_{1i}x_{2i} + b_2 \sum_{i=1}^{n} x_{2i}^2 + \ldots + b_k \sum_{i=1}^{n} x_{2i} \cdot x_k = \sum_{i=1}^{n} x_{2i}y_i$$

$$\vdots \qquad \vdots \qquad \vdots \qquad \vdots \qquad \vdots$$

$$a\sum_{i=1}^{n} x_k + b_1 \sum_{i=1}^{n} x_{1i} \cdot x_k + b_2 \sum_{i=1}^{n} x_{2i} \cdot x_k + \ldots + b_k \sum_{i=1}^{n} x_{ki}^2 = \sum_{i=1}^{n} x_k y_i$$

Zur Bestimmung der Regressionskoeffizienten nutzt man vereinfachend die **Matrix**-Schreibweise:

$$\begin{bmatrix} n & \sum x_{1i} & \sum x_{2i} & \cdots & \sum x_{ki} \\ \sum x_{1i} & \sum x_{1i}^2 & \sum x_{2i}x_{1i} & \cdots & \sum x_{1i}x_{ki} \\ \sum x_{2i} & \sum x_{1i}x_{2i} & \sum x_{2i}^2 & \cdots & \sum x_{2i}x_{ki} \\ \vdots & \vdots & \vdots & & \vdots \\ \sum x_{ki} & \sum x_{1i}x_{ki} & \sum x_{2i}x_k & \cdots & \sum x_{ki}^2 \end{bmatrix} \begin{bmatrix} a \\ b_1 \\ b_2 \\ \vdots \\ b_k \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum x_{1i}y_i \\ \sum x_{2i}y_i \\ \vdots \\ \sum x_k y_i \end{bmatrix}$$

Die erste Matrix lässt sich mit Hilfe der Datenmatrix X darstellen, indem man die *transponierte Matrix* von X, also X′, die aus X durch Vertauschen von Zeilen und Spalten hervorgeht, mit der Datenmatrix X multipliziert:

$$X'X = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_{11} & x_{12} & \cdots & x_{1n} \\ \vdots & \vdots & & \vdots \\ x_{k1} & x_{k2} & \cdots & x_{kn} \end{bmatrix} \begin{bmatrix} 1 & x_{11} & \cdots & x_{k1} \\ 1 & x_{12} & \cdots & x_{k2} \\ \vdots & \vdots & & \vdots \\ 1 & x_{1n} & \cdots & x_{kn} \end{bmatrix}$$

Die erste Zeile bzw. Spalte wird jeweils gleich 1 gesetzt, weil dem konstanten Glied kein Faktor zugeordnet werden kann.

$$X'X = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_{11} & x_{12} & \cdots & x_{1n} \\ \vdots & \vdots & & \vdots \\ x_{k1} & x_{k2} & \cdots & x_{kn} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

Das Gleichungssystem kann nun ganz in Matrixschreibweise dargestellt werden:

$$\boxed{(X'X) \cdot b = X'y}$$

Hierbei ist b der Spaltenvektor der Regressionskoeffizienten: b′ = (a, b_1, ..., b_k)

Multipliziert man beide Seiten der Matrizengleichung mit der Inversen von X′X, also jener Matrix, für die gilt: $(X'X)^{-1}\,(X'X) = I$ (I = neutrales Element der Matrixmultiplikation = *Einheitsmatrix*), kann man die Gleichung nach dem gesuchten Spaltenvektor b auflösen.

$$(X'X)^{-1}\,(X'X)\,b = (X'X)^{-1}\,X'y$$
$$b = (X'X)^{-1}\,X'y$$

Das symmetrisch aufgebaute Gleichungssystem zur Bestimmung der Regressionskoeffizienten vereinfacht sich, wenn man die Variablen zuerst standardisiert:

$$\tilde{a} + \tilde{b}_1 \underbrace{\bar{z}_{x_1}}_{0} + \cdots + \tilde{b}_k \underbrace{\bar{z}_{x_k}}_{0} = \underbrace{\bar{z}_y}_{0}$$

Somit wird $\tilde{a} = 0$, weshalb jeweils die erste Gleichung entfällt.

Die allgemeine Gleichung $b_1 s_{x_1,x_i} + b_2 s_{x_2,x_i} + \ldots + b_i s_{x_i}^2 + \ldots + b_k s_{x_k,x_i} = s_{y,x_i}$ lässt sich nach Multiplikation mit $\dfrac{1}{s_y \cdot s_{x_i}}$ umformen zu:

$$\underbrace{b_1 \frac{s_{x_1}}{s_y}}_{\beta_1} \underbrace{\frac{s_{x_1,x_i}}{s_{x_1} \cdot s_{x_i}}}_{r_{x_1,x_i}} + \underbrace{b_2 \frac{s_{x_2}}{s_y}}_{\beta_2} \underbrace{\frac{s_{x_2,x_i}}{s_{x_2} \cdot s_{x_i}}}_{r_{x_2,x_i}} + \ldots + \underbrace{b_i \frac{s_{x_i}}{s_y}}_{\beta_i} + \ldots + \underbrace{b_k \frac{s_{x_k}}{s_y}}_{\beta_k} \underbrace{\frac{s_{x_k,x_i}}{s_{x_k} \cdot s_{x_i}}}_{r_{x_k,x_i}} = \underbrace{\frac{s_{y,x_i}}{s_y \cdot s_{x_i}}}_{r_{y,x_i}}$$

(Allgemein lautet der Zusammenhang von standardisierten und anstandardisierten Regressionskoeffizienten: $\beta_{yx} = b_{yx} \cdot \dfrac{s_x}{s_y}$)

Um die Bezeichnungen zu vereinfachen, sei $r_{ij} := r_{x_i,x_j}, r_{y_i} := r_{y,x_i}$

Mit diesen Bezeichnungen lauten die Gleichungssysteme für die *standardisierten* Variablen (Ansatz: $\hat{y} = \beta_1 x_1 + \ldots + \beta_k x_k$):

k = 1: $\beta_1 = r_{y_1}$

k = 2:
$$\beta_1 + \beta_2\, r_{12} = r_{y_1}$$
$$\beta_1\, r_{21} + \beta_2 = r_{y_2}$$

k = 3:
$$\beta_1 + \beta_2\, r_{12} + \beta_3\, r_{13} = r_{y_1}$$
$$\beta_1\, r_{21} + \beta_2 + \beta_3\, r_{23} = r_{y_2}$$
$$\beta_1\, r_{31} + \beta_2\, r_{32} + \beta_3 = r_{y_3}$$

In Matrixschreibweise:

$$\mathrm{R} := \begin{pmatrix} 1 & & r_{ji} \\ & \ddots & \\ r_{ij} & & 1 \end{pmatrix}_{(k,k)} \qquad \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} \qquad \mathrm{r} = \begin{pmatrix} r_{y_1} \\ \vdots \\ r_{y_k} \end{pmatrix}$$

Also: $\underset{(k,k)}{R} \cdot \underset{(k,l)}{\beta} = \underset{(k,l)}{r}$ oder: $\beta = R^{-1} \cdot r$, wobei R^{-1} die inverse Matrix von R ist.

Die rechnerische Lösung der multiplen Regression besteht also darin, die **Korrelationsmatrix R** der unabhängigen Variablen zu invertieren bzw. auf die andere Seite zu bringen und mit dem Vektor der Korrelationen von Prädikatioren und abhängigen Variablen zu multiplizieren.

Beweis:

$$X := (x_1, ..., x_k)_{(n,k)}, R := \frac{X'X}{n}, y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, r := \frac{X' \cdot y}{n} \beta := \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}$$

$$\hat{y} = X\beta$$

$$\| y - \hat{y} \|^2 = (y - \hat{y})' (y - \hat{y}) = (y - X\beta)' (y - X\beta) = y'y + \beta'X'X\beta - 2\beta' y \qquad (*)$$

$$\frac{\partial}{\partial \beta'} = 2X' X\beta - 2X'y = 0$$

Für jede symmetrische Matrix A gilt (vgl. van de Geer 1971:56):

$$\frac{\partial (z' Az)}{\partial z'} = 2 Az$$

$$\Rightarrow \quad X'X\beta = X'y : n$$

$$R\beta = r$$

$$\Rightarrow \quad \beta = R^{-1} \cdot r$$

R^{-1} = inverse Matrix zu R ($R \cdot R^{-1}$ = I = Einselement der Matrixmultiplikation)

3.1.4 Multipler Korrelationseffekt R

Der multiple Korrelationskoeffizient $R_{y; x_1, ..., x_k}$ ist definiert als der einfache Korrelationskoeffizient $r_{y,\hat{y}}$, wobei $\hat{y}$ die Regressionsschätzung von y aufgrund von x_1, ..., x_k ist. Im Falle k = 1 handelt es sich also um den einfachen Korrelationskoeffizienten $r_{y x_1}$, weil:

$$\left| r_{y,\hat{y}} \right| = \left| r_{y, a + bx_1} \right| = \left| r_{y, x_1} \right|$$

Die Streuungszerlegung lautet wie in der einfachen Regression, die Regressionsschätzung in der multiplen Regression basiert aber auf k Prädiktoren statt einem einzigen: $\hat{y} = a + b_1 x_1 + ... + b_k x_k$

$SS_y = SS_{reg} + SS_{res}$

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2$$

$$R^2_{y; x_1, ..., x_k} = \frac{SS_y - SS_{res}}{SS_y} = \frac{SS_{reg}}{SS_y}$$

SS_y = sum of squares total $= \sum (y_i - \bar{y})^2$

SS_{res} = sum of squares residual $= \sum (y_i - \hat{y}_i)^2$

SS_{reg} = sum of squares regression $= \sum (\hat{y}_i - \bar{y})^2$

Wie im Falle der einfachen Regression gilt also die Interpretation:

$$R^2_{y; x_1, ..., x_k} = \frac{\text{durch die Regression erklärte Varianz}}{\text{Gesamtvarianz}}$$

Das Quadrat des multiplen Korrelationskoeffizienten $R^2_{y;x_1,\ldots,x_k}$ gibt also das Ausmaß wieder, in dem die abhängige Variable y durch die Regression auf die unabhängigen Variablen x_1, ..., x_k erklärt wird.

Der multiple Korrelationskoeffizient lässt sich nach der Gleichung (*) unter 3.1.3 berechnen:

$$\frac{\sum(y_i-\hat{y}_i)^2}{n}=\frac{y'y}{n}+r'R^{-1}\underbrace{\frac{X'X}{n}}_{\text{R}}R^{-1}r-2r'R^{-1}\underbrace{\frac{X'y}{n}}_{\text{r}}$$

Also: $$\frac{\sum(y_i-\hat{y}_i)^2}{n}=\frac{y'y}{n}+\underbrace{r'R^{-1}r}_{r'\beta} \qquad (**)$$

$$\frac{y'y}{n}=\frac{\sum(y_i-\hat{y}_i)^2}{n}+r'R^{-1}r$$

Dabei ist:

- $\frac{y'y}{n}$ = Varianz von y,
- $\frac{\sum(y_i-\hat{y}_i)^2}{n}$ = durch die Regression nicht erklärte Varianz und
- $r'R^{-1}r=r'\beta=$ durch die Regression erklärte Varianz und zugleich das Quadrat der multiplen Korrelationskoeffizienten. Dieses Ergebnis erhält man auch, wenn man benutzt:

$$\text{Multiple } R^2=s^2_{\hat{y}}=\frac{\hat{y}'\hat{y}}{n}=\frac{(X\beta)'(X\beta)}{n}=\beta'\left(\frac{X'X}{n}\right)\beta=\beta'r, \text{ weil } X'X/n=R.$$

Im Falle k = 1 wäre: $r'R^{-1}r=r_{y1}\cdot 1\cdot r_{y1}=r^2_{y1}$

In diesem Fall ist der multiple Korrelationskoeffizient gleich dem einfachen Korrelationskoeffizienten.

Sind die Variablen zentriert $(\bar{x}_i=0)$, so lässt sich der multiple Korrelationskoeffizient bei zwei Prädiktoren (k = 2) wie folgt berechnen:

$$R^2_{y;x_1x_2}=\frac{s^2_{\hat{y}}}{s^2_y}=b_1^2\frac{s^2_{x_1}}{s^2_y}+b_2^2\frac{s^2_{x_2}}{s^2_y}+2b_1b_2\frac{s_{x_1x_2}}{s^2_y}$$

Denn: $\hat{y}<0b_1x_1+b_2x_2 \; und \; E\left[(b_1x_1+b_2x_2)^2\right]=b_1^2s^2_{x_1}+b_2^2s^2_{x_2}+2b_1b_2s_{x_1x_2}$

Im Falle von standardisierten Variablen erhält man:

$$R^2_{y;x_1x_2}=\beta_1{}^2+\beta_2{}^2+2\beta_1\beta_2\,r_{x_1x_2}$$

Das Quadrat des multiplen Korrelationskoeffizienten lässt sich nach (**) auch folgendermaßen berechnen (mittels der Korrelationsmatrix R):

$$R^2_{y;x_1,\ldots,x_k} = r' R^{-1} r = r' \beta$$
$$= \sum_{i=1}^{k} \beta_i \, r_{y,x_i}$$
$$= \beta_1 \, r_{y,x_1} + \ldots + \beta_k \, r_{y,x_k}$$

3.1.5 Interpretation der Koeffizienten

Die Koeffizienten b_i geben den Einfluss der unabhängigen Variablen x_i (Prädiktoren) auf die abhängige Variable y (Kriterium) wieder, wenn der Einfluss der übrigen unabhängigen Variablen kontrolliert wird, was für den Fall k = 2 gezeigt werden soll:

$\hat{y} = a + b_1 x_1 + b_2 x_2$

So ist b_1 der einfache Regressionskoeffizient von y auf die Residualvariable $x_1 - \hat{x}_1$, wobei $\hat{x}_1 = \tilde{a} + \tilde{b} x_2$ die Regressionsschätzung von x_1 auf x_2 ist. (Die Variable x_2 hat keinen Einfluss mehr auf die Residualvariable $x_1 - \hat{x}_1$, weil der Einfluss von x_2 in $x_1 - \hat{x}_1$ eliminiert ist: $r_{x_2, x_1 - \hat{x}_1} = 0$.)

Beweis:

Dieser einfache Regressionskoeffizient b_1 ist gleich $\frac{s_{x_1-\hat{x}_1,y}}{s^2_{x_1-\hat{x}_1}}$.

$$s_{x_1-\hat{x}_1,y} = s_{x_1 y} - \tilde{b} s_{x_2 y} = s_{x_1 y} - \frac{s_{x_2 x_1} s_{x_2 y}}{s^2_{x_2}} \text{ wobei : } \tilde{b} = \frac{s_{x_2 x_1}}{s^2_{x_1}}$$

$$s^2_{x_1-\hat{x}_1} = s^2_{x_1} - 2\tilde{b} s_{x_1 x_2} + \tilde{b}^2 s^2_{x_2}$$
$$= s^2_{x_1} - \frac{s^2_{x_2 x_1}}{s^2_{x_2}}$$

Nach dem Gleichungssystem der multiplen Regression erhält man aus:

$$b_1 s^2_{x_1} + b_2 s_{x_1 x_2} = s_{y,x_2}$$
$$b_1 s_{x_2 x_1} + b_2 s^2_{x_2} = s_{y,x_2}$$

gerade: $b_1 = \dfrac{s_{y x_1} - \frac{s_{x_1 x_2} s_{y x_2}}{s^2_{x_2}}}{s^2_{x_1} - \frac{s^2_{x_1 x_2}}{s^2_{x_2}}}$. Dies ist das gleiche Ergebnis.

Entsprechend gibt b_2 den Einfluss von x_2 auf y an, wenn der Einfluss von x_1 auf x_2 kontrolliert ist.

Im Falle von standardisierten Variablen ergibt sich aus der Formel für den Regressionskoeffizienten:

$$\beta_1 = \frac{r_{yx_1} - r_{x_1x_2} r_{yx_2}}{1 - r^2_{x_1x_2}}$$

Die **Regressionskoeffizienten β_i** der standardisierten Variablen sind die *einfachen* Regressionskoeffizienten von y auf die bereinigten Prädiktoren $x_i - \hat{x}_i$, d.h. diese Koeffizienten sind um den Einfluss der anderen Prädiktoren auf x_i bereinigt.

Die Korrelationskoeffizienten von y mit diesem bereinigten Prädiktoren heißen **Part Correlations** oder auch **Semi-partial Correlations** (Bezeichnung z.B.: $r_{y(1.2)}$).

Da für die einfache Regression gilt: $r = b \cdot \frac{s_x}{s_y}$, erhält man:

$$r_{y(1.2)} = \beta_1 \cdot s_{x_1 - \hat{x}} = \beta_1 \sqrt{s^2_{x_1} - \frac{s^2_{x_2,x_1}}{s^2_{x_2}}}$$

$$r_{y(1.2)} = \frac{r_{y,x_1} - r_{x_1,x_2} r_{y,x_2}}{\sqrt{1 - r^2_{x_1,x_2}}}, \text{ wobei } \sqrt{s^2_{x_1} - \frac{s^2_{x_2x_1}}{s^2_{x_2}}} = \sqrt{1 - r^2_{x_1,x_2}} \text{ bei standardisierten Variablen.}$$

Das Quadrat des Korrelationskoeffizienten $r_{y,\,x_i - \hat{x}_i}$ gibt also an, welcher Anteil der Varianz von y durch den um den Einfluss der Prädiktoren $x_{j,\,j \neq i}$, bereinigten Prädiktor x_i erklärt wird.

Formal:

$$r^2_{y(1.2)} = \frac{r^2_{y,x_1} - 2 r_{y,x_1} r_{y,x_2} r_{x_1,x_2} + r^2_{y,x_2} r^2_{x_1,x_2}}{1 - r^2_{x_1,x_2}}$$

$$= \frac{r_{y,x_1}\left(r_{y,x_1} - r_{y,x_2} r_{x_1,x_2}\right) + r_{y,x_2}\left(r_{y,x_2} - r_{y,x_1} r_{x_1,x_2}\right) - r^2_{y,x_2}\left(1 - r^2_{x_1,x_2}\right)}{1 - r^2_{x_1,x_2}}$$

$$= r_{y,x_1}\beta_1 - r_{y,x_2}\beta_2 - r^2_{y,x_2}$$

$$= R^2_{y.12} - r^2_{y,x_2} \text{ (nach (*) aus Punkt 3.1.3)}$$

D. h. in diesem Fall: $x_1 - \hat{x}_1$ erklärt das, was noch nicht durch x_2 erklärt worden ist.

Während beim partiellen Korrelationskoeffizienten $r_{yx_1 \cdot x_2}$ der Einfluss von x_2 auf *beide* Variablen y und x kontrolliert wird, wird beim Part Correlation Coefficient nur der Einfluss der unabhängigen Variablen x_2 auf die unabhängigen Variable x_1 kontrolliert:

$r_{y,x_1 - \hat{x}_1(x_2)}$ vs. $r_{yx_1 \cdot x_2} = r_{y - \hat{y}(x_2),x_1 - \hat{x}_1(x_2)}$

Das Vorzeichen der Regressionskoeffizienten drückt die Richtung des Einflusses aus (positiver oder negativer Einfluss), die Größe der Regressionskoeffizienten die Stärke des Einflusses der unabhängigen Variablen auf die abhängige Variable, nachdem alle übrigen unabhängigen Variablen kontrolliert wurden.

Im Gegensatz zu den Faktoren der Faktorenanalyse sind die unabhängigen Variablen in der Regressionsanalyse i.a. nicht unabhängig voneinander. Die Beta-Koeffizienten sind deshalb i.a. nicht identisch mit den einfachen Korrelationskoeffizienten der abhängigen Variablen mit der unabhängigen Variablen, was bei untereinander unabhängigen Variablen x_i der Fall wäre. Das Quadrat des Beta-Gewichtes einer unabhängigen Variablen gibt nicht einfach den Anteil der durch die unabhängige Variable erklärten Varianz der abhängigen Variablen wieder. Die Quadrate der Beta-Gewichte können deshalb auch nicht einfach addiert werden, um den Anteil der insgesamt erklärten Varianz zu bestimmen, da sich die Erklärungen überschneiden können sowie Suppressor- und Distorter- Phänomene möglich sind.

Die Unterschiede unter den einfachen Korrelationskoeffizienten, zwischen den einfachen Korrelationskoeffizienten und den Beta-Koeffizienten sowie zwischen den Beta-Koeffizienten untereinander und vor und nach Einführen einer weiteren unabhängigen Variablen in die Regressionsgleichung geben Hinweise auf die Art der Überschneidungen etc. der Erklärungen. Eine präzisere Beschreibung der Überschneidungen etc. ist mit Hilfe von weiteren Koeffizienten möglich, wie in Punkt 3.1.7 dargestellt wird.

3.1.6 Schrittweise Regression

Eine multiple Regression einer abhängigen Variablen y auf unabhängige Variablen $x_1,...,x_k$ kann auch derart durchgeführt werden, dass nicht alle unabhängigen Variablen $x_1,...,x_k$ auf einmal (simultan) in die Regressionsgleichung eingeführt werden, sondern in der Reihenfolge, dass sie einen maximalen Anteil der jeweils noch nicht erklärten Varianz erklären.

Die erste unabhängige Variable wäre also die, welche am meisten Varianz der abhängigen Variablen erklärt, d. h. am höchsten mit y korreliert. Die nächste unabhängige Variable ist die, welche am meisten der jetzt noch verbleibenden Varianz von y erklärt $\left(x_j, j \neq i, \text{mit}: r_{y,(x_j \cdot x_i)} \text{ maximal}\right)$. Etc. Mit jeder neuen unabhängigen Variablen ändert sich dabei i. a. die ganze Regressionsgleichung.

Der Anteil der durch x_i erklärte Varianz ist r_y^2, x_i, die erklärte Varianz also: $s_y^2 \cdot r_{y,x_i}^2$.

Die nicht erklärte Varianz $s_y^2\left(1 - r_{y,x_i}^2\right)$ wird durch x_j zu einem Anteil $r_{y(x_j \cdot x_i)}^2$ erklärt, also wird $s_y^2\left(1 - r_{y,x_i}^2\right) r_{y(x_j \cdot x_i)}^2$ zusätzlich erklärt.

Insgesamt erklären die beiden nach der Bedeutung für y wichtigsten unabhängigen Variablen x_i und x_j also:

$s_y^2\left(r_{y,x_i}^2 + \left(1 - r_{y,x_i}^2\right) \cdot r_{y(x_j \cdot x_i)}^2\right)$

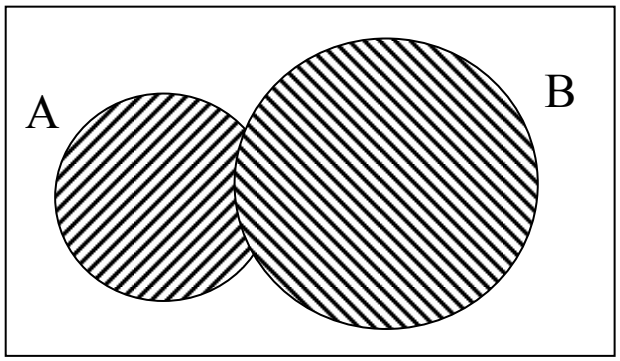
Etc.

3.1.7 Zuordnung der gesamten erklärten Varianz zu den Prädiktoren

3.1.7.1 Einführung in die Problemstellung

Das wesentliche Interpretationsproblem der multiplen Regression besteht wohl darin, wie die gesamte erklärte Varianz den verschiedenen Prädiktoren $x_1,...,x_k$ zugeordnet werden kann. Anschaulich lässt sich das Problem von Überschneidungen an Wahrscheinlichkeitsmaßen in Form von Flächenmaßen demonstrieren (Abbildung 3-2).

Abbildung 3-2: Veranschaulichung von Erklärungsbeiträgen in Form von Flächenmaßen



Im einfachsten Fall von zwei Mengen (sie entsprechen den Prädiktoren) gilt für das Maß der Vereinigungsmenge:

(1) $$P(A \cup B) = P(A - B) + P(B - A) + P(A \cap B)$$

oder

(2) $$P(A \cup B) = P(A) \quad + P(B) \quad - P(A \cap B)$$

Zu 1): Die schraffierten Flächen symbolisieren insgesamt die um Überschneidungen mit anderen Prädiktoren bereinigten Einflüsse, man erhält so nicht 100 % der Gesamterklärung. In der multiplen Regression erhält man, wie im Folgenden genauer erklärt wird, als entsprechende Größe:

$$\sum_{i=1}^{k} \text{Part Correlation}^2 \left(y, x_i\right) = \sum_{i=1}^{k} \left(R^2 - R^2(i)\right)$$

Zu 2): Addiert man die Flächenmaße der einzelnen Mengen auf, so erhält man i.a. wegen der Überschneidungen insgesamt über 100 % der Gesamtfläche. Bildlich gesprochen wird der Überschneidungsbereich doppelt gezählt. In der multiplen Regression entspricht dem die Größe:

$$\sum_{i=1}^{k} r_{y,x_i}^2$$

Insgesamt erhält man also eine teilweise Ausschöpfung „von unten“ und eine wegen der Überschneidungen i.a. „zu große“ Summe der Gesamterklärungskraft der einzelnen Prädiktoren. (Dies gilt für den Fall, dass nur Überschneidungen vorliegen. Im komplizierteren Fall gibt es auch noch Distorter-phänomene etc.) Ferner gibt es noch die rechnerische Lösung:

$$\text{Multiple } R^2 = \sum_{i=1}^{k} \beta_i \, r_{y,x_i}$$

Die Summanden ergeben hier insgesamt genau die Gesamterklärungskraft Multiple R^2, aber die Summanden können auch negativ sein und lassen sich deshalb nicht als Anteile der erklärten Varianz interpretieren, wohl aber als Resultat von korrelierten Effekten, wie im Kapitel 3.1.7.4 ausgeführt wird.

3.1.7.2 Charakterisierung der Koeffizienten mit Hilfe von Residuen

Die ganze Problematik lässt sich besonders transparent darstellen, wenn man die relevanten Koeffizienten in der multiplen Regression mit Hilfe von Residuen charakterisiert.

Wegen der größeren Übersichtlichkeit der Formeln werden standardisierte Variablen y, x_1, ... x_k vorausgesetzt, die Zurückrechnung auf nicht standardisierte Variablen ist ja immer möglich. Die geometrische Vorstellung der **Orthogonalität** von Dimensionen wird auch als **Unabhängigkeit** bezeichnet. Sie lässt sich algebraisch behandeln durch das innere Produkt (oder Skalarprodukt) $\langle a,b\rangle := \sum_{i=1}^{n} a_i b_i$ für 2 beliebige Vektoren $a=(a_1,...,a_n), b=(b_1,...,b_n)$. Die Kovarianz von standardisierten Variablen a, b steht in einem einfachen Zusammenhang mit dem inneren Produkt: $s_{a,b} = \frac{\langle a,b\rangle}{n}$ (Die Varianz $s_{a,a}$ wird wie üblich als s_a^2 bezeichnet.)

Die Lösung des Regressionsproblems besteht darin, $\sum_{j=1}^{n}(y_j - \hat{y}_j)^2$ (wobei $\hat{y} = \sum_{i=1}^{k} \beta_i x_i$) als Funktion der Koeffizienten β_i zu minimieren. Dies erfordert bekanntlich die Auflösung des Gleichungssystems $R \cdot \beta = r$, wobei R der Korrelationsmatrix der Prädiktoren, r den Spaltenvektor $\begin{pmatrix} r_y, x_1 \\ \vdots \\ r_y, x_k \end{pmatrix}$ und β den Koeffizientenvektor $\begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}$ bezeichnen.

Ausführlicher formuliert (für i = 1, ..., k):

$$r_{y,x_i} = \sum_{j=1}^{k} r_{x_i,x_j}\ \beta_j$$

Unter Verwendung des Skalarprodukts erhält man:

$\langle y, x_i\rangle = \langle \sum_{j=1}^{k} x_j \beta_j, x_i\rangle = \langle \hat{y}, x_i\rangle$ oder:

$\langle y - \hat{y}, x_i\rangle = 0$ (für i=1, ..., k)

Das Residuum $y - \hat{y}$ ist orthogonal zu allen Prädiktoren x_1, ..., x_k, korreliert also insbesondere auch nicht mehr mit ihnen. Der Grund dafür besteht darin, dass die gesamte zur Schätzung von y geeignete Information der Prädiktoren in $\hat{y}$ eingegangen ist, also im Residuum nichts mehr verbleibt, was noch einen Zusammenhang zu den Prädiktoren aufweist.

Da das Argument bezüglich der Orthogonalität und den Lösungsgleichungen für die Koeffizienten umkehrbar ist, wird die Regressionslösung $\hat{y}$ also genau dadurch charakterisiert, dass das Residuum $y - \hat{y}$ zu allen Prädiktoren orthogonal ist.

Da das Residuum $y - \hat{y}$ orthogonal ist zu allen Prädiktoren, ist es auch orthogonal zu jeder Linearkombination von Prädiktoren, also auch zu $\hat{y}$, d.h.: $\langle y - \hat{y}, \hat{y} \rangle = 0$ *oder* $\langle y, \hat{y} \rangle = \langle \hat{y}, \hat{y} \rangle$.[6]

Die gesamte Erklärungskraft der multiplen Regression, Multiple R^2, lässt sich also folgendermaßen charakterisieren:

$$\text{Multiple } R^2 = s^2_{\hat{y}} = \frac{\langle \hat{y}, \hat{y} \rangle}{n} = \frac{\langle y, \hat{y} \rangle}{n} = s_{y,\hat{y}}$$

Oder:

$$\text{Multiple } R^2 = \frac{\langle \hat{y}, \hat{y} \rangle}{n} = \sum_{i=1}^{k} \frac{\langle y, \beta_i x_i \rangle}{n} = \sum \beta_i r_{y,x_i}$$

Dies ist die rechnerische Lösung der Zerlegung der Erklärungskraft.
Als Varianz des Residuums erhält man den Anteil der durch die Regression nicht erklärten Varianz von y:

$$s^2_{y-\hat{y}} = \frac{\langle y - \hat{y}, y - \hat{y} \rangle}{n} = \frac{\langle y, y - \hat{y} \rangle}{n} = 1 - R^2$$

Im folgenden möge $\hat{x}_i$ abkürzend die Regressionsschätzung von x_i auf die übrigen Prädiktoren $x_j, j \neq i$, bezeichnen. Also ist $x_i - \hat{x}_i$ orthogonal zu allen $x_j, j \neq i$. Das Residuum $y - \hat{y}$ ist orthogonal zu $x_i - \hat{x}_i$, da letzteres nur eine Linearkombination der Prädiktoren ist. Also erhält man:
$\langle y, x_i - \hat{x}_i \rangle = \langle \hat{y}, x_i - \hat{x}_i \rangle = \beta_i \langle x_i, x_i - \hat{x}_i \rangle = \beta_i \langle x_i - \hat{x}_i, x_i - \hat{x}_i \rangle$

Somit lässt sich der **Regressionskoeffizient β_i** *mit Hilfe von Residuen* charakterisieren:

$$\beta_i = \frac{s_{y,x_i - \hat{x}_i}}{s_{x_i, x_i - \hat{x}_i}} = \frac{s_{y,x_i - \hat{x}_i}}{s^2_{x_i - \hat{x}_i}} \qquad \text{(analog zu } b_{yx} = \frac{s_{yx}}{s^2_x} \text{ in der einfachen Regression)}$$

Aus letzterem folgt, dass β_i der *einfache* Regressionskoeffizient der Regression von y auf den um die Einflüsse der übrigen Prädiktoren bereinigten Prädiktor $x_i - \hat{x}_i$ ist.

Die einfachste Interpretation des Regressionskoeffizienten b_i scheint darin zu bestehen, dass er die geschätzte Differenz der abhängigen Variablen angibt, wenn die unabhängige Variable x_i um eine Einheit geändert wird und alle übrigen unabhängigen Variablen x_j ($j \neq i$) konstant gehalten werden. In diesem Sinne gibt b_i (bzw. β_i bei standardisierten Variablen) den relativen Einfluss des Prädiktors x_i auf die abhängige Variable y wieder.

---

[6] Die Regressionsschätzung ist dadurch charakterisiert, dass der Abstand $\| y - \hat{y} \|$ minimiert wird. Aber:

$$\frac{\|y - \hat{y}\|^2}{n} = \frac{1 - \langle \hat{y}, \hat{y} \rangle}{n} \quad \text{Ferner: } r_{y,\hat{y}} \frac{\frac{\langle y, \hat{y} \rangle}{n}}{\sqrt{\frac{\langle \hat{y}, \hat{y} \rangle}{n}}} = \sqrt{\frac{\langle \hat{y}, \hat{y} \rangle}{n}}$$

Eine äquivalente Bedingung für die Regressionsschätzung ist also, dass die Korrelation von y und $\hat{y}$ maximiert wird.

Für $y' = a + b_1 x_1 + ... + b_k x_k$ gilt (mit Δ als Symbol für die Differenz):

$$\Delta Y' = a \underbrace{(1-1)}_{0} + b_1 \underbrace{\Delta x_1}_{0} + ... + b_{i-1} \underbrace{\Delta x_{i-1}}_{0} + b_i \underbrace{\Delta x_i}_{1} + b_{i+1} \underbrace{\Delta x_{i+1}}_{0} + ... b_k \underbrace{\Delta x_k}_{0}$$

Wie Küchler (1979: 45) richtig bemerkt, ist diese Interpretation insofern nicht angemessen, als aus einer Änderung in x_i um eine Einheit in der Regel wegen der möglichen Korrelation der unabhängigen Variablen untereinander sich auch Änderungen in den anderen unabhängigen Variablen ergeben.

Da die Beta-Koeffizienten jedoch einfache Regressionskoeffizienten von y auf $x_i - \hat{x}_i$ sind, kann man schreiben: $\Delta y = \beta_i \, \Delta (x_i - \hat{x}_i)$.

Das Problem der Interkorrelation von Prädiktoren entfällt hier, weil es in der einfachen Regression nur einen Prädiktor gibt. Man kann nun sagen, dass eine Änderung des bereinigten Prädiktors $x_i - \hat{x}_i$ um eine Einheit (d.h. $\Delta(x_i - \hat{x}_i) = 1$) eine Änderung der Größe β_i in y bewirkt.
β_i spiegelt also den Einfluss des bereinigten Prädiktors wieder.
Der Part Correlation Coefficient $r_{y, x_i - \hat{x}_i}$ steht in einem einfachen Zusammenhang mit dem β-Koeffizienten, weil $r_{a,b} = \frac{s_{a,b}}{s_a s_b}$ und $s_y = 1$:

$$\beta_i = \frac{r_{y, x_i - \hat{x}_i}}{s_{x_i - \hat{x}_i}}$$

Mit $\hat{y}_{(i)}$ wird hier die Regressionsschätzung von y auf $\{x_1, ... x_k\} - \{x_i\}$ bezeichnet, d.h. auf alle Prädiktoren x_j, ($j \neq i$). Der partielle Korrelationskoeffizient $r_{y, x_i \cdot x_1, ..., x_{i-1}, x_{i+1}, ..., x_k}$ zwischen y und x_i unter Kontrolle aller Prädiktoren x_j, ($j \neq i$), lässt sich dann auch schreiben als: $r_{y - \hat{y}(i), x_i - \hat{x}_i}$. Im Unterschied zur Semi-partiellen bzw. Part Correlation ist hier auch der Einfluss der übrigen Prädiktoren auf die Kriteriumvariable herausgerechnet. Den Zusammenhang zwischen dem β-Koeffizienten und dem partiellen Korrelationskoeffizienten erhält man, wenn man den Zusammenhang verwendet, dass das Residuum $x_i - \hat{x}_i$ zu allen Linearkombinationen von x_j, ($j \neq i$), orthogonal ist, also auch zu $\hat{y}_{(i)}$.

$$\beta_i = \frac{s_{y - \hat{y}(i), x_i - \hat{x}_i}}{s^2_{x_i - \hat{x}_i}} = r_{y - \hat{y}(i), x_i - \hat{x}_i} \cdot \frac{s_{y - \hat{y}(i)}}{s_{x_i - \hat{x}_i}}$$

$$= r_{y, x_i \cdot x_1, ..., x_{i-1}, x_{i+1}, ..., x_k} \cdot \frac{s_{y - \hat{y}(i)}}{s_{x_i - \hat{x}_i}}$$

Part Correlation und **Partial Correlation** stehen deshalb in dem Zusammenhang:

$$r_{y, x_i - \hat{x}_i} = r_{y, x_i \cdot x_1, ..., x_{i-1}, x_{i+1}, ..., x_k} \cdot s_{y - \hat{y}(i)}$$

Aus den Orthogonalitätseigenschaften der Residuen folgt, dass die verwendeten drei Regressionsschätzungen in folgendem Zusammenhang stehen:

$$\hat{y} = \hat{y}_{(i)} + \beta_i (x_i - \hat{x}_i)$$

Γ

Zum Beweis kann man zeigen, dass das innere Produkt des Differenzvektors beider Seiten mit sich selbst gleich 0 ist:

$\langle \hat{y} - \hat{y}_{(i)} - \beta_i (x_i - \hat{x}_i), \hat{y} - (\hat{y}_{(i)} + \beta_i (x_i - \hat{x}_i)) \rangle$ $\quad$ (Für alle j: $\langle \hat{y}, x_j \rangle = \langle y, x_j \rangle$)

$= \langle (y - \hat{y}_{(i)}) - \beta_i (x_i - \hat{x}_i), \hat{y} - (\hat{y}_{(i)} + \beta_i (x_i - \hat{x}_i)) \rangle$

($y - \hat{y}_{(i)}$ ist orthogonal zu allen $x_j, j \neq i.$

$x_i - \hat{x}_i$ ist orthogonal zu allen $x_j, j \neq i.$)

$= \langle (y - \hat{y}_{(i)}) - \beta_i (x_i - \hat{x}_i), \underbrace{\beta_i x_i - \beta_i x_i}_{0} \rangle = 0$

Hierbei lässt sich β_i ausführlicher beschreiben als:

$\beta_{y, x_i . x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_k}$

L

Bezeichnet $R^2_{(i)}$ den Anteil der Varianz von y, der durch die Prädiktoren $x_j, (j \neq i)$, erklärt wird, so gilt entsprechend für die Berechnung von Multiple R^2:

$$R^2_{(i)} = \frac{\langle y, \hat{y}_{(i)} \rangle}{n}$$

Aus dem Zusammenhang $\hat{y} = \hat{y}_{(i)} + \beta_i (x_i - \hat{x}_i)$ folgt:

$$\text{Multiple } R^2 = \frac{\langle y, \hat{y} \rangle}{n} = \frac{\langle y, \hat{y}_{(i)} \rangle}{n} + \beta_i \frac{\langle y, x_i - \hat{x}_i \rangle}{n}$$

$$= R^2_{(i)} + r_{y, x_i - \hat{x}_i} \frac{s_{y, x_i - \hat{x}_i}}{s_{x_i - \hat{x}_i}}$$

$$= R^2_{(i)} + r^2_{y, x_i - \hat{x}_i}$$

Das Quadrat des Part Correlation Coefficient gibt also genau den Anteil der Varianz von y an, der durch Prädiktor x_i zusätzlich zu den anderen Prädiktoren $x_j, (j \neq i)$, erklärbar ist. In anderen Worten misst das Quadrat von $r_{y, x_i - \hat{x}_i}$ also den Erklärungszuwachs, den man erhält, wenn man den Prädiktor x_i zusätzlich zu den anderen Prädiktoren $x_j, (j \neq i)$, für die Regressionsschätzung verwendet:

$r^2_{y, x_i - \hat{x}_i} = R^2 - R^2_{(i)}$ $\quad$ (**bereinigte Erklärungskraft** des i-ten Prädiktors)

Daraus folgt nun, dass das Quadrat des partiellen Korrelationskoeffizienten $r_{y - \hat{y}_{(i)}, x_i - \hat{x}_i}$ die proportionale Reduktion des Fehlers bei der Schätzung von y durch Hinzufügen des Prädiktors x_i zu den Prädiktoren $x_j, (j \neq i)$, wiedergibt:

$$r^2_{y - \hat{y}_{(i)}, x_i - \hat{x}_i} = \frac{r^2_{y, x_i - \hat{x}_i}}{s^2_{y - \hat{y}_{(i)}}} = \frac{R^2 - R^2_{(i)}}{1 - R^2_{(i)}} = \frac{(1 - R^2_{(i)}) - (1 - R^2)}{1 - R^2_{(i)}}$$

Wendet man den Zusammenhang $\hat{y} = \hat{y}_{(i)} + \beta_i (x_i - \hat{x}_i)$ rekursiv an, so erhält man (für i = 2, ..., k):

$$\hat{y}(x_1, \ldots, x_i) = \hat{y}(x_1, \ldots, x_{i-1}) + \beta_{y, x_i . x_1, \ldots, x_{i-1}} \cdot (x_i - \hat{x}_i(x_1, \ldots, x_{i-1}))$$

Deshalb lässt sich der multiple Regressionsansatz
$\hat{y} = \beta_{y,x_1} \cdot x_1 + \beta_{y,x_2 .x_1,x_3,...,x_k} \cdot x_2 + ... + \beta_{y,x_k .x_1,...,x_{k-1}} \cdot x_k$ auch darstellen als schrittweise Erweiterung der einfachen Regression mit Hilfe von um den Einfluss der vorangehenden Prädiktoren bereinigten unabhängigen Variablen:
$\hat{y} = \beta_{y,x_1} \cdot x_1 + \beta_{y,x_2 .x_1} \cdot \left(x_2 - \hat{x}_2(x_1)\right) + ... + \beta_{y,x_k .x_1,...,x_{k-1}} \cdot \left(x_k - \hat{x}_k(x_1,...,x_{k-1})\right)$

Hieraus folgt auch eine hierarchische Zerlegung von Multiple R^2:

(1) $$R^2 = \frac{\langle y, \hat{y} \rangle}{n} = r^2_{y,x_1} + r^2_{y,x_2 - \hat{x}_2(x_1)} + ... + r_{y,x_k - \hat{x}_k(x_1,...,x_{k-1})}$$
$$= r^2_{y,x_1} + r^2_{y-\hat{y}(x_1),x_2-\hat{x}_2(x_1)}\left(1 - R^2 y, x_1\right) + ... + r^2_{y-\hat{y}(x_1,...,x_{k-1}),x_k-\hat{x}_k(x_1,...,x_{k-1})}\left(1 - R^2_{y,x_1,...,x_{k-1}}\right)$$

Dies folgt, weil: $\beta_{y,x_1} s_{y,x_1} = r^2_{y,x_1}$ und für i = 2, ..., k:

$$\beta_{y,x_i \cdot x_1,...,x_{i-1}} s_{y,x_i - \hat{x}_i(x_1,...,x_{i-1})} = \frac{r_{y,x_i - \hat{x}_i(x_1,...,x_{i-1})}}{s_{x_i - \hat{x}_i(x_1,...,x_{i-1})}} s_{y,x_i - \hat{x}_i(x_1,...,x_{i-1})}$$
$$= r^2_{y,x_i - \hat{x}_i(x_1,...,x_{i-1})} = r^2_{y-\hat{y}(x_1,...,x_{i-1}),x_i - \hat{x}_i(x_1,...,x_{i-1})} s^2_{y-\hat{y}(x_1,...,x_{i-1})} \text{ und}$$
$$s^2_{y-\hat{y}(x_1,...,x_{i-1})} = 1 - R^2_{y,x_1,...,x_{i-1}}$$

Werden die Prädiktoren in der Reihenfolge eingeführt, dass man zunächst den Prädiktor mit maximaler Erklärungskraft r^2_{y,x_i} berücksichtigt und anschließend jeweils den Prädiktor x_j, der von der noch zu erklärenden Varianz $1 - R^2_{y,x_1,...,x_{j-1}}$ den maximalen Anteil $r^2_{y-\hat{y}(x_1,...,x_{j-1}),x_j-\hat{x}_j(x_1,...,x_{j-1})}$ erklärt, so handelt es sich um die **schrittweise multiple Regression**, die die Prädiktoren nach ihrer zusätzlichen Erklärungskraft ordnet. Der Anteil der bei jedem Schritt zusätzlich bzw. insgesamt erklärten Varianz lässt sich aus Formel (1) ablesen, wobei die Prädiktoren in der Reihenfolge der abnehmenden zusätzlichen Erklärungskraft durchnummeriert sind.

3.1.7.3 Zwei Zerlegungen von Multiple R^2

Das in 3.1.7.1 formulierte Problem der Zerlegung von Multiple R^2 im Fall von Überschneidungen lässt sich mit dem dargestellten Instrumentarium leicht lösen. Die Zerlegung der gesamten Erklärungskraft „von unten“ und „von oben“ lassen sich berechnen als:

(1) $$R^2 = \sum_{i=1}^{k} r^2_{y,x_i - \hat{x}_i} + \sum_{i=1}^{k} \beta_i s_{\hat{y}(i),x_i}$$

(2) $$R^2 = \sum_{i=1}^{k} r^2_{y,x_i} - \sum_{i=1}^{k} r_{y,x_i} \sum_{\substack{j=1 \\ j \neq i}}^{k} r_{x_i,x_j} \beta_j$$

Die beiden wichtigsten Koeffizienten bei dem Problem der Zuordnung der erklärten Varianz zu den einzelnen Prädiktoren sind der Part Correlation Coefficient $r_{y,x_i - \hat{x}_i}$ und der einfache Korrelationskoeffizient r_{y,x_i}. Das Quadrat des Part Correlation Coefficient $r_{y,x_i - \hat{x}_i}$ gibt den Zuwachs in der Erklärung von y wieder, den man erzielt, wenn man den Prädiktor x_i zusätzlich zu den übrigen Prädiktoren $x_j, (j \neq i)$, berücksichtigt.

Es gilt: $$\boxed{r^2_{y,x_i - \hat{x}_i} = R^2 - R^2_{(i)}}$$

Summiert man alle quadrierten Part Correlation Coefficients, so erhält man nicht genau Multiple R^2, weil alle Überschneidungen nicht berücksichtigt werden. Die verbleibende Differenz zu Multiple R^2 ergibt sich aus Formel (1).

Das Quadrat des einfachen Korrelationskoeffizienten r_{y,x_i} gibt wieder, welcher Anteil der Varianz von y durch den Prädiktor x_i insgesamt erklärt wird. Addiert man alle diese quadrierten einfachen Korrelationskoeffizienten auf, so erhält man wegen der mehrfachen Berücksichtigung der Überschneidungen als Summe nicht genau Multiple R^2. Die Differenz wird in der o.g. Formel (2) charakterisiert.

Gibt es eine Hierarchie der Prädiktoren x_1, ..., x_k, so ist noch die entsprechende Zerlegung von Multiple R^2 relevant:

$$R^2 = r^2_{y,x_1} + r^2_{y,x_2-\hat{x}_2(x_1)} + ... + r^2_{y,x_k-\hat{x}_k(x_1,...,x_{k-1})}$$

Die quadrierten Part Correlation Coefficients geben genau an, was die um den Einfluss der in der Hierarchie voranstehenden Prädiktoren bereinigten unabhängigen Variablen zusätzlich zu dem erklären, was die vorher berücksichtigten Prädiktoren allein erklären. Sie messen also den Zuwachs in der Erklärungskraft.

Die Verwendung der Residuen erhöht die Übersichtlichkeit, insofern der Zusammenhang zwischen den verschiedenen Koeffizienten besonders transparent wird. Zur Berechnung der Koeffizienten braucht man jedoch weiterhin Determinanten, deren Benutzung zur Lösung des Regressionsproblems üblich ist. Die Koeffizienten $\beta_i, r_{y,x_i-\hat{x}_i} = r_{y-\hat{y}(i),x_i-\hat{x}_1}$, Multiple R^2 etc. lassen sich alle mit Hilfe von Determinanten ausdrücken. Dies ist im Anhang ausgeführt. Die Darstellung mit Residuen scheint mir übersichtlicher und eleganter.

3.1.7.4 Darstellung der erklärten Varianz durch Kovarianzen und Effekte

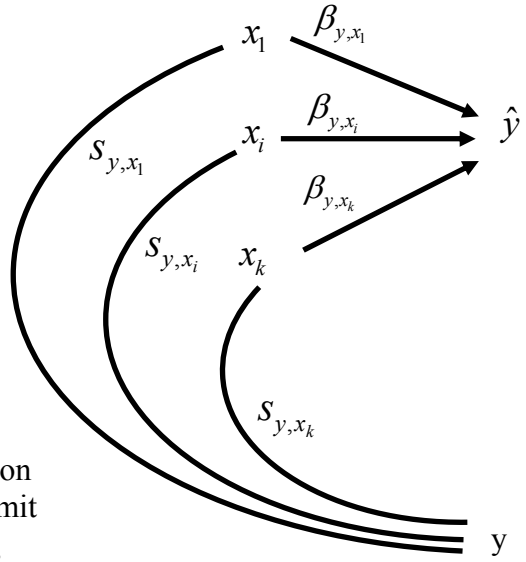
Multiple R^2 = $s^2_{\hat{y}} / s^2_y = s^2_{\hat{y}}$ (für standardisiertes y)

= $s_{\hat{y},\hat{y}} = s_{\hat{y},y}$ (weil $y - \hat{y}$ orthogonal ist zu $\hat{y}$)

Eine neue Interpretation bestände darin, Multiple R^2 durch „erklärte Kovarianz“, die auch negativ sein kann, auf die einzelnen Prädiktoren zurückzuführen.

Multiple $R^2 = s_{y,\hat{y}} = \sum \beta_{y,x_i} s_{y,x_i}$

Abbildung 3-3:



Die Kovarianz zwischen y und $\hat{y}$ (d.h. R^2) ist das Resultat der Kovariation von y und den Prädiktoren x_i, welche mit der Gewichtung β_{y,x_i} (direkter Effekt, bereinigter Effekt) in die lineare Modellschätzung $\hat{y}$ einfließen.

3.1.8 Interaktion in der Regression

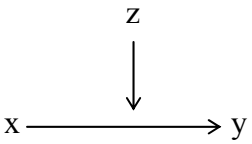
Die Regressionsschätzung $\hat{y} = \beta_1 x_1 + \beta_2 x_2$ kann man dadurch erweitern, dass Effekte von Kombinationen (nach Blalock/Blalock 1968) zugelassen werden, in diesem einfachsten Fall also: $\hat{y} = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$. Man würde eine neue Variable definieren: $x_3 := x_1 \cdot x_2$. Anschließend lässt sich dann eine Regression auf x_1, x_2, x_3 durchführen.

Bei drei Variablen x_i erhält man bereits $\binom{3}{2} = 3$ Interaktionsterme 2. Ordnung $[(x_1, x_2); (x_1, x_3); (x_2, x_3)]$ und $\binom{3}{3} = 1$ Interaktionsterm 3. Ordnung $[(x_1, x_2, x_3)]$. Für 4 Variablen entsprechend $\binom{4}{2} = 6$ Interaktionen 2. Ordnung, $\binom{4}{3} = 4$ Interaktionen 3. Ordnung und $\binom{4}{4} = 1$ Interaktion 4. Ordnung.

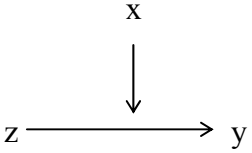
Das Problem der Multikollinearität (vgl. Abschnitt 3.1.10.1) spielt bei der Berücksichtigung von Interaktionstermen eine größere Rolle als sonst, weil z. B. x_1 und (x_1; x_2) in der Regel hoch korrelieren dürften.

Additive und multiplikative Effekte bei metrischen Variablen lassen sich wie folgt charakterisieren:
Ist y ein **additiver Effekt** von x und z, d.h. y = ax + bz, so sind $\frac{d_y}{d_x}$ unabhängig von z und $\frac{d_y}{d_T}$ unabhängig von x.
Ist y ein **multiplikativer Effekt** von x und z, d.h. y = ax · z, so hängt $\frac{d_y}{d_x}$ von z ab $\left(\frac{d_y}{d_x} = f(z)\right)$ und $\frac{d_y}{d_z}$ von x ab $\left(\frac{d_y}{d_z} = g(x)\right)$.

Dies ist der symmetrische Ansatz zur Untersuchung der Interaktion (vgl. Blalock/Blalock 1968). In Kapitel 1 wurde ein asymmetrischer Ansatz zur Untersuchung der Interaktion behandelt, nämlich die Spezifikation:



z modifiziert die Beziehung zwischen x und y. Mit vertauschten Rollen:



x modifiziert die Beziehung zwischen z und y, woraus folgt, dass für verschiedene Werte von x sich die Beziehung zwischen z und y unterschiedlich darstellt.

3.1.9 Anknüpfung an die Tabellenanalyse: Regressionsanalyse der Lebenszufriedenheit

In der Darstellung der Tabellenanalyse wurde in Kapitel 2.2.4.7 gezeigt, dass man bei einem Beispiel von Mayntz et al. (1978) zur Erklärung der Lebenszufriedenheit durch die Zufriedenheit mit den Primärbeziehungen und die Zufriedenheit mit dem Beruf die Interaktion der beiden Erklärungsfaktoren zusätzlich berücksichtigen muss. Der Effekt der Interaktion soll im Folgenden mit Hilfe der multiplen Regression herausgearbeitet werden.

Tabelle 3-2: Untersuchung der Lebenszufriedenheit mit dem linearen Regressionsansatz

	Nicht standardis. Koeffizienten B	Standardisierte Koeffizienten Beta	T	Signifikanz
Konstante	,188		5,347	,000
Zufriedenheit Primärbez. (x)	,295	,278	6,759	,000
Berufszufr. (z)	,670	,677	13,044	,000
Interaktion ($x \cdot z$)	-,430	-,423	-6,923	,000

R-Quadrat: 0,196

Die Lebenszufriedenheit wird zu 19,6 % durch die beiden Pädiktoren erklärt, d.h. durch die Zufriedenheit mit den Primärbeziehungen und durch die Berufszufriedenheit.

Tabelle 3-3: Regression von y auf x bzw. auf z bzw. auf x, z bzw. auf xz bzw. auf x, z, xz

	y auf x		y auf z		y auf x, z		y auf xz		y auf x, z, xz	
Prädik-toren	Multiple R^2 $\frac{2,733}{231,579} = 0,012$		Multiple R^2 $\frac{34,601}{231,579} = 0,149$		Multiple R^2 $\frac{35,993}{231,579} = 0,155$		Multiple R^2 $\frac{11,899}{231,579} = 0,051$		Multiple R^2 $\frac{45,425}{231,579} = 0,196$	
	B	Beta	B	Beta	B	Beta	B	Beta	B	Beta
x	0,500 0,115	0,109			0,325 0,083	0,078			0,188 0,295	0,278
z			0,378 0,382	0,387	0,376	0,380			0,670	0,677
xz							0,492 0,231	0,227	-0,430	-0,423

B = Unstandardisierte Koeffizienten; Beta = Standardisierte Koeffizienten.

Diese Regressionen werden im Folgenden ausführlicher dargestellt.

Einfache Regression von y auf x (Primärbeziehungen)

	x		
y	150	250	400
	150	400	550
	300	650	950

$[yx] = 22.500$ Die Differenz der Kreuzprodukte ist die einfachste Quasi-Maßzahl.

$s_x^2 = s_{xx} = \frac{300}{950} \cdot \frac{650}{950} = 0{,}316 \cdot 0{,}684 = 0{,}2160$ Varianz von x

$s_{yx} = [yx] / n^2 = 0{,}025$ Die Kovarianz ist eine besonders wichtige Normierung der Differenz der Kreuzprodukte.

$b_{yx} = \frac{s_{yx}}{s_x^2} = \frac{0{,}025}{0{,}216} = 0{,}115$ Den Regressionskoeffizient erhält man mit Hilfe von Kovarianz und Varianz des Prädiktors.

$a = \bar{y} - b\bar{x} = 0{,}579 - 0{,}115 \cdot 0{,}684 = 0{,}500 = \frac{150}{300} = \bar{y}_{x=0}$ Regressionskonstante

$s_y^2 = s_{yy} = \frac{400}{950} \cdot \frac{550}{950} = 0{,}421 \cdot 0{,}579 = 0{,}244 = \frac{231{,}579}{950} = \frac{SS_y}{n}$ Varianz von y

$\beta_{yx} = b_{yx} \cdot \frac{s_x}{s_y} = \frac{s_{yx}}{s_y \cdot s_x} = r_{yx}$ Beta-Koeffizient

$= \frac{0{,}025}{0{,}494 \cdot 0{,}465} = 0{,}109$

Prozentsatzdifferenz $d_{yz} = \frac{380}{500} - \frac{170}{450} = 76{,}00 - 37{,}78 = 38{,}22\%$

b_{yx} lässt sich also interpretieren als Unterschied in der Lebenszufriedenheit (y) zwischen den Gruppen (x = 0) und (x = 1).
Bzw.: b_{yx} ist der „Effekt“ in y aufgrund von x.

Einfache Regression von y auf z (Berufszufriedenheit)

	z		
y	280	120	400
	170	380	550
	450	500	950

$[yz] = 86.000$

$s_z^2 = s_{zz} = \frac{450}{950} \cdot \frac{500}{950} = 0{,}474 \cdot 0{,}526 = 0{,}2493; s_z = 0{,}4993$

$s_{yz} = [yz] / n^2 = 0{,}0953$

$b_{yz} = \frac{s_{yz}}{s_x^2} = \frac{0{,}095}{0{,}249} = 0{,}382$

$a = \bar{y} - b\bar{z} = 0{,}579 - 0{,}382 \cdot 0{,}526 = 0{,}378 = \frac{170}{450} = \bar{y}_{z=0}$

Prozentsatzdifferenz $d_{yz} = \frac{380}{500} - \frac{170}{450} = 76{,}00 - 37{,}78 = 38{,}22\%$

Regression von y auf x, z

Zusammenhang der beiden Prädiktoren x, z:

	Zufriedenheit mit den Primärbeziehungen (x)		
Berufszufriedenheit (z)	160	290	450
	140	360	500
	300	650	950

$[zx] = 17.000$

$s_{tx} = [zx] / n^2 = 0{,}019$

$s_{zx}^2 = 0{,}00036$

Um den Einfluss von z bereinigter Regressionskoeffizient $b_{yx.z}$:

$$b_{yx.z} = \frac{s_{xz} - \frac{s_{xz}s_{yz}}{s_z^2}}{s_x^2 - \frac{s_{xz}^2}{s_z^2}}$$

$$= \frac{0{,}025 - \frac{0{,}019 \cdot 0{,}095}{0{,}249}}{0{,}216 - \frac{0{,}00036}{0{,}249}} = \frac{0{,}01775}{0{,}21455} = 0{,}0827$$

Um den Einfluss von x bereinigter Regressionskoeffizient $b_{yz.x}$:

$$b_{yz.x} = \frac{s_{yz} - \frac{s_{xz}s_{yx}}{s_x^2}}{s_z^2 - \frac{s_{xz}^2}{s_x^2}}$$

$$= \frac{0{,}0953 - \frac{0{,}019 \cdot 0{,}025}{0{,}216}}{0{,}2493 - \frac{0{,}00036}{0{,}216}} = \frac{0{,}0931}{0{,}2476} = 0{,}3760$$

Bezugspunkt:

$$a = \bar{y} - b_{x.z}\bar{x} - b_{z.x}\bar{z}$$
$$= 0{,}579 - 0{,}0827 \cdot \frac{650}{950} - 0{,}3760 \cdot \frac{500}{950} = 0{,}325$$

Erklärte Varianz:

$$s_{y,\hat{y}} = s_{yx} \cdot b_{yx} + s_{yz} \cdot b_{yz}$$

$$= 0{,}025 \cdot 0{,}0827 + 0{,}0953 \cdot 0{,}3760$$

$$= 0{,}002 \quad + 0{,}0358$$

$$= \quad 0{,}0378$$

$$\textit{Multiple } R^2 = s_{y,\hat{y}} / s_y^2 = 0{,}0378 / 0{,}244 = 15{,}5\%$$

Regression von y auf xz

Dies ist eine einfache Regression mit einem komplizierteren Prädiktor. Es handelt sich um den Kontrast der Kombination (1,1) vs. dem Rest.

Typ (1,1): $\frac{260}{360} = 72{,}22\%$

Typ ((x, z) ≠ (1,1)): $\frac{30+140+120}{160+290+140} = \frac{290}{590} = 49{,}15\%$

$$B_0 = B_{(x,z)\neq(1,1)} = 49{,}15\%$$

$$B_1 = B_{(x,z)=(1,1)} = 72{,}22 - 49{,}15 = 23{,}07\%$$

(B-Koeffizienten: Nicht standardisierte Koeffizienten im SPSS-Ausdruck.)

Regression von y auf x, z, xz („Saturiertes Modell“)

$$B_0 = B_{x=0,\,z=0} = 18{,}75\%$$
$$B_{x=1,\,z=0} = 48{,}28 - 18{,}75 = 29{,}53\%$$
$$B_{x=0,\,z=1} = 85{,}71 - 18{,}75 = 66{,}96\%$$
$$B_{x=1,\,z=1} = (72{,}22 - 18{,}75) - (29{,}53 + 66{,}96)$$
$$= 53{,}47 - 96{,}49 = -43{,}02\%$$

(Der „multiplikative Effekt“ ist kleiner als die „Summe der additiven Effekte“.)

$$Beta_x = B_x \cdot \frac{s_x}{s_y} = 0{,}295 \cdot \frac{0{,}465}{0{,}494} = 0{,}278$$

$$Beta_z = B_z \cdot \frac{s_z}{s_y} = 0{,}670 \cdot \frac{0{,}499}{0{,}494} = 0{,}677$$

$$Beta_{(xz)} = B_{(xz)} \cdot \frac{s_{(xz)}}{s_y} = -0{,}430 \cdot \frac{0{,}4851}{0{,}4937} = -0{,}423$$

Erklärte Varianz:

$$s_{y,\hat{y}} = s_{yx} \cdot b_{yx} + s_{yx} \cdot b_{yz} + s_{y,xz} \cdot b_{y,xz}$$

$$s_{yx} = \frac{nx}{n} \cdot (\bar{y}_x - \bar{y}) \quad ; b_{yx} = \bar{y}_{x=1,\,z=0} - \bar{y}_{0,0}$$

$$s_{yx} \cdot b_{yx} = \frac{nx}{n} \cdot (61{,}54\% - 57{,}89\%) \cdot (48{,}28\% - 18{,}75\%)$$
$$= 0{,}6842 \cdot (+3{,}65\%) \qquad \cdot (+29{,}53\%)$$
$$= 0{,}00238$$

$$s_{yz} = \frac{nz}{n} \cdot (\bar{y}_z - \bar{y}) \;; b_{yz} = \bar{y}_{x=0,\,z=1} - \bar{y}_{0,0}$$

$$\begin{aligned} s_{yz} \cdot b_{yz} &= 0{,}5263 \cdot (76{,}00\% - 57{,}89\%) \cdot (85{,}71\% - 18{,}75\%) \\ &= 0{,}5263 \cdot (+18{,}11\%) \qquad \cdot (+66{,}96\%) \\ &= 0{,}0638 \end{aligned}$$

		x z = 0	x z = 1	
y	klein	300	100	
	groß	290	260	
		950-360 = 590	360	950

$$s_{y,xz} = \frac{49.000}{950 \cdot 950} = 0{,}0543$$

$$s_{y,xz} = \frac{n_{xz=1}}{n} \cdot (\bar{y}_{xz=1} - \bar{y}) \;\; ; b_{y,xz} = \bar{y}_{x=1,\,z=1} - \left(\bar{y}_{1,1} + (\bar{y}_{1,0} - \bar{y}_{0,0}) + (\bar{y}_{0,1} - \bar{y}_{0,0})\right)$$

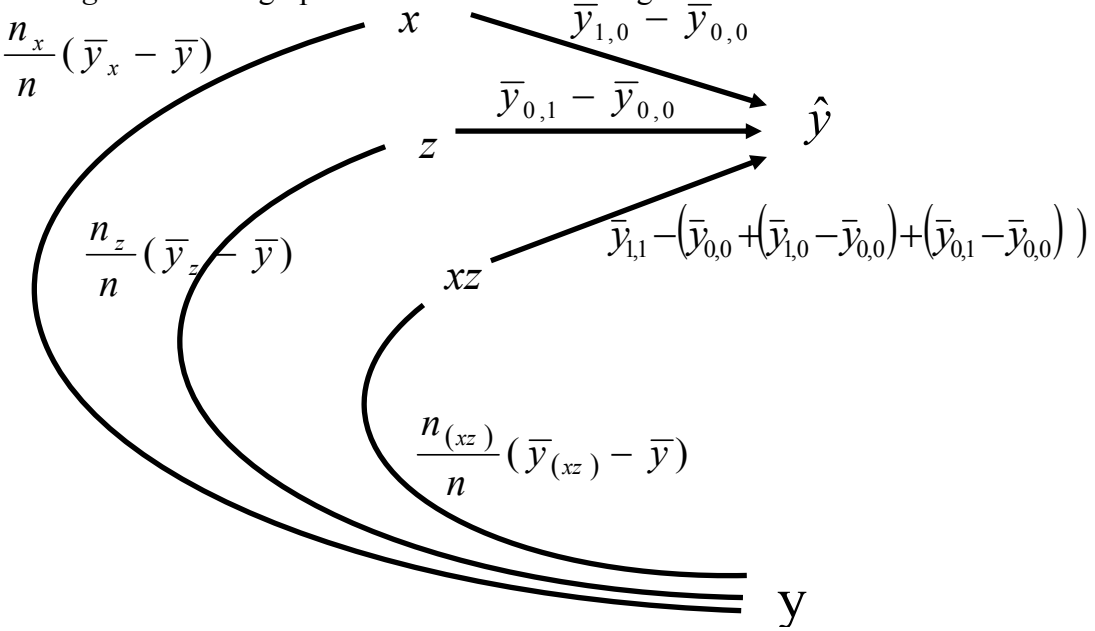
$$\begin{aligned} s_{y,xz} \cdot b_{y,xz} &= 0{,}3789 \cdot (72{,}22\% - 57{,}89\%) \cdot (72{,}22\% - 115{,}24\%) \\ &= 0{,}3789 \cdot (+14{,}33\%) \qquad \cdot (-43{,}02\%) \\ &= -0{,}02336 \end{aligned}$$

Also:

$$s_{y,\hat{y}} = 0{,}00238 + 0{,}0638 - 0{,}02336 = 0{,}0478$$

$$\textit{Multiple}\;\; R^2 = s_{y,\hat{y}} / s_y^2 = 0{,}0478 / 0{,}244 = 19{,}6\%$$

Abbildung 3-4: Meine graphische Veranschaulichung der erklärten Varianz



Die erklärte Varianz ergibt sich daraus, dass y mit den Prädiktoren x und z sowie xz kovariiert, die bei der 1-0-Codierung die angegebenen Effekte haben. Die erklärte Varianz ist also die mit den Kovarianzen gewichtete Summe der Effekte.

3.1.10 Logistische Regression

In der Varianzanalyse will man die Unterschiede von Gruppenmittelwerten schätzen und/oder testen, wobei man voraussetzt, dass die Gruppen eine gleiche Variation aufweisen, damit die unterschiedliche Variation (σ) nicht als Störfaktor bei dem Vergleich wirkt.

Wenn die abhängige Variable (y) eine Dichotomie ist, so gilt für die Gruppe $G_i : \sigma_i^2 = p_i(1-p_i)$, so dass die Streuung mit den Mittelwertunterschieden variieren würde. (Hierbei ist p_i der Anteil der Einheiten mit Ausprägung y_1 in der Gruppe G_i.)

Aus diesem Grund wird die abhängige Variable vor der Analyse erst durch Logit-Transformation $\ln \frac{p}{1-p}$ „geglättet". Denn während ein Anteil p im Intervall [0,1] variiert, variiert $\frac{p}{1-p}$ im Intervall]0, ∞[und $\ln \frac{p}{1-p}$ im Intervall]-∞, +∞[. Die abhängige Variable passt dadurch besser zu einem linearen Modell. Eine Gerade *y = a + bx* ist nur dann in den Werten von y beschränkt, wenn b = 0, d. h. in dem wenig interessanten Fall, dass es keinen Zusammenhang gibt.

Der Modellansatz entspricht ansonsten dem üblichen linearen Regressionsansatz:

$$\ln \frac{p}{1-p} = b_0 + b_1 x_1 + \ldots + b_k x_k$$

Beispiel für die logistische Regression: Lebenszufriedenheit in Abhängigkeit von der Zufriedenheit mit den Primärbeziehungen (x), der Berufszufriedenheit (z) und ihrer Interaktion (x · z)

$y = \ln \frac{p}{1-p}$, wobei p die Wahrscheinlichkeit für Lebenszufriedenheit und 1-p die komplementäre Wahrscheinlichkeit ist, nicht lebenszufrieden zu sein.

Variablen in der Gleichung

	Regressions-koeffizient B	Signifikanz	Exp. (B)
x (1)	1,397	,000	4,044
z(1)	3,257	,000	25,984
x (1) by z (1)	-2,233	,000	,107
Konstante	-1,466	,000	,231

Cox und Snell R-Quadrat: 0,187

Mit einer logistischen Regression lassen sich die relativen Effekte verschiedener Variablen *innerhalb eines Modells* vergleichen. Für verschiedene Gruppen oder Modelle sind die Effekte nicht zu vergleichen (vgl. Mood 2010). Also lässt sich die logistische Regression auch nicht einfach pfadanalytisch erweitern wie die lineare Regression.

Interpretation der Quotienten („Odds“) und der logarithmierten Quotienten („Log Odds“ oder auch „Logits“)

Ausgangspunkt: Kreuztabelle für zwei dichotome Merkmale (Vierfeldertafel)

	z.B. zufrieden mit Beziehungen	nicht zufrieden mit Beziehungen
	a	b
	c	d

a/c nennt man Odds, d.h. Chance, und zwar hier die spaltenbezogene Chance, dass ein Fall in die erste Zeile fällt, und nicht in die zweite Zeile. (Eine Art Wette um das Eintreten der ersten gegenüber der zweiten Möglichkeit.) Z.B. Gesamtzufriedenheit ja vs. nein.

Analog ist b/d die Chance bzgl. der zweiten Spalte.

$\frac{a / c}{b / d} = \frac{ad}{bc}$ nennt man Odds-Ratio.

Wenn dieses multiplikative Konzept des Odds-Ratio durch Logarithmieren in eine additive Form gebracht wird, so spricht man von Log-Odds oder Logits:

$\ln \frac{ad}{bc}$

Zur Interpretation in dem konkreten Beispiel

In einem Gedankenexperiment kann man sich quasi-experimentell vorstellen, was für einen Effekt es hat, wenn eine Person zufrieden mit ihren Primärbeziehungen wird. Als direkter Effekt würde sich $\ln \frac{p}{1-p}$ um 1,397 erhöhen und $\frac{p}{1-p}$ um 4,044. (Die Änderung in x kann neben dem direkten Effekt auch indirekte Effekte anstoßen, wie es in dem Beispiel mit dem Interaktionseffekt $x \cdot z$ ja auch offensichtlich der Fall ist.) Da $\frac{p}{1-p}$ einfacher zu formulieren ist, wird die Erklärung für das Chancenverhältnis $\frac{p}{1-p}$ formuliert. Erklärt wird in dem Beispiel also die Chance der Gesamtzufriedenheit mit dem Leben (statt insgesamt nicht zufrieden zu sein): $\frac{p}{1-p}$

p ist hierbei die Chance der Gesamtzufriedenheit. 1 – p ist die Chance, insgesamt nicht zufrieden zu sein.
Eine Person, die zufrieden ist mit den Primärbeziehungen, hat insgesamt (d.h. in der gesamten Zufriedenheit mit dem Leben) eine ca. 4 mal so große Chance zufrieden zu sein (statt unzufrieden) wie eine Person, die nicht zufrieden ist mit den Primärbeziehungen.

Entsprechend: Eine Person, die zufrieden ist mit dem Beruf, hat insgesamt eine ca. 26 mal so große Chance zufrieden zu sein wie eine Person, die nicht zufrieden ist mit dem Beruf.

Ferner:
Eine Person, die sowohl zufrieden ist mit den Primärbeziehungen als auch mit dem Beruf, weist nach den vorliegenden Daten von Mayntz et al. in der Gesamtzufriedenheit nur ein Zehntel der Zufriedenheits-Chance der Personen mit den übrigen Kombinationen auf. In diesem Beispiel liegt dies wohl daran, dass ein Teil der Personen stärker berufsorientiert und ein Teil der Personen stärker beziehungsorientiert ist, so dass sich die Effekte nicht einfach kumulieren.

3.1.11 Statistische Inferenz

Der Ansatz der einfach linearen Regression lautet: $\hat{y}_i = \alpha + \beta\, x_i$

Der Fehlerterm beträgt: $y_i - \hat{y}_i = u_i$

Die Methode der kleinsten Quadrate besteht im Minimieren der Abweichungsquadrate.

Für die statistischen Inferenz von der Stichprobe auf die Grundgesamtheit müssen folgende Annahmen gemacht werden: Die Fehler u_i sind statistisch unabhängige Zufallsvariablen mit dem Mittelwert 0 – d.h.: es gibt keine systematischen Über- bzw. Unterschätzungen – und haben die gleichen Varianz σ^2. (Äquivalent: Die y_i sind statistisch unabhängige Zufallsvariablen mit Mittelwert $\alpha + \beta\, x_i$ und Varianz σ^2.)

Auf Grund der Beobachtungswerte $(x_1, y_1), \ldots, (x_n, y_n)$ lassen sich a und b für die entsprechenden Parameter α und β in der Grundgesamtheit schätzen. Diese Schätzungen haben die Eigenschaften:

$$E(a) = \alpha$$

$$Var(a) = \frac{\sigma^2}{n}$$

$$E(b) = \beta$$

$$Var(b) = \frac{\sigma^2}{\sum (x_i - \bar{x})^2}$$

Beweis für b:

Die y_i sind unabhängige Zufallsvariablen.

$b = \frac{s_{xy}}{s_x^2} = \sum_i \frac{x_i - \bar{x}}{\sum_j (x_j - \bar{x})^2} (y_i - \bar{y})$ ist eine gewichtete Summe von Zufallsvariablen.

$$E(b) = \sum_i \frac{x_i - \bar{x}}{\sum_j (x_j - \bar{x})^2} E(y_i - \bar{y}) = \beta \quad \text{,wobei}$$

$$E(y_i - \bar{y}) = E(u_i) + \beta (x_i - \bar{x}) \quad [\text{mit } E(u_i) = 0]$$

Da $\sigma^2 = Var(y_i - \bar{y})$, so gilt:

$$Var(b) = \sum_i \frac{(x_i - \bar{x})^2}{\left(\sum_j (x_j - \bar{x})^2\right)^2} Var(y_i - \bar{y}) = \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2}$$

Bei der *einfachen linearen Regression* $y = \alpha + \beta x$ hat also die Schätzung b unter diesen Annahmen den Mittelwert β und die Varianz $\frac{\sigma^2}{\left(\sum (x_i - \bar{x})^2\right)}$, wobei σ die Varianz von y ist.

Die normierte Variable $z = \frac{b - \beta}{\sqrt{\frac{\sigma^2}{\left(\sum (x_i - \bar{x})^2\right)}}}$ ist normalverteilt mit dem Mittelwert 0 und der Varianz 1 – d.h.: standardnormalverteilt –, falls die y_i normalverteilt sind oder der Stichprobenumfang hinreichend groß ist n > 200.

Da die Varianz von y oft nicht bekannt ist, wird sie durch die „residuale Varianz“ geschätzt:
$s^2 = \frac{1}{n-2} \sum (y_i - \hat{y}_i)^2$

Die neue Variable $t = \frac{b - \beta}{\sqrt{\frac{s^2}{\left(\sum (x_i - \bar{x})^2\right)}}}$ ist verteilt[7] nach t_{n-2}.

Wie in der Inferenzstatistik üblich, lassen sich denn Konfidenzintervalle für β berechnen:
Konfidenzintervall $(b - z_{1-\alpha/2}\, s_b) < \beta < (b + z_{1-\alpha/2}\, s_b)$

Mit Wahrscheinlichkeit 1 – α liegt der Koeffizient β in dem angegebenen Intervall. (Bei kleiner Stichprobe (n ≤ 200) muss man mit der t-Verteilung statt mit der Normalverteilung arbeiten: $\pm t_{n-2;1-\alpha/2}$.)

Durch den Signifikanztest der Regressionskoeffizienten wird geprüft, ob diese überzufällig von Null verschieden sind.

Die Signifikanz von b lässt sich mit einem Quotienten von χ^2-verteilten Größen (jeweils dividiert durch die Anzahl der Freiheitsgrade) testen, der gemäß Ergebnissen der Inferenzstatistik F-verteilt ist.

Der Quotient $\frac{\left(\frac{\sum (\hat{y}_i - \bar{y})^2}{1}\right)}{\left(\frac{\sum (y_i - \hat{y}_i)^2}{(n-2)}\right)}$ ist $F_{1,n-2}$ verteilt.

---

[7] Unter Hypothesen „$\beta = 0$“: $t^2_{n-2} = \frac{b^2}{\frac{s^2}{\sum (x_i - \bar{x})^2}} = \frac{b^2 \left(\sum (x_i - \bar{x})^2\right)}{s^2} = \frac{\sum (\hat{y}_i - \bar{y})^2}{s^2}$ ist $F_{1,n-2}$ verteilt. T-Test und F-Test sind also austauschbar.

Die Hypothese „H_0 : β = 0“ – d.h.: kein linearer Zusammenhang zwischen y und x – wird zurückgewiesen, wenn der F-Wert größer ausfällt, als er bei gegebenem Signifikanzniveau von z.B. 5 % wahrscheinlich wäre. Dieser kritische Wert lässt sich einer F-Tabelle entnehmen.

Im Falle der *multiplen Regression* von y auf x_1, ..., x_k ändert sich die Anzahl der Freiheitsgrade. Die Testgröße für „H_0: Der multiple Korrelationskoeffizient R² ist gleich Null“ lautet:

$$F = \frac{\left(\frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{k}\right)}{\left(\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{(n-k-1)}\right)} = \frac{\left(\frac{R^2}{k}\right)}{\left(\frac{(1-R^2)}{n-k-1}\right)}$$ ist $F_{k,n-k-1}$ verteilt.

Für k = 1 erhält man gerade den Spezialfall der einfachen Regression mit der entsprechenden Testgröße.

F-Test für die einzelnen Regressionskoeffizienten

Ein F-Test für die einzelnen Regressionskoeffizienten kann nach folgenden zwei verschiedenen Verfahren durchgeführt werden:

1) der Standardmethode für die Regressionsanalyse und
2) der hierarchischen Methode.

1) <u>Standardmethode</u>

Jede Variable x_j wird daraufhin untersucht, ob sie noch einen *signifikanten Beitrag* zur Erklärung leisten kann, *wenn alle anderen unabhängigen Variablen bereits in die Regressionsanalyse eingeführt worden sind.*

$$F = \frac{\left(\frac{\left(\text{Zuwachs in } \sum(\hat{y}_i - \bar{y})^2 \text{ aufgrund von } x_j\right)}{1}\right)}{\left(\frac{\sum(y_i - \hat{y}_i)^2}{(n-k-1)}\right)}$$ ist $F_{1,n-k-1}$ verteilt.

Dieser Ausdruck kann auch geschrieben werden als:

$$F = \frac{\left(\frac{r^2_{y,(x_j \cdot x_1,\ldots,x_{j-1},x_{j+1},\ldots,x_k)}}{1}\right)}{\left(\frac{(1-R^2_y; x_1,\ldots,x_k)}{(n-k-1)}\right)} = \frac{\frac{r_{y,x_i-\hat{x}_i}}{1}}{\frac{1-R^2}{n-k-1}}$$

Hierbei ist $r^2_{y,(x_j \cdot x_1,\ldots,x_k)}$ die Änderung in R^2 aufgrund Einführung von x_j:

$$r^2_{y,(x_j \cdot x_1,\ldots,x_{j-1},x_{j+1},\ldots,x_k)} = R^2_{y;x_1,\ldots,x_{j-1}x_{j+1},\ldots x_k} - R^2_{y;x_1,\ldots,x_{j-1},x_{j+1},\ldots,x_k}$$

2) Hierarchische Methode

Die Signifikanz der einzelnen Koeffizienten lässt sich auch derart testen, ob der Zuwachs an Erklärung durch *Hinzufügen einer Variablen* x_j *gemäß einer durch den Forscher vorgegebenen Reihenfolge* eine signifikante Verbesserung der Erklärung liefert.

$$F = \frac{\left(\frac{\left(\text{Zuwachs in } \sum(\hat{y}_i - \bar{y})^2 \text{ aufgrund von } x_j\right)}{1}\right)}{\left(\frac{\sum(y_i - \hat{y}_i)^2}{(n-k-1)}\right)} \text{ ist } F_{1,n-k-1} \text{ verteilt.}$$

Dieser Ausdruck beträgt für die nach der vorgegebenen Ordnung eingeführte erste Variable:

$$F = \frac{\left(\frac{r^2_{y,x_1}}{1}\right)}{\left(\frac{\left(1-R^2_{y;x_1,\ldots,x_k}\right)}{(n-k-1)}\right)}$$ Hierbei ist r^2_{y,x_1} der durch die erste eingeführte Variable erklärte Anteil der Varianz.

Für die zweite nach der vorgegebenen Ordnung eingeführte Variable:

$$F = \frac{\left(\frac{r^2_{y,(x_2 \cdot x_1)}}{1}\right)}{\left(\frac{\left(1-R^2_{y;x_1,\ldots,x_k}\right)}{(n-k-1)}\right)}$$ Hierbei ist $r^2_{y;(x_2 \cdot x_1)}$ der Zuwachs in R^2 aufgrund des Hinzufügens von x_2.

Für die dritte Variable:

$$F = \frac{\left(\frac{r^2_{y,(x_3 \cdot x_1,x_2)}}{1}\right)}{\left(\frac{\left(1-R^2_{y;x_1,\ldots,x_k}\right)}{(n-k-1)}\right)}$$. Etc.

Im SPSS-Programm wird nach der Standardmethode vorgegangen. Die Werte nach der hierarchischen Methode lass sich aber mittels des Outputs berechnen.

Voraussetzung der statistischen Inferenz in der Regressionsanalyse waren Annahmen über die Verteilung der Schätzfehler, also der Residuen $y_i - \hat{y}_i = u_i$. Im SPSS-Programm kann man anhand eines Streudiagramms inspizieren, ob die Rediduen die für die statistischen Inferenz erforderlichen Voraussetzungen erfüllen (Streudiagramm: $y_i - \hat{y}_i$ als Funktion von $\hat{y}_i$).

Varianzanalyse als Test für die Erklärungskraft des gesamten Regressionsansatzes

Die Güte der Regressionsanalyse kann durch die Varianzanalyse geprüft werden. Die Nullhypothese lautet dann „$H_0 : R^2 = 0$“.
Dazu wird die Gesamtvarianz zerlegt in den Anteil der durch die Regression nicht erklärten Varianz (Residuen) und den multiplen Korrelationskoeffizienten R^2:

$$1 = \frac{\sum (y_i \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} + \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2}$$

Tabelle 3-4: Varianzanalyse für die *multiple Regression* von y auf $x_1, ..., x_k$ mit den Freiheitsgraden k bzw. n – k – 1

	„Sum of Squares“	Freiheitsgrade (df)	Varianz („Mean Square“)
Durch die Regression Erklärt	$\sum (\hat{y}_i - \bar{y})^2$	k	$\frac{\sum (\hat{y}_i - \bar{y})^2}{k}$
Nicht erklärter Rest (Residuum)	$\sum (y_i - \hat{y}_i)^2$	$n-k-1$	$\frac{\sum (y_i - \hat{y}_i)^2}{(n-k-1)}$
Insgesamt	$\sum (y_i - \bar{y})^2$	$n-1$	

$$F = \frac{\frac{\sum (\hat{y}_i - \bar{y})^2}{k}}{\frac{\sum (y_i - \hat{y}_i)^2}{n-k-1}}$$ ist $F_{k, n-k-1}$ verteilt.

Gemäß der Varianzzerlegung kann man F auch charakterisieren durch:

$$F = \frac{\frac{R^2}{k}}{\frac{1-R^2}{(n-k-1)}}$$

3.1.12 Die Verletzung der Modellannahmen

Es ist zu prüfen, ob Verletzungen der Änderungsvoraussetzungen vorliegen. Als hier nicht weiter zu erörternde Grundvoraussetzungen gelten das theoriegeleitete Vorgehen bei der Zusammenstellung von Pädiktoren und Kriterium sowie die Normalverteilung der Variablen in der Grundgesamtheit.

(1) **Multikollinearität**

Multikollinearität bezieht sich auf Modellverletzungen innerhalb der Gruppe der unabhängigen Variablen.

Multikollinearität liegt dann vor, wenn eine unabhängige Variable mit einer oder mehreren unabhängigen Variablen stark korreliert, so dass sich eine Variable nahezu als eine Linearkombination der übrigen darstellen lässt. Bei einer zu starken linearen Abhängigkeit der unabhängigen Variablen untereinander lässt sich die Regressionsanalyse nicht mehr durchführen.

Das liegt daran, dass zur Berechnung der Regressionskoeffizienten die Inverse der Korrelationsmatrix benutzt wird. Besteht eine zu starke lineare Abhängigkeit unter den unabhängigen Variablen, dann ist die Korrelationsmatrix nicht mehr invertierbar.

Die Multikollinearität hat folgende Auswirkungen (vgl. z.B. Opp und Schmidt 1976: 168-184):

- Die Standardfehler der Koeffizienten (bei der statistischen Inferenz) nehmen stark zu.
- Das Vorzeichen der Koeffizienten kann falsch werden und die standardisierten Koeffizienten können sehr große Werte (also auch natürlich > 1) haben.

(Da $\beta_{y,x_1 \cdot x_2} = \frac{y_{y,x_1} - r_{x_1,x_2} r_{y,x_2}}{1 - r_{x_1,x_2}^{\ 2}}$,so kann der Nenner beliebig klein, β also beliebig groß werden (wenn $r^2_{x_1,x_2} \to 1$).)

Als Daumenregel empfehlen Opp und Schmidt: $r^2_{x_1,x_2} < 0{,}6$

Außerdem kann das Bestimmtheitsmaß R^2 als signifikant ausgegeben werden, obwohl kein Koeffizient signifikant ist.

Als Auswege bieten sich an:

- Nur eine Variable aus dem Bündel der hoch miteinander korrelierenden Variablen wird als Repräsentant für dieses Bündel von Variablen in die Regressionsgleichung eingeführt.

- Es wird ein Index aus den hoch miteinander korrelierenden Variablen gebildet, der dann in die Regressionsanalyse eingeführt wird. Allerdings treten dann die üblichen Probleme der Indexbildung auf (Gewichtung, Verknüpfung etc.)

Weiterhin könnten weitere Berechnungen des Regressionsmodells mit (teilweise) anderen Prädiktoren vorgenommen werden.

Multikollinearität kann mit einer speziellen Kollinearitätsdiagnose im SPSS-Programm ermittelt werden (vgl. das Beispiel 3.1.13).

(2) **Nichtlinearität**

Die Beziehung zwischen einer abhängigen Variablen (Kriterium) und einer oder mehreren unabhängigen Variable(n) muss nicht immer linear sein (weiterhin denkbar sind quadratische, kubische, logarithmische, exponentielle oder logistische Zusammenhänge). Für die multiple Regression ist Linearität jedoch eine notwendige Voraussetzung.

Nichtlinearität führt zu Ungültigkeit der Testergebnisse (F- und t-Test. Zu identifizieren ist sie im SPSS-Programm mit Punkt-(Scatter-)Diagramm. Nichtlinearität kann z.B. durch Transformation der nichtlinearen in lineare Beziehungen begegnet werden. Ihre Existenz ist statistisch testbar.

(3) **Heteroskedastizität**

Ist die Varianz der Fehler e nicht für alle Variablen homogen, kommt es vor, dass die Streuung der Residuen vom Betrag oder der Reihenfolge der Beobachter der Prädiktoren abhängt. Anders gesagt müssen alle Residuen paarweise voneinander linear unabhängig (unkorreliert) sein und alle in ihren Verteilungen denselben Mittelwert 0 und die gleiche Varianz haben. Ist dies nicht der Fall, liegt Heteroskedastizität vor.

Heteroskedastizität führt zur Verzerrung der Konfidenzintervall-Schätzungen. Man erkennt die Heteroskedastizität entweder mittels eines Punktdiagramms oder mittels des Geldfeld/Quandt-Tests.

(4) **Keine Normalverteilung der Residuen**

Um bei der Signifikanzprüfung zu unverzerrten Ergebnissen zu kommen, ist die Voraussetzung der Normalverteilung der Residuen von Bedeutung. Die Residuen (die Differenz zwischen den beobachteten und den nach der Regressionsschätzung theoretisch zu erwartenden Werten, also die Fehler e) dürfen nicht systematisch auftreten, da sonst die Schätzung und die Signifikanzprüfung zu verzerrten Ergebnissen führen. Die Residuen müssen in ihrer Verteilung der Normalverteilung folgen. Dies lässt sich im SPSS-Programm wiederum durch verschiedene Grafiken prüfen.

(5) **Autokorrelation**

Korrelieren die Residuen in der Grundgesamtheit, liegt Autokorrelation vor. Mit anderen Worten bezeichnet Autokorrelation systematische Verbindungen zwischen den Residuen benachbarter Fälle; z.B. hängen die Residuen vom jeweils vorherigen Beobachtungswert ab, was eine verzerrte Bestimmung des Standardfehlers nach sich zieht.

Die Untersuchung der Autokorrelation empfiehlt sich bei Längenschnittdaten, da hier aufeinanderfolgende Zeitpunkte die Fälle darstellen, weshalb Autokorrelation wahrscheinlicher ist als bei Querschnittsdaten.

Im SPSS-Programm ist der Durbin-Watson-Test für die Aufdeckung von Autokorrelation verfügbar.

3.1.13 Beispiel für die Regressionsanalyse

Die Wohlfahrtsforschung beschäftigt sich unter anderem mit der Lebensqualität der Bürger und Bürgerinnen in der Bundesrepublik Deutschland. Neben sogenannten objektiven Indikatoren (Einkommen, Bildung, Berufsprestige, Geschlecht etc.) steht auf der Seite der subjektiven Indikatoren vor allen die allgemeine Lebenszufriedenheit im Zentrum des Interesses. Welche objektiven oder subjektiven Determinanten bedingen die Lebenszufriedenheit? Welche sind sehr wichtig, welche weniger wichtig und welche tragen nichts zur Erklärung der Lebenszufriedenheit bei?

Diesen Fragen soll anhand von Daten der Sozio-ökonomischen-Panels für 1995 nachgegangen werden, wobei folgende Hypothesen aufgestellt wurden:

a) Die allgemeine Lebenszufriedenheit ist abhängig von der Zufriedenheit in einzelnen Lebensbereichen.

b) Die Lebenszufriedenheit wird primär durch die Zufriedenheit mit dem Lebensstandard und mit dem Haushaltseinkommen strukturiert, da in das Konzept des Lebensstandards und des Haushaltseinkommens sowohl die Befriedigung materieller Bedürfnisse als auch soziale Vergleichsprozesse einfließen.
c) Die Lebenszufriedenheit wird stark durch die Zufriedenheit mit der Arbeit strukturiert, da die Arbeit für die Mehrheit der Menschen einen der wichtigsten Lebensmittelpunkte darstellt. Viele Menschen verbringen nahezu die Hälfte ihrer Zeit mit der Arbeit. Schlechte Arbeitsbedingungen könnten sich somit, vermittelt über die Zufriedenheit mit der Arbeit, auf die Lebenszufriedenheit auswirken.
d) Neben der Arbeit ist der Bereich der Freizeittätigkeit relevant für die Lebenszufriedenheit. Der Bereich der Freizeittätigkeit ermöglicht sowohl die Distanz zur zeitaufwendigen Erwerbstätigkeit als auch die Verwirklichung persönlicher Neigungen und Ziele. Beeinträchtigungen im Bereich der Freizeittätigkeit schlagen sich demnach auf die Lebenszufriedenheit nieder.
e) Als letzter Indikator für die Lebenszufriedenheit wird die Zufriedenheit mit der Gesundheit in Betracht gezogen. Anhaltende oder auch kurzfristige Beeinträchtigungen der Gesundheit sollten einen Einfluss auf die Lebenszufriedenheit haben, da diese Beeinträchtigungen (chronische Krankheiten, Behinderungen, Unfallfolgen etc.) das tägliche Leben in seiner Durchführung verkomplizieren.

Um die Bedeutung der Indikatoren für die allgemeine Lebenszufriedenheit zu bestimmen, wird eine lineare multiple Regression durchgeführt.

1. LP0108 Zufriedenheit Lebensstandard
2. LP0101 Zufriedenheit Gesundheit
3. LP0102 Zufriedenheit Arbeit
4. LP0107 Zufriedenheit Freizeittätigkeit
5. LP0104 Zufriedenheit Haushaltseinkommen

Abhängige Variable: LP10401 Lebenszufriedenheit gegenwärtig

1. Prüfung der Voraussetzungen:

a) Prüfung auf Multikollinearität

Man erhält u.a. den folgenden Teil des Ausdrucks, der das Vorliegen von Multikollinearität prüft.

---------- Variables in the Equation ----------

Variable	**Tolerance**	VIF	T	Sig T
LP0104	**,538170**	1,858	10,935	,0000
LP0101	**,768031**	1,302	24,072	,0000
LP0102	**,709773**	1,409	19,629	,0000
LP0107	**,735366**	1,360	12,500	,0000
LP0108	**,465430**	2,149	24,686	,0000
(Constant)			27,381	,0000

Equation Number 1 Dependent Variable: LP10401 Lebenszufriedenheit gegenwärtig

Collinearity Diagnostics

Numbers	Eigenval	Cond Index	Variance Constant	Proportions LP0104	LP0101	LP0102	LP0107	LP0108
1	5,74227	1,000	,00133	,00170	,00176	,00188	,00203	,00095
2	,0790	8,524	,01971	,41614	,09066	,00263	,21550	,01898
3	,06810	9,183	,00452	,02504	,16494	,29096	,42535	,03069
4	,04878	10,849	,05359	,02503	,37251	,70050	,12983	,00520
5	,03784	12,319	,81340	,05462	,36987	,00109	,09323	,00616
6	,02398	15,475	,10744	,47748	,00027	,00293	,13406	,93801

End Block Number 1. All requested variables entered.

Im oberen Teil des Ausdrucks ist eine Spalte mit der Überschrift *Tolerance* zu finden. Die Toleranz ist folgendermaßen definiert:

$\text{Toleranz}_i = 1 - R_i^2$

Der multiple Korrelationskoeffizient R_i^2 bezieht sich hierbei auf den Sachverhalt, dass die i-te unabhängige Variable durch die anderen unabhängigen Variablen erklärt wird. Eine sehr kleine Toleranz spricht für das Vorliegen von Multikollinearität. Brosius (Brosius 2002: 564) schlägt vor, bei Toleranzwerten kleiner 0,1 davon auszugehen, dass Multikollinearität vorliegt, Werte < 0,01 weisen mit hoher Sicherheit auf Multikollinearität hin.

Anhand dieser Faustregel zeigen unsere Prädiktoren keine Multikollinearität, liegt doch auch der kleinste Toleranzwert mit rund 0,47 weit über der Grenze von 0,1.

Die Spalte VIF (Variance Inflation factor) gibt den Kehrwert der Toleranz an, ist also umgekehrt zur Toleranz zu interpretieren. Hier deuten hohe Werte auf Multikollinearität hin.

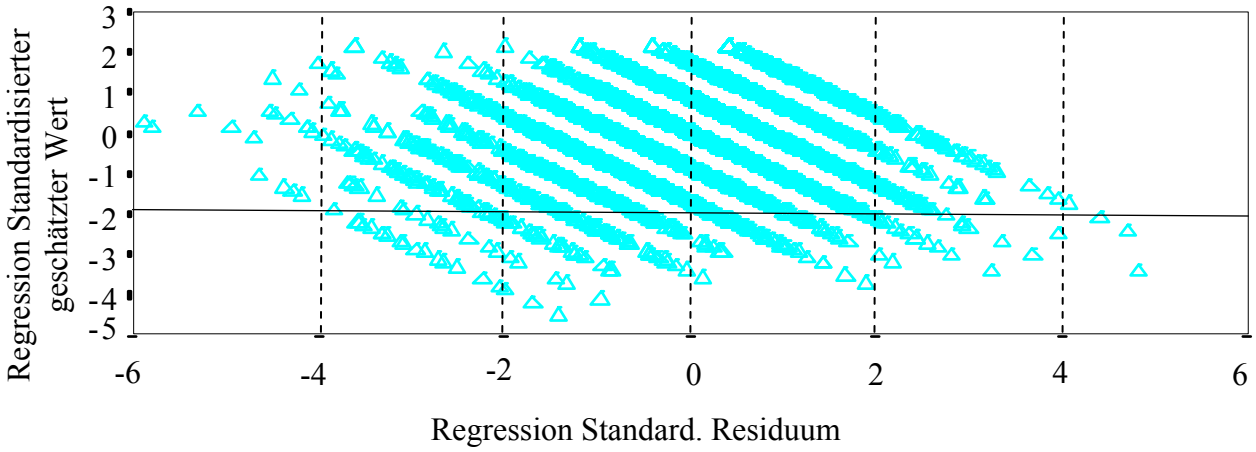
Im unteren Teil des Ausdrucks erfolgt die eigentliche Kollinearitätsdiagnose.

Die Spalte mit der Überschrift *Eigenval* bezeichnet die Eigenwerte der Varianz-Kovarianz-Matrix. Die nächste Spalte bezieht sich auf den Konditionsindex (*Cond Index*). Er wird berechnet als die Quadratwurzel aus dem Quotienten aus größtem Eigenwert (Zähler) und dem zugehörigen Eigenwert (Nenner). Demnach sprechen große Konditionsindizes für Multikollinearität. Brosius (Brosius 2002: 589) erwähnt eine grobe Daumenregel, wonach Werte zwischen 10 und 30 für den Konditionsindex auf mäßige, Werte über 30 auf starke Multikollinearität hinweisen. In unserem Beispiel sind drei Indizes geringfügig größer als 10. Lediglich der letzte Wert in der Spalte könnte auf Multikollinearität hinweisen. In der letzten Spalte sind die Varianzanteile (Variance Proportions) zu finden. Hier werden die Varianzen der Regressionskoeffizienten, zerlegt in Komponenten, die den Eigenwerten zuzurechnen sind, dargestellt. Lineare Abhängigkeiten von unabhängigen Variablen sind dadurch identifizierbar, dass pro Zeile (pro Eigenwert) hohe Werte für alle Variablen ausgewiesen werden. Auch dieses letzte Kriterium lässt in unserem Beispiel nicht auf starke Multikollinearität schließen.

b) Prüfung auf Nichtlinearität

Wie oben bereits erwähnt, ist die Linearitätsannahme eine Voraussetzung, um das Verfahren der linearen multiplen Regression anwenden zu dürfen. Deswegen empfiehlt es sich, die unterstellte lineare Beziehung zu überprüfen. Für diese Prüfung inspiziert man ein Streudiagramm, in dem die Beziehung zwischen den standardisierten Vorhersagewerten und den standardisierten Residuen dargestellt wird.

Abbildung 3-5: Scatterplot – Abhängige Variable: Lebenszufriedenheit gegenwärtig



Die obige Grafik stellt eine Punktwolke dar, die keinen systematischen Kurvenverlauf aufweist. Bei einem quadratischen Verhältnis etwa wie $Y = b_0 + b_1 * X + b_2 * X^2$ müsste eine quadratische Verteilung der Punktwolke in Form einer Parabel erscheinen. Die Betrachtung des Streudiagramms führt zu dem Ergebnis, dass es sich hier tatsächlich um einen linearen Zusammenhang handelt.

c) Heteroskedastizität

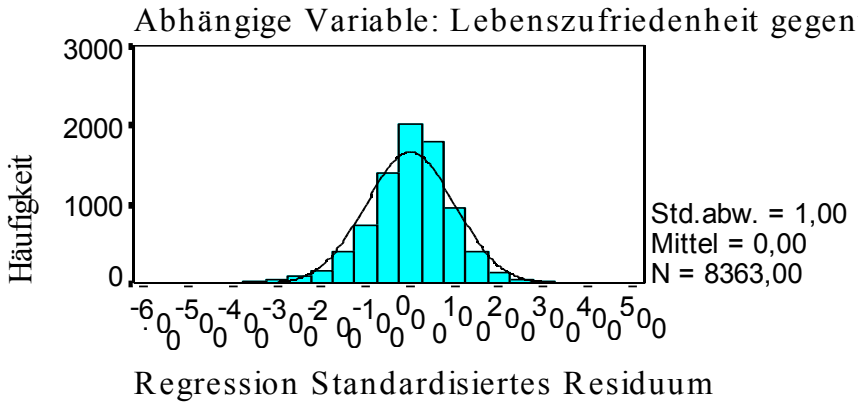
Sowohl der Geldfeld/Quant-Test als auch die bereits diskutierte Grafik können Heteroskedastizität ermitteln. Auf den Geldfield/Quant-Test soll an dieser Stelle nicht eingegangen. Viel mehr wird ein weiteres Mal das Streudiagramm betrachtet. Wie wir erkennen, ist weder die Größe noch die Reihenfolge der Vorhersagewerte für die Streuung der Residuen verantwortlich, denn sonst müsste die Punktwolke in ihrer Breite wachsen oder abnehmen. Daraus folgt, dass keine auffällige Streuung des Residuen zu beobachten sind. Die Annahme der Heteroskedastizität kann also fallengelassen werden. Mit anderen Worten, es liegt keine Heteroskedastizität vor, eine weitere Voraussetzung für die Anwendung der multiplen linearen Regression ist erfüllt.

Auch die Annahme, dass die Mittelwerte der Residuen gleich Null sind, lässt sich mit der dargestellten Grafik überprüfen. Diese Annahme gilt als gewährleistet, wenn die Zentren der Punktkonzentration, die auf Parallelen der Y-Achse liegen (gepunktete Hilfslinien), in etwa auf der Parallelen der X-Achse durch den Wert 0 (durchgezogene Hilfslinie in der Grafik) treffen. Dieser Sachverhalt scheint in unserem Beispiel ebenfalls gewährleistet.

d) Normalverteilung der Residuen

Für die Prüfung der Normalverteilung der Residuen sind prinzipiell zwei Tests möglich. Der erste Test zählt die Häufigkeit der Residuen (Histogramm) und vergleicht sie mit der theoretischen Normalverteilung:

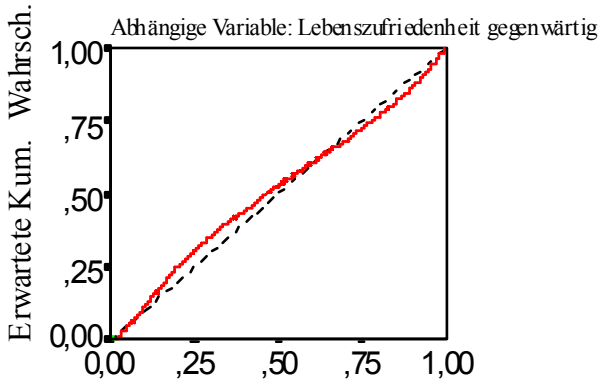
Abbildung 3-6: Histogramm – Abhängige Variable: Lebenszufriedenheit



Es sieht so aus, dass die empirische Verteilung der Residuen nicht deutlich von der theoretischen Normalverteilung abweicht. Die Bedingung, dass die Residuen normalverteilt sein müssen, also zufällig streuen, ist nach dieser Grafik erfüllt.

Ein zweiter Test bietet der sogenannte P-P-Plot (Proportion-Proportion-Plot). Der P-P-Plot stellt die empirischen kumulierte Häufigkeitsverteilung der standardisierten Residuen der zu erwartenden kumulierten Wahrscheinlichkeitsverteilung unter der Annahme der Normalverteilung gegenüber. Die Werte dieser Gegenüberstellung müssen sich auf einer Geraden befinden, wenn die Normalverteilungsannahme gewährleistet sein soll.

Abbildung 3-7: Normal P-P-Plot von Regression Standardisiertes Residuum – Abhängige Variable: Lebenszufriedenheit gegenwärtig



Wie schon das Histogramm verrät, zeigt auch der P-P-Plot leichte Abweichungen von der Normalverteilung. Allerdings sind diese relativ gering und somit tolerierbar. Die Wertekombinationen von beobachteter und erwarteter kumulierter Wahrscheinlichkeit liegen fast genau auf einer Geraden. Die Normalverteilung der Residuen ist zulässigen Einschränkungen gegeben.

e) Autokorrelation

Sollte Autokorrelation existieren, gibt der Durbin-Watson-Test Auskunft. Diese Testgröße wird mit folgender Zeile ausgegeben.

Durbin-Watson Test = 1,81115

Der Durbin-Watson-Koeffizient variiert zwischen 0 und 4. Der Wert 0 steht für positive, der Wert 4 für negative Autokorrelation. Nur der Wert 2 spricht für die Einhaltung der Voraussetzungen, dass keine Autokorrelation besteht. Brosius (Brosius 2002: 584/585) berichtet jedoch von Faustregeln, wonach Werte des Durbin-Watson-Koeffizienten zwischen 1,5 und 2,5 akzeptabel sind.

Mit einem Wert von 1,81 für den Durbin-Watson-Koeffizient in unserem Beispiel besteht also kein Anhaltspunkt für Autokorrelation, die Unabhängigkeit der Residuen ist gewährleistet.

Damit ist auch die fünfte Voraussetzung für die Durchführung der multiplen linearen Regression erfüllt.

Demnach können wir die Schätzungen mit den Prädiktoren und dem Kriterium ohne die Befürchtung durchführen, verzerrte Signifikanzen bzw. Testgrößen zu erhalten.

Bevor wir uns dem SPSS-Output widmen, soll der Vollständigkeit halber noch gesagt werden, dass die fehlenden Werte listenweise ausgeschlossen werden: Wenn ein Wert zu einer Person fehlt, wird der Fall der Person im ganzen ausgeschlossen.

Output * MULTIPLE REGRESSION *****

Listwise Deletion of Missing Data

	Mean	Std Dev	Label
LP10401	6,994	1,697	Lebenszufriedenheit gegenwärtig
LP0104	6,185	2,172	Zufriedenheit Haushaltseinkommen
LP0101	6,986	2,046	Zufriedenheit Gesundheit
LP0102	6,892	2,191	Zufriedenheit Arbeit
LP0107	6,647	2,168	Zufriedenheit Freizeittätigkeit
LP0108	6,874	1,868	Zufriedenheit Lebensstandard

N of Cases = 8363

Correlation, 1-tailed Sig:

	LP10401	LP0104	LP0101	LP0102	LP0107	LP0108
LP10401	1,000	,486	,471	,482	,416	,578
	,	,000	,000	,000	,000	,000
LP0104	,48	1,000	,293	,426	,288	,656
	,000	,	,000	,000	,000	,000
LP0101	,471	,293	1,000	,413	,326	,353
	,000	,000	,	,000	,000	,000
LP0102	,482	,426	,413	1,000	,297	,414
	,000	,000	,000	,	,000	,000
LP0107	,416	,288	,326	,297	1,000	,479
	,000	,000	,000	,000	,	,000
LP0108	,578	,656	,353	,414	,479	1,000
	,000	,000	,000	,000	,000	,

*** MULTIPLE REGRESSION ***

Equation Number

Equation Number 1 Dependent Variable.. LP10401 Lebenszufriedenheit gegenwärtig

Descriptive Statistics are printed on Page 25

Block Number 1. Method: Enter
LP0104 LP0101 LP0102 LP0107 LP0108

Variable(s) Entered on Step Number
1.. LP0108 Zufriedenheit Lebensstandard
2.. LP0101 Zufriedenheit Gesundheit
3.. LP0102 Zufriedenheit Arbeit
4.. LP0107 Zufriedenheit Freizeittätigkeit
5.. LP0104 Zufriedenheit Haushaltseinkommen

Multiple R	,68298
R Square	,46646
Adjusted R Square	,46614
Standard Error	1,24002

Analysis of Variance

	DF	Sum of Squares	Mean Square
Regression	5	11234,56968	2246,91394
Residual	8357	12850,11931	1,53765

F = 1461,26735 Signif F = ,0000

---------- Variables in the Equation ----------

Variable	B	SE B	95 % Confdnce Intrvl B		Beta
LP0104	0,93057	,008510	,076375	,109739	,119098
LP0101	,182065	,007563	,167239	,196890	,219475
LP0102	,144234	,007348	,129831	,158638	,186167
LP0107	,091159	,007293	,076864	,105455	,116468
LP0108	,262684	,010641	,241825	,283543	,289121
(Constant)	1,740672	,063572	1,616056	1,865288	

---------- Variables in the Equation ----------

Variable	T	Sig T
LP0104	10,935	,0000
LP0101	24,072	,0000
LP0102	19,629	,0000
LP0107	12,500	,0000
LP0108	24,686	,0000
(Constant)	27,381	,0000

2. Interpretation der Ergebnisse

Der Output liefert die Angaben, die zur Interpretation einer Regressionsschätzung gebraucht werden. Wir erhalten zuerst eine Aufstellung der benutzten Variablen und deren univariate Statistiken (Mittelwert und Standardabweichung), gefolgt von der Anzahl der einbezogenen Fälle. Diese Angaben haben zwar keine direkte Bedeutung für die Schätzung, dennoch dienen sie zur Überprüfung der Fallzahlen sowie der benutzten Variablen, ein Problem, das sich insbesondere bei Längsschnittuntersuchungen bzw. bei Variablen mit kodierten Namen ergibt.

Es folgt die einseitige Korrelationsmatrix, an der schon im Vorfeld das Phänomen der Multikollinearität identifiziert werden kann. In unserem Fall liegen die Korrelationen zwischen .28 und .66. Die Größenordnung ist beachtlich. Inhaltlich liegen die Erklärungen allerdings auf der Hand (wer zufrieden mit seinem Lebensstandard ist, ist wahrscheinlich auch mit dem Haushaltseinkommen zufrieden, da die Höhe des Haushaltseinkommen ein wichtiger Punkt bei der Befriedigung materieller Bedürfnisse ist).

Bezogen auf die Modellgüte lässt sich sagen, dass die **erklärte Varianz**, d.h. der Anteil an der Gesamtvarianz, der durch die ausgewählten Variablen erklärt wird, weiter unten im Output als *R Square* (Bestimmtheitsmaß) ausgedrückt wird und 46,6 % beträgt. Das von uns gewählte Modell erklärt also fast die Hälfte der Gesamtstreuung. Das ist eine relativ gute Schätzung, auch wenn fünf Prädiktoren verwendet werden. Demnach ist das *Adjusted R Square*, das sich verkleinert, wenn zu viele Prädiktoren in die Schätzung einfließen, nur um .00032 kleiner als das *R Square*.

Die *Analysis of Variance* ergibt einen signifikanten Erklärungsbeitrag für das ganze Modell. Der empirische Signifikanzwert von F ist kleiner als die übliche Irrtumswahrscheinlichkeit von 5 % (0,05).

Bis hierhin haben wir ausschließlich Informationen über die Güte und Anpassung des ganzen Modells erhalten, ohne auf die Besonderheiten der einzelnen Prädiktoren einzugehen. Der folgende Teil von Output 4 bezieht sich nun gerade auf die Stellung und die Funktion der unabhängigen Variablen. Die Zeilen stellen die einzelnen Prädiktoren dar und die Spalten die Regressionskoeffizienten („B“), Standardfehler der Regressionskoeffizienten („SE B“), Konfidenzintervall für die Irrtumswahrscheinlichkeit von 5 % („95 % Confidence Intrv. B“), standardisierte Regressionskoeffizienten („Beta“), T-Werte („T“) und die Signifikanz der T-Werte („Sig T“).

Auf den ersten Blick fällt auf, dass alle unabhängigen Variablen dem Signifikanzniveau von 0,05 genügen, wir es also mit signifikanten Effekten zu tun haben. Denn die Spalten „T“ und „Sig T“ beschreiben die Werte zur Überprüfung der Nullhypothese, dass der Regressionskoeffizient b_i gleich Null ist, mit dem zugehörigen Signifikanzniveau, das in allen Fällen kleiner als 0,05 ist. Die gleiche Information erhalten wir, wenn wir uns die Konfidenzintervalle ansehen. Auch sie beinhalten in keinem Fall bei 5%iger Irrtumswahrscheinlichkeit den Wert Null, was zur Ablehnung der Nullhypothese, dass die Regressionskoeffizienten gleich Null sind, führt.

Die Regressionskoeffizienten liefern Informationen über die relative Einflussstärke eines jeden Prädiktors. Wenn sich die Verteilung dieser Werte als eine zufällige erweist, ist der Standardfehler der Regressionskoeffizienten („SE B“) eine Schätzung der Standardabweichung dieser Zufallsvariablen.

In den meisten Anwendungsfällen sind die standardisierten Regressionskoeffizienten wichtig als Anhaltspunkte für die relative **Effektstärke** der Prädiktoren. Denn die unstandardisierten Regressionskoeffizienten B_i sind abhängig vom Maßstab der Variablen. In unserem Beispiel ergeben sich durch die Standardisierung für alle Regressionskoeffizienten nur geringe Veränderungen, was darauf zurückzuführen ist, dass alle Variablen auf der gleichen Skala gemessen wurden.

Betrachten wir aus diesen Gründen nun die standardisierten Regressionskoeffizienten $Beta_i$. Wir erkennen, dass die Zufriedenheit mit dem Lebensstandard den größten Effekt auf die allgemeine Lebenszufriedenheit hat. Die Zufriedenheit mit der Gesundheit weist den zweitstärksten Effekt auf die allgemeine Lebenszufriedenheit auf. Den dritten Platz nimmt die Zufriedenheit mit der Arbeit ein und am Ende der Einflussfaktoren stehen die Zufriedenheiten mit Freizeittätigkeit und Haushalteinkommen.

Wie verhalten sich diese empirischen Ergebnisse nun zu den oben formulierten Hypothesen?

Da fast die Hälfte der Varianz der allgemeinen Lebenszufriedenheit durch die fünf Prädiktoren erklärt wird, ist Hypothese a) bestätigt. Die Vermutung liegt nahe, dass bei der Hinzunahme von weiteren Bereichszufriedenheiten die Modellgüte (*R Square*) weiter steigt.

Hypothese b) ist teilweise anzunehmen und teilweise abzulehnen. Der stärkste Prädiktor ist die Zufriedenheit mit dem Lebensstandard, sie ist damit am wichtigsten für die Erklärung der allgemeinen Lebenszufriedenheit. Oder anders gesagt, Menschen, die zufrieden mit ihrem Lebensstandard sind, sind auch eher allgemein mit ihrem Leben zufrieden. Allerdings hat die Zufriedenheit mit dem Haushalteinkommen nicht die formulierte Wirkung auf die allgemeine Lebenszufriedenheit. Hier kann vermutet werden, dass sowohl Lebensstandard als auch Haushalteinkommen eine gemeinsame Schnittmenge besitzen, also in Teilen die gleiche Erklärungsfunktion erfüllen. Das würde die hohe Korrelation von .66 erklären. Andererseits sind sie nicht vollkommen linear abhängig, was auf Unterschiede hinweist.

Hypothese c) kann als bestätigt angesehen werden, da der zugehörige Prädiktor mit .19 einen mittleren, aber hochsignifikanten Effekt aufweist.

Die Hypothese d) und e) sind schwach bestätigt. Sowohl Zufriedenheit mit Freizeittätigkeit als auch Zufriedenheit mit Gesundheit haben einen Einfluss auf die allgemeine Lebenszufriedenheit. Es kann nun anhand der Ergebnisse der multiplen Regression gezeigt werden, wie groß der relative Einfluss dieser Variablen ist. Die Zufriedenheit mit der Gesundheit liegt mit .22 hinter der Zufriedenheit mit dem Lebensstandard. Ein Effekt, der so nicht in unseren Hypothesen zu finden ist. Die Annahme, die Zufriedenheit mit dem Haushalteinkommen habe einen starken Einfluss auf die Lebenszufriedenheit, lässt sich ebenfalls mit den Daten nicht halten. Sie hat zwar einen Einfluss, aber die anderen Einflüsse sind stärker.

3.1.14 Die multivariate Regression

In der multiplen Regression betrachtet man *eine* Variable y in Abhängigkeit von l unabhängigen Variablen $x_1,...,x_l$. In der *multivariaten* Regression werden *k abhängige* Variable y_1, ..., y_k berücksichtigt. Im Unterschied zur kanonischen Korrelation (vgl. Kapitel 1) werden die Regressionen von y_i auf $x_1,...,x_l$ alle *einzeln* behandelt.

Ist y_i der (n, 1)-Vektor $\begin{pmatrix} y_{i1} \\ \vdots \\ y_{in} \end{pmatrix}$, X die (n, l)-Matrix $(x_1,...,x_l)$ und β_i der $(l,1)$-Koeffizientenvektor $\begin{pmatrix} \beta_{i1} \\ \vdots \\ \beta_{il} \end{pmatrix}$, so ist der Koeffizientenvektor β_i für die Regression von y_i auf $x_1,...,x_l$ gegeben durch:

$$\beta_i = (X'X)^{-1}(X'y_i) = R^{-1}{}_{xx}\, r_{xy_i},\ wobei\ R_{xx} = \frac{(X'X)}{n},\ r_{xy_i} = \frac{X'y_i}{n}.$$

Alle k Regressionslösungen lassen sich zusammenfassend formulieren:

$$\underset{(l,\,k)}{\beta} = \underbrace{R_{xx}^{-1}}_{(l,\,l)}\underbrace{R_{xy}}_{(l,\,k)},\ \text{wobei}\ R_{xy} = \frac{X'Y}{n}.$$

3.1.15 „Weighted least squares“

Bei inferenzstatistischen Überlegungen zur multiplen Regression geht man von folgendem Ansatz aus:

$$y_j = \sum_{i=1}^{k} \beta_i x_{ji} + u_j \qquad \text{(für j = 1, ..., n).}$$

Über die Fehlerterme u_j der Beobachtungseinheiten (z.B. Individuen) werden folgende Annahmen gemacht: Die Fehlerterme müssen erstens statistisch unabhängig voneinander sein (für verschiedene Einheiten) und zweitens müssen die Mittelwerte aller Fehlerterme Null und die Varianzen aller Fehlerterme gleich sein.

Zusammenfassend kann man mit Hilfe der *Kovarianzmatrix* (= Matrix der Kovarianzen) der u_j (mit j = 1, ..., n) formulieren:

$E(u) = 0,\ E(uu') = \sigma^2 E_n = \begin{pmatrix} \sigma^2 & & 0 \\ & \ddots & \\ 0 & & \sigma^2 \end{pmatrix}$, wobei E_n die n-reihige *Einheitsmatrix* $\begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix}$ ist und

$u = \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix}$ der Vektor der Fehler für die verschiedenen Einheiten.

Diese Bedingungen lassen sich entsprechend für y formulieren:

Die y_j (j = 1, ..., n) sind unabhängige Zufallsvariablen mit Erwartungswert $\sum_{i=1}^{k} \beta_i \, x_{ij}$ und gleicher Varianz σ^2. In Matrixschreibweise:

$$E(Y) = X\beta, s_{yy} = E\left[(y-\bar{y})(y-\bar{y})'\right] = \sigma^2 \, E_n, \text{wobei } y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, X = \begin{pmatrix} x_{11} \cdots x_{1k} \\ \vdots \quad \vdots \\ x_{n1} \ldots x_{nk} \end{pmatrix}.$$

Die Methode der kleinsten Quadrate $\left(u'u = \sum_{j=1}^{n} u_j^2 \Rightarrow \min. \right)$ liefert die Lösung:

$$\hat{\beta} = (X'X)^{-1} X'y \, .$$

Die Lösung erhält man auch durch die Maximum Likelihood-Schätzung, wenn man Normalverteilung der Fehler voraussetzt.
Bezeichnet S die Kovarianzmatrix der Fehler u_j (j = 1, ..., n) bzw. der y_j, so ist die Dichte der multivariaten Normalverteilung:

$$f(u) = \frac{1}{\left(\sqrt{2\pi}\right)^n \sqrt{\det(S)}} e^{-\frac{1}{2} u' S^{-1} u}$$

Die Likelihoodfunktion L (β, σ^2) wird als Funktion von β maximiert, wenn man $u' S^{-1} u = (y - X\beta)' S^{-1} (y - X\beta)$ minimiert. Da $S = \sigma^2 \cdot E_n$, so ist also $\sigma^2 (y - X\beta)' (y - X\beta)$ zu minimieren als Funktion von β, wobei man σ^2 vernachlässigen kann. Dies ist aber genau der Ansatz der Methode der kleinsten Quadrate, weshalb die ML-Schätzung ebenfalls liefert:

$$\hat{\beta} = (X'X)^{-1} X'y \, .$$

Dabei ist $\hat{\beta}$ normalverteilt mit:

Erwartungswert $E(\hat{\beta}) = \beta$ (Vektoren!)

Kovarianzmatrix $S_{\hat{\beta}\hat{\beta}} = \sigma^2 (X'X)^{-1}$

Die ML-Schätzung von σ^2 ist:

$$\hat{\sigma}^2 = \frac{1}{n} (y - \hat{y})' (y - \hat{y}), \text{ wobei } \hat{y} = X\hat{\beta} \, .$$

Einer erwartungstreue Schätzung ist $\hat{\sigma}^2 = \frac{1}{n-k} (y - \hat{y})' (y - \hat{y})$.

Die Methode der kleinsten Quadrate von Gauss (1809) und Markoff (1990) bzw. Fischer (1921) ML-Schätzung, die im Falle der Normalverteilung die gleiche Lösung liefert, sind von Aitken (1934) auf den Fall angewendet worden, dass eine beliebige (invertierbare) Kovarianzmatrix S vorliegt. Z.B.: Ungleiche Varianzen der Fehler oder korrelierte Fehler.

Die verallgemeinerte „Summe der Abweichungsquadrate“ lautet in diesem Fall:

$u' S^{-1} u = (y - X\beta)' S^{-1} (y - X\beta) = Y' S^{-1} y + \beta' X' S^{-1} X\beta - 2\,\beta' X' S^{-1} y$

Will man dies als Funktion von β minimieren, so berechnet man:

$$0 = \frac{\sigma u' S^{-1} u}{\sigma \beta'} = 2X' S^{-1} X\beta - 2X' S^{-1} y$$

Also: $\hat{\beta} = \left(X' S^{-1} X\right)^{-1} X' S^{-1} y$

Setzt man wieder voraus, dass der Fehlervektor u der multivariaten Normalverteilung folgt, so ist das Maximieren der Likelihoodfunktion L = ln f wieder äquivalent zu dem Minimieren von $(u' S^{-1} u)$, weshalb die ML-Methode das gleiche Ergebnis liefert.

Somit ist $\hat{\beta}$ dann normalverteilt mit:

Erwartungswert $E\left(\hat{\beta}\right) = \beta$ (Vektoren!)

Kovarianzmatrix $S_{\hat{\beta}\hat{\beta}} = \left(X' S^{-1} X\right)^{-1}$

Unterscheiden sich nur die Varianzen der Fehler (d.h. sind die Fehler unabhängig und nur die Varianzen verschieden), so ist:

$$S = \begin{pmatrix} \sigma_1^2 & 0 \\ \ddots & \\ 0 & \sigma_n^2 \end{pmatrix} \text{ und } S^{-1} = \begin{pmatrix} \sigma_1^{\frac{1}{2}} & 0 \\ \ddots & \\ 0 & \sigma_n^{\frac{1}{2}} \end{pmatrix}$$

In diesem Fall ist die Funktion, die minimiert wird, einfach:

$$u' S^{-1} u = \sum_{j=1}^{n} \frac{u_j^2}{\sigma_j^2}$$

D.h. die Fehler werden standardisiert, ihre Größenordnung wird vergleichbar gemacht.

In dem speziellen Fall, dass es genau so viele Untersuchungseinheiten wie zu schätzende Regressionskoeffizienten gibt (dies wird für saturierte Modelle in der Varianzanalyse gelten), führen OLS (Ordinary Least Squares) und WLS (Wighted Least Squares) zum gleichen Ergebnis. Denn in diesem Fall ist X eine quadratische Matrix. Falls der Regressionsansatz lösbar ist, so ist X invertierbar:

OLS: $\hat{\beta} = (X'X)^{-1} X' y = X^{-1} (X')^{-1} X' y = X^{-1} y$

WLS: $\hat{\beta} = \left(X' S^{-1} X\right)^{-1} X' S^{-1} y = X^{-1}\, S\left(X'\right)^{-1} X' S^{-1} y = X^{-1} y$

In diesem Fall führen OLS und WLS also zum gleichen Ergebnis.

Literaturverzeichnis

Aitken, A.C., 1934: *On least squares and linear combinations of observations*. In: Proc. Royal Soc. Edin. A 55: 42-48.

Anderson, T.W., 2003[3]: *An Introduction to Multivariate Statistical Analysis*. New York: Wiley InterScience.

Backhaus, K., Erichson, B., Plinke, W., Weiber, R., 2008[12]: *Multivariate Analysemethoden*. Eine anwendungsorientierte Einführung. Berlin: Springer.

Blalock, H.M., Blalock, A.B., 1968: *Methodology in social research*. New York: McGraw-Hill.

Brosius, F., 2004: *SPSS 12*. Bonn: mitp.

Goldberger, A.S., 1964: *Econometric theory*. New York: Wiley.

Holm, K., 1977: *Lineare multiple Regression und Pfadanalyse*. In: Ders. (Hg.): Die Befragung. Band 5. München: UTB.

Johnston, J., 1996[4]: *Econometric Methods*. New York: McGraw-Hill/Irwin.

Kerlinger, F.N., Pedhazur, E.J., 1973: *Multiple regression in behavioral research*. New York: Holt.

Küchler, M., 1979: *Multivariate Analyseverfahren*. Stuttgart: Teubner.

Mayntz, R., Holm, K., Hübner, P., 1978[5]: *Einführung in die Methoden der empirischen Soziologie.* Opladen: Westdeutscher Verlag.

Mood, Carina: Logistic regression: Why we cannot do what we think we can do, and what we can do about it. In: European Sociological Review 26, 2010: 67-82.

Nie, N.H. et al., 1975[2]: *Statistical package for the social sciences* (SPSS). New York: McGraw-Hill.

Opp, K.-D., Schmidt, P., 1976: *Einführung in die Mehrvariablenanalyse*. Grundlagen der Formulierung und Prüfung komplexer sozialwissenschaftlicher Aussagen. Reinbek bei Hamburg: Rowohlt.

Smillie, K.W., 1966: *An Introduction to Regression and Correlation*. New York: Academic Press.

Van de Geer, J.P., 1971: *Introduction to Multivariate Analysis for the Social Sciences*. San Francisco: Freeman.

Wonnacott, R.J., Wonnacott, T.H., 1979[2]: *Econometrics*. New York: Wiley.

Wonnacott, Th.H., Wonnacott, R.J., 1990[4]: *Introductory Statistics for Business and Economics*. New York: Wiley.

3.2 Pfadanalyse

Die Pfadanalyse ist von Sewall Wright (1934, 1960) in der Genetik entwickelt worden. Sie wurde im soziologischen Bereich zuerst von Otis D. Duncan (1966) angewendet. Die Pfadanalyse besteht in der wiederholten Anwendung der multiplen Regression in dem Fall, dass eine (schwache) kausale Ordnung der Variablen x_1, ..., x_k vorausgesetzt wird. Dies betrifft oft eine Zeitordnung oder den Grad der Veränderbarkeit, darf jedoch nicht darauf beschränkt werden.

3.2.1 Ein klassisches Beispiel von Blau und Duncan

Blau/ Duncan (1967) haben die Pfadanalyse in den Sozialwissenschaften bekannt gemacht durch eine Untersuchung des „status attainment process“ in den USA. Dabei verwendeten sie folgende Variablen zur Abbildung des sozialen Status:

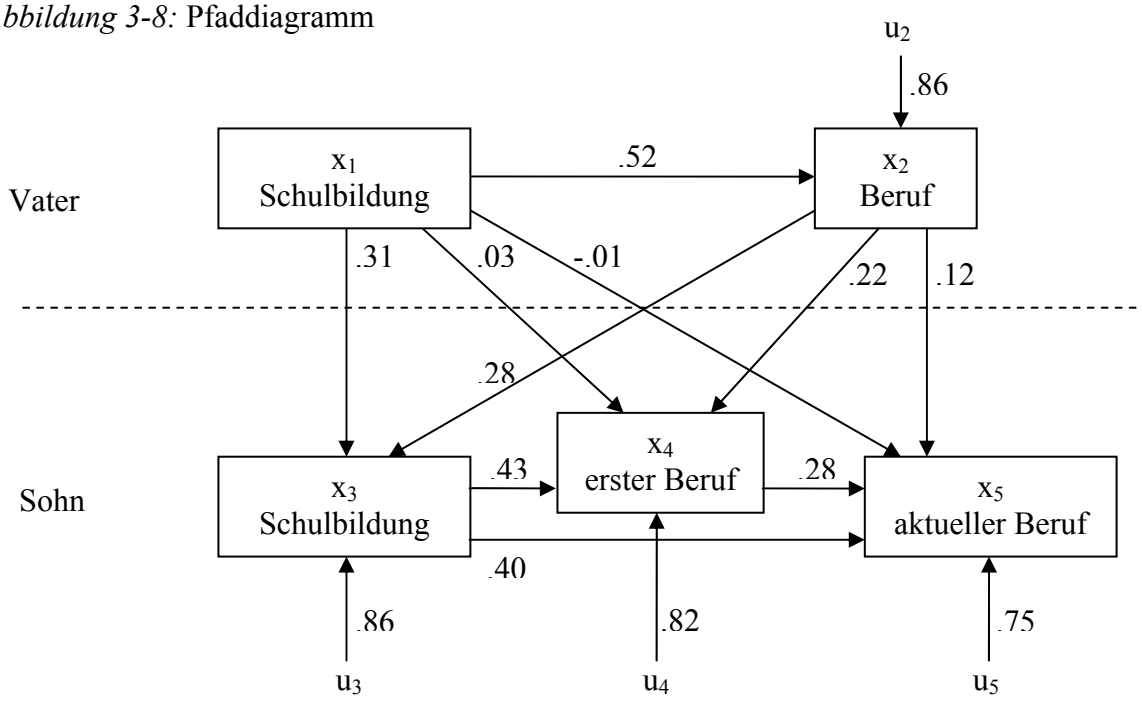
x_1 = Schulbildung des Vaters
x_2 = Berufsstatus des Vaters
x_3 = Schulbildung des Sohnes
x_4 = Berufsstatus der ersten Beschäftigung des Sohnes
x_5 = Berufsstatus der späteren Beschäftigung des Sohnes

Die abhängigen Variablen berechnen sich nach folgenden Regressionsgleichungen:

$x_2 = .52\ x_1 + .86\ u_2$
$x_3 = .31\ x_1 + .28\ x_2 + .86\ u_3$
$x_4 = .03\ x_1 + .22\ x_2 + .43\ x_3 + .82\ u_4$
$x_5 = -.01\ x_1 + .12\ x_2 + .40\ x_3 + .28\ x_4 + .75\ u_5$

Im folgenden Pfaddiagramm (Abbildung 3-8) kann u.a. gezeigt werden, dass der Vater sowohl hinsichtlich seiner Schulbildung als auch seines Berufsstatus den Status des Berufs des Sohnes weniger stark beeinflusst als die Schulbildung des Sohnes selbst.

Abbildung 3-8: Pfaddiagramm



Die Regressionsschätzung der jeweiligen abhängigen Variablen aufgrund der Prädiktoren (die Variablen, die kausal vorangehen) wird in dem Diagramm dadurch dargestellt, dass jeweils ein Pfeil von jedem Prädiktor zur abhängigen Variablen verläuft, charakterisiert durch die β - Koeffizienten als Einflusskoeffizienten. Der durch die Regressionsschätzungen nicht erklärte Rest wird jeweils den exogenen Fehlertermen u zugeschrieben, was ebenfalls in Form eines Pfeiles dargestellt wird.

3.2.2 Kausale Ordnung und Rekursivität

Kausale Ordnung

Eine Ordnungsrelation für eine Menge von Einheiten ist zunächst dadurch charakterisiert, dass für je zwei Elemente a und b eindeutig feststehen muss, ob entweder a kleiner b (im Sinne der Ordnung) oder b kleiner a oder a gleich b. Ferner muss eine Ordnungsrelation die Bedingungen der Nicht-Reflexivität, Asymmetrie und Transitivität erfüllen.

Eine („starke“) kausale Ordnung von Variablen wäre analog zu definieren. Aber: Für den speziellen Anwendungszusammenhang von Kausalmodellen ist es günstiger, die Anforderungen an eine kausale Ordnung abzuschwächen, um die Anwendungsmöglichkeiten zu erhöhen.

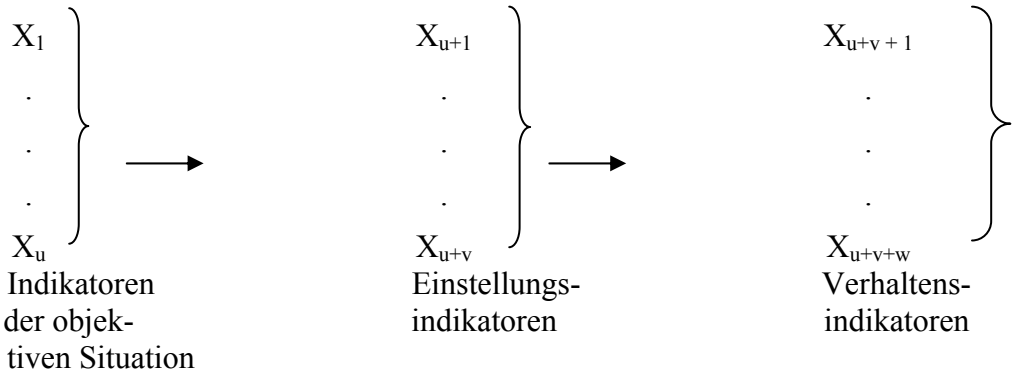
(Schwache) kausale Ordnung von Variablen

In der Reihenfolge der Nummerierung vorhergehende Variablen können einen (Kausal-) Effekt auf in der Reihenfolge folgende Variablen haben, müssen aber nicht.

Ferner: Es gibt keine Wechselwirkungen bzw. allgemeiner keinen Effekt gegen diese Reihenfolge.

(In der Reihenfolge nachfolgende Variable dürfen also nicht als Ursachen modelliert werden.)

Abbildung 3-9: Beispiel für die kausale Ordnung von Indikatoren



Innerhalb der drei Gruppen von Indikatoren werden keine kausalen Einflüsse unterstellt, sondern nur *zwischen* den Indikatoren verschiedener Gruppen (zusammenfassend durch nur einen Pfeil zwischen den Gruppen symbolisiert). **Rekursivität** bedeutet, dass keine Rückkopplungen oder reziproke Verursachungen vorliegen. Dies wäre im o.g. Modell der Fall, wenn z.B. das Verhalten auf die Einstellungen (kausal) zurückwirken würde.

3.2.3 Vollständiges Modell und unvollständiges Modell

Bei der Pfadanalyse werden aufgrund theoretischer Gesichtspunkte jene Verknüpfungen aus der sozialen Realität zusammengefügt, deren Auftreten und Reihenfolge Voraussagen über die Veränderungen von Variablen aufgrund anderer Variablen ermöglichen.

Geht x_1 der Variablen x_2 kausal voran und gehen x_1, x_2 der Variablen x_3 kausal voran, allgemein x_1, ..., x_{i-1} der Variablen x_i kausal voran (i = 2, ..., k), so kann man allgemein x_i durch Regression auf x_1, ..., x_{i-1} schätzen:

$$\hat{x}_2 = f(x_1)$$
$$\hat{x}_3 = f(x_1, x_2)$$

Während in der schrittweisen multiplen Regression ein und dieselbe Variable als abhängige Variable $\hat{y} = f(x_1), \hat{y} = f(x_1, x_2)$ etc. betrachtet wird, ändert sich hier die abhängige Variable.

Allgemein (für i = 2, ..., k) gilt:

$$\hat{x}_1 = f(x_1, x_2, ..., x_{i-1})$$

Für sämtliche Variablen (i = 2, ..., k) werden die Werte der jeweils als abhängig betrachteten Variablen aufgrund der kausal vorangehenden Variablen geschätzt.

Dies ist das **vollständige Modell** der Pfadanalyse für die Variablen x_1, ..., x_k. Weil die Pfadanalyse auf Regressions- und Korrelationsrechnungen beruht, setzen die verwendeten Variablen Intervallskalenniveau voraus. Die Anwendbarkeit lässt sich dadurch erhöhen, dass die Regressionsschätzung jeder abhängigen Variablen x_i sich nur auf *die* Variablen x_j $\left(j \in \{1,...,i-1\}\right)$ erstreckt, die nach der Modellannahme einen kausalen Einfluss auf x_i haben (i = 2, ..., k). Es handelt sich dann um das **unvollständige Modell**, weil nicht jeweils alle einer abhängigen Variablen x_i kausal vorangehenden Variablen x_j $\left(j \in \{1,...,i-1\}\right)$ nach den Modellannahmen eine kausale Wirkung auf x_i haben (i = 2, ..., k). Im folgenden soll zunächst das vollständige Modell behandelt werden.

In der Pfadanalyse wird wegen der größeren Übersichtlichkeit mit den Regressionskoeffizienten der *standardisieren Variablen* gearbeitet, wobei die Zurückrechnung auf nicht standardisierte Variablen ja immer möglich ist. Diese standardisierten Regressionskoeffizienten werden **Pfadkoeffizienten** genannt und mit p_{ij} bezeichnet, wobei i der Index der abhängigen Variablen und j der Index der unabhängigen Variablen ist (i = 2, ..., k; j = 1, ..., i - 1).

Das allgemeine Erklärungsmodell in der Pfadanalyse lässt sich dann formulieren als:
$\hat{x}_i = p_{i,1}x_1 + p_{1,2}x_2 + ... + p_{i,i-1}x_{i-1}$ (für i = 2, ..., k)

Für j = 1, ..., i-1 (i = 2, ..., k) ist $x_i - \hat{x}_i$ orthogonal zu (unabhängig von) x_j, d.h.:
$< x_i - \hat{x}_i, x_j >= 0$ oder $< x_i, x_j >=< \hat{x}_i, x_j >$

Also erhält man:

$$r_{ij} = \frac{< x_i, x_j >}{n} = \frac{< \hat{x}_i, x_j >}{n} = p_{i1}\frac{< x_1, x_j >}{n} + ... + p_{i,i-1}\frac{< x_{i-1}, x_j >}{n} = p_{i1}r_{1,j} + ... + p_{i,i-1}r_{i-1,j}$$

Die Gleichung $r_{ij} = \sum_{l-1}^{i-1} p_{il} r_{lj}$ gilt als „Grundgleichung“ oder **Fundamentaltheorem** der Pfadanalyse. Dies ist natürlich nichts anderes als die übliche Gleichung für die Berechnung der Regressionskoeffizienten aus den Korrelationskoeffizienten.

Somit setzt die Pfadanalyse die Regressionsanalyse voraus. Wegen der vorgängigen Standardisierung gehen alle Werte nicht als Originalwerte, sondern als Abweichungen von ihrem Mittelwert in die Analyse ein.

Mittels Matrixschreibweise und –rechnung lassen sich die Bestandteile der Regressionsgleichungen folgendermaßen ausdrücken.

Die Korrelationsmatrix von x_1, ..., x_j (für j = 1, ..., k) lautet:

$$R_j = \begin{pmatrix} 1 & \dots & r_{1j} \\ r_{j1} & \dots & 1 \end{pmatrix}$$

Der Vektor r_i ist der Vektor der Korrelationen von x_i mit den kausal vorangehenden Variablen:

$$r_i = \begin{pmatrix} r_{i,1} \\ \vdots \\ r_{i,i-1} \end{pmatrix}$$

p_i sei der Vektor der Pfadkoeffizienten der Pfade der kausal vorangehenden Variablen nach x_i, d.h. der Vektor der Regressionskoeffizienten von x_i auf x_1, ..., x_{i-1}:

$$p_i = \begin{pmatrix} p_{i,1} \\ \vdots \\ p_{i,i-1} \end{pmatrix}$$

Dann lauten die angegebenen Gleichungen in Matrixschreibweise (entsprechend der Regressionsanalyse):

$r_i = R_{i-1} \cdot p_i$ (Anzahl der Ausprägungen: (i – 1, 1), (i – 1, i – 1), (i – 1, 1))

Wie in der üblichen Regression: $p_i = R_{i-1}^{-1} \cdot r_i$ (für i = 2, ..., k)

Da in der Realität vieles mit vielem kausal zusammenhängt, sind Modelleinschränkungen notwendig. Eine erste Einschränkung betrifft die Definition **exogener und endogener Variablen**. Eine Variable, auf die keine andere Variable einwirkt, heißt exogen. In Pfaddiagrammen findet man sie häufig ganz links. Ihre Werte werden nicht durch das Pfadmodell erklärt, sondern werden im Modell als gegeben betrachtet. Endogene Variablen sind diejenigen, auf die eine oder mehrere andere Variable des Pfadmodells einwirken. Diese anderen Variablen können sowohl exogene als auch endogene Variablen sein. Im Unterschied zur Regressionsanalyse kann also in der Pfadanalyse ein und dieselbe Variable einmal abhängige und ggfs. Mehrmals unabhängige Variable innerhalb eines Modells sein.

Die zweite wichtige Einschränkung befasst sich mit der Frage, wie diejenigen Kausalwirkungen behandelt werden, die zwar auf die endogenen Modellvariablen einwirken, jedoch selbst nicht Bestandteil des Modells sind. Solche Wirkungen werden dann einfach sogenannten **Rest- bzw. Residualvariablen u** zugeschrieben, wobei inhaltlich nicht präzisiert wird, um welche Variablen es

sich konkret handelt. Modelliert wird allein die Stärke der Wirkung der Residualvariablen auf jeweils eine bestimmte endogene Variable.
Es ist üblich, jeweils den nicht erklärten Rest $x_i - \hat{x}_i$ der abhängigen Variablen x_i (i = 2, ..., k) einer exogenen Fehler-Variablen u zuzuschreiben: $x_i - \hat{x}_i = p_{i,u} u_i$

Also folgt: $\frac{< x_i - \hat{x}_i, x_i - \hat{x}_i >}{n} = p_{i,u}^2 \frac{< u_i, u_i >}{n}$, wobei: $\frac{< u_i, u_i >}{n} = 1$

Somit ist $p_{i,u} = \sqrt{1 - R^2_{x_i \cdot x_1, ..., x_{i-1}}}$ (für i = 2, ..., k).

Aus der Schreibweise $x_i - \hat{x}_i = p_{i,u} u_i$ ist ersichtlich, dass u_i orthogonal (unabhängig) ist zu allen Prädiktoren x_1, ..., x_{i-1}, d.h. auch unkorreliert mit diesen Prädiktoren ist (für i = 2, ..., k). Also ist u_i auch orthogonal zu u_2, ..., u_{i-1}, da dies nur Linearkombinationen von x_1, ..., x_{i-1} sind, d.h. u_i ist auch unkorreliert mit allen Residualvariablen u_2, ..., u_{i-1} (für i = 3, ..., k). Der nicht erklärte Rest wird also immer einer mit den kausal vorangehenden Variablen unkorrelierten exogenen Variablen zugeordnet. Dadurch ergibt sich keinerlei zusätzlicher Erkenntnisgewinn über die übliche Regressionsanalyse hinaus, jedoch steht dann in dem Modell (und der graphischen Darstellung, dem Pfeildiagramm) neben den Effektkoeffizienten der kausal vorangehenden Variablen der Effekt der Residualvariablen, woraus i.a. ersichtlich ist, dass das Modell weniger erklärt als das, was noch nicht in dem Modell erfasst ist.

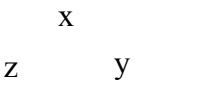
Dieser Rest besagt auch, dass die Abhängigkeit nicht deterministisch erklärt werden kann.[8]

3.2.3.1 Kausale Geschlossenheit eines Modells gegenüber weiteren Einflussfaktoren

Geht z der Variablen x kausal voran und x der Variablen y und unterscheidet man nur: vorhandener Kausaleinfluss versus nicht vorhanden, so gibt es für diese 3 Variablen $2^3 = 8$ mögliche Kausalstrukturen. (Allgemein: Für k Variablen mit gegebener kausaler Ordnung gibt es 2^k mögliche Kausalstrukturen.)

Abbildung 3-10:

Fall 1

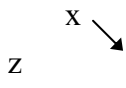


Keine Beziehung.

Fall 2

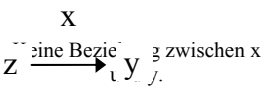
Keine Beziehung zwischen x und y.

Fall 3



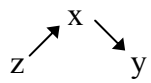
Die Einführung von z modifiziert die Beziehung zwischen x und y nicht („Bestätigung“).

Fall 4



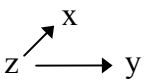
Keine Beziehung zwischen x und y.

Fall 5



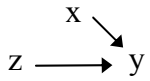
Die Kovariation von x und y wird durch z nicht geändert.

Fall 6



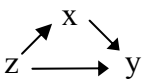
„Scheinkorrelation“ zwischen x und y.

Fall 7



Die Kovariation von x und y wird durch z nicht geändert.

Fall 8



Die Kovariation zwischen x und y ist zum Teil nicht auf die Kausalbeziehung zwischen x und y zurückzuführen, sondern auf den Einfluss von z.

---

[8] Hieraus folgt auch:
$p_{i,u} = p_{i,u} \frac{< u_i, u_i >}{n} = \frac{< x_i - \hat{x}_i, u_i >}{n} = \frac{< x_i, u_i >}{n} = r_{x_i, u_i}$ (weil $u_i \perp x_j, j = 1, ..., i-1$)
Der Koeffizient $p_{i,u}$ der Fehlervariablen ($x_i = \hat{x}_i + p_{i,u} u_i$) ist also gleich dem Korrelationskoeffizient der zu erklärenden und der Fehlervariablen.

Beschränkt man das betrachtete Modell auf die Merkmale x und y, so lässt sich diskutieren, welche Auswirkungen die Nicht-Berücksichtigung weiterer Einflussfaktoren – in diesem Fall z – auf die Angemessenheit der Modellbildung haben könnte. In Fall 6 und 8 ist die Kovariation von x und y nicht geschlossen gegenüber dem weiteren Einflussfaktor z (vgl. auch Nie et al.: 385). Die Pfadanalyse setzt kausale Geschlossenheit des Modells voraus, denn die Residualvariable u_i jeder abhängigen Variablen x_i darf nicht mit den kausal vorangehenden Variablen x_j und ihren Residuen u_j korrelieren (i = 2, ..., k; j = 1, ..., i – 1; wobei $u_1 := 0$).

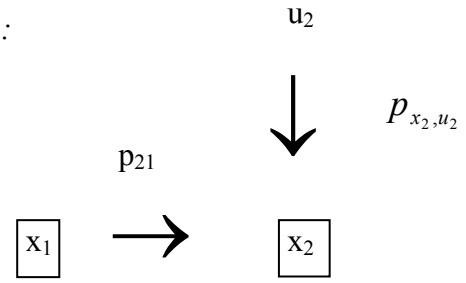
3.2.3.2 Pfaddiagramm und Effekte für das vollständige Modell mit 2,3 und 4 Variablen

Ein kausaler Einfluss von x_1 und x_2 und von x_2 auf x_3 lässt sich durch den Pfad $x_1 \rightarrow x_2 \rightarrow x_3$ graphisch darstellen, wobei eine Verbindung zweier Variablen durch eine Gerade einen Kausalzusammenhang symbolisiert und die Richtung des Pfeils die Richtung der Kausalwirkung angibt. Die Pfadanalyse ist eine Synthese aus Gleichungssystemen (mittels Regression gewonnen) und Kausalstruktur (graphisch dargestellt durch ein Pfeildiagramm). Jede abhängige Variable wird durch Regression auf *die* kausal vorangehenden Variablen geschätzt, von denen ein Pfeil (d.h. ein kausaler Einfluss) zur abhängigen Variablen verläuft. Bei dem vollständigen Modell, das zunächst behandelt wird, sind dies alle einer abhängigen Variablen kausal vorangehenden Variablen.

(1) Einfachster Fall: 2 Variablen

Wie in der folgenden Abbildung zu erkennen ist, geht von der Variablen x1 ein kausaler Einfluss mit der Stärke p21 auf x2 aus, wobei x2 auch von weiteren Einflüssen abhängig ist, die als Residuen u2 mit der Stärke p_{x_2,u_2} in die Berechnung eingehen. Die Größe $p^2_{x_2,u_2}$ ist also die durch x1 nicht erklärte Varianz der endogenen Variablen x2.

Abbildung 3– 11:

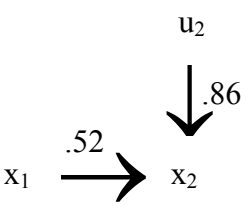


Hierbei ist der Pfadkoeffizient p_{21} gleich dem Korrelationskoeffizienten r_{21} und $p_{x_2,u_2} = \sqrt{1 - r_{21}^2}$. Die Kausalbeziehung und ihre Richtung werden durch eine Verbindungslinie und die Pfeilrichtung dargestellt. Der Pfadkoeffizient neben dem Pfeil gibt die Stärke des „direkten Einflusses“ an (β - Koeffizient).

In dem Beispiel von Blau und Duncan:

Regressionsschätzung: $\hat{x}_2 = .52\ \hat{x}_1$; $R^2 = 26\ \%$; $1 - R^2 = 74\ \%$.

$1\text{-}R^2 = .74 = .86^2$



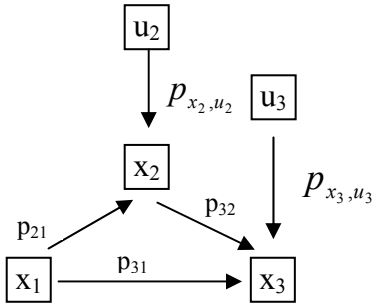
74 % der Variation wird nicht erklärt durch das Modell; diese nicht erklärte Variation wird in der Darstellung der Variablen u_2 (= nicht berücksichtigte Faktoren) zugeordnet.

Erklärte Varianz: $.52^2 = .26$

(2) Vollständiges Modell mit 3 Variablen

Die exogene Variable x_1 wirkt direkt und indirekt auf die endogene Variable x_3. Als **direkter Effekt** wird die Wirkung $x_1 \rightarrow x_3$ bezeichnet. Der **indirekte Effekt** wird durch x_2 vermittelt: $x_1 \rightarrow x_2 \rightarrow x_3$.

Abbildung 3-12:



Bezogen auf die von x_1 direkt ausgehende Effekte erhält man zwei Regressionen:

1. Regression: $p_{21} = r_{21}$ und $p_{x_2,u_2} = \sqrt{1-r_{21}^2}$

2. Regression: $p_{31} = \beta_{31}, p_{32} = \beta_{32}$ und $p_{x_3,u_3} = \sqrt{1-R^2_{x_3.x_1,x_2}}$

Eine Änderung in x_j ($j = 1, ..., i-1$) kann direkt zu einer Änderung von x_i führen oder zunächst zu einer Änderung in x_1 $\left(1 \in \{1,...,i-1\}\right)$ und dadurch schließlich zu einer Änderung in x_i. Der (lineare additive) **kausale Effektkoeffizient C_{ij}** fasst den direkten kausalen Effekt und die indirekten kausalen Effekte über andere Pfade zusammen (Δ bezeichne die Differenz).

$\Delta x_i = C_{ij} \cdot \Delta x_j$

Im Falle von 3 Variablen mit strenger Kausalordnung erhält man folgende Gleichungen, in die jeweils die Variablen eingehen, die einen direkten Einfluss auf die abhängige Variable ausüben:

a) $x_2 = p_{21}x_1 + p_{x_2,u_2}u_2$, also $\Delta_{x_2} = p_{21} \cdot \Delta_{x_1}$

Eine Veränderung Δ_{x_1} bewirkt eine kausale Veränderung $p_{21}\,\Delta_{x_1}$ in x_2, d.h. $C_{21} = p_{21}$.

b) $x_3 = p_{31}\,x_1 + p_{32}\,x_2 + p_{x_3,u_3}\,u_3$, also $\Delta_{x_3} = p_{32} \cdot \Delta_{x_2}$

Eine Veränderung Δ_{x_2} bewirkt direkt eine kausale Veränderung $p_{32}\,\Delta_{x_2}$ in x_3, d.h. $C_{32} = p_{32}$.

Zur Berechnung des kausalen Effekts von x_1 auf x_3 muss man den indirekten kausalen Effekt über den Pfad $x_1 \rightarrow x_2 \rightarrow x_3$ berücksichtigen:

$$\left.\begin{aligned} x_3 &= p_{31}\,x_1 + p_{32}\,x_2 + p_{x_3,u_3}\,u_3 \\ x_2 &= p_{21}\,x_1 + p_{x_2,u_2}\,u_2 \end{aligned}\right\} \quad x_3 = (p_{31} + p_{32}p_{21})x_1 + p_{32}p_{x_2,u_2}u_2 + p_{x_3,u_3}u_3$$

Also: $\Delta_{x_3} = (p_{31} + p_{32}\,p_{21})\Delta_{x_1}$, weshalb: $C_{31} = p_{31} + p_{32}\,p_{21}$.

Der gesamte kausale Effekt zerfällt also in den direkten kausalen Effekt p_{31} und den indirekten kausalen Effekt $p_{32}\ p_{21}$.

In anderen Worten: Der kausale Effektkoeffizient C_{ij} misst den kausalen Effekt in x_i aufgrund von x_j, der kausale Effekt zerfällt in den direkten kausalen Effekt p_{ij} und die Summe der Produkte der Pfadkoeffizienten entlang der Pfade, die von x_i nach x_j führen:

$$\begin{array}{ccccccccc} & p_{j_1,j} & & p_{j_2,j_1} & & p_{j_m,j_{m-1}} & & p_{i,j_m} & \\ x_j & \rightarrow & x_{j_1} & \rightarrow & x_{j_2} \ldots x_{j_{m-1}} & \rightarrow & x_{j_m} & \rightarrow & x_i \end{array}$$

Jeder solche Pfad führt zu einem Produkt: $p_{j_1,j} p_{j_2,j_1} \cdots p_{j_m,j_{m-1}} p_{i,j_m}$. Der indirekte Kausaleffekt ist die Summe dieser Produkte.

Die Zerlegung der Korrelation in dem vollständigen Modell für 3 Variablen (vgl. Nie et al.: 388) stellt die folgende Tabelle zusammenfassend dar.

Tabelle 3-5:

Variablenpaar	Gesamtkorrelation r_{ij} (A)	Direkter Kausaleffekt (B)	Indirekter Kausaleffekt (C)	Gesamter Kausaleffekt C_{ij} (D) = B + C	Nicht kausaler Effekt $r_{ij} - C_{ij}$ (E) = A - D
x_2, x_1	r_{21}	p_{21}	keiner	p_{21}	0
x_3, x_1	r_{31}	p_{31}	$p_{32}p_{21}$	$p_{31} + p_{22}p_2$	0
x_3, x_2	r_{32}	p_{32}	keiner	p_{32}	$r_{32} - p_{32}$
					*
					**

* $r_{y,x_i} = \sum_j r_{x_i,x_j} \cdot \beta_j$ lautet in diesem Fall: $r_{31} = r_{11}\ p_{31} + r_{12}\ p_{32} = p_{31} + r_{12}\ p_{32}$

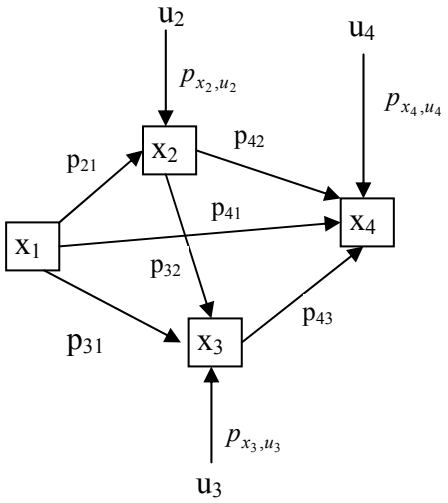
Ferner: $(r_{12} =) r_{21} = p_{21}$ im Falle der einfachen Regression.

** Es handelt sich hier nicht um eine einfache Regression, deshalb braucht p_{32} nicht gleich r_{32} zu sein.

(3) Vier Variablen mit strenger Kausalordnung

Bei drei endogenen und einer exogenen Variablen wird das Kausalmodell bereits komplexer (Abbildung 3-13).

Abbildung 3-13:



Die Zerlegung der Korrelation in dem vollständigen Modell für vier Variablen (vgl: auch Nie et al.: 389) wird in Tabelle 3-6 dargestellt.

Tabelle 3-6: Vollständiges Modell im Fall von 4 Variablen

Variablenpaar	Gesamtkorrelation r_{ij} (A)	Direkter Kausaleffekt (B)	Indirekter Kausaleffekt (C)	Gesamter Kausaleffekt C_{ij} (D) = B + C	Nicht kausaler Effekt r_{ij} - C_{ij} (E) = A - D
x_2,x_1	r_{21}	p_{21}	keiner	p_{21}	0
x_3,x_1	r_{31}	p_{31}	$p_{32}p_{21}$	$p_{31}+p_{22}p_{21}$	0
x_3,x_2	r_{32}	p_{32}	keiner	p_{32}	$r_{32}-p_{32}$
x_4,x_1	r_{41}	p_{41}	$p_{42}p_{21}+$ $p_{43}p_{32}p_{21}+$ $p_{43}p_{31}$	r_{41} (*)	0
x_4,x_2	r_{42}	p_{42}	$p_{43}p_{32}$	$p_{42}+p_{32}p_{43}$	$r_{42}-p_{42}-p_{32}p_{43}$
x_4,x_3	r_{43}	p_{43}	*keiner*	p_{43}	$r_{43}-p_{43}$

* $r_{y,x_i} = \sum_j r_{x_i,x_j} \cdot \beta_j$ lautet in diesem Fall: $r_{41} = r_{13}\ p_{43} + r_{12}\ p_{42} + r_{11}\ p_{41}$, wobei $r_{12} = p_{21}$ und $r_{11} = 1$. Ferner war (s.o.): $r_{13} = p_{31} + p_{21}\ p_{32}$

3.2.3.3 Unvollständiges Modell

Sobald ein rekursives System unvollständig ist, d.h. nicht alle einer abhängigen Variablen kausal vorangehenden Variablen eine kausale Wirkung auf die abhängige Variable haben, ist das System überdeterminiert und testbar. Es kann jedoch durch folgende Annahme genau determiniert werden (vgl. auch Hummell/Ziegler 1976: 72/73): Die Regression der abhängigen Variablen erstreckt sich jeweils auf genau die kausal vorangehenden Variablen, die einen kausalen Einfluss auf die abhängige Variable haben. Die Residualvariable jeder abhängigen Variablen ist also nicht zu allen der abhängigen Variablen kausal vorangehenden Variablen orthogonal, sondern nur zu solchen, die einen kausalen Einfluss auf die abhängige Variable haben.

Für jede abhängige Variable x_i (i = 2, ..., k) gibt es Variablen $x_{i_1},...,x_{i_m}$ $\left(\text{mit } i_1,...,i_{m_i} \in \{1,...,i-1\}\right)$, die einen kausalen Einfluss auf x_i haben. Die Regression erstreckt sich auf genau diese Variablen: $\hat{x}_i = f\left(x_{i_1},...,x_{i_{m_i}}\right)$

Genauer: $\hat{x}_i = p_{i,i_1} x_{i_1} + ... + p_{i,i_{m_i}} x_{i_{m_i}}$

$r_{i,i_j} = p_{i,i_1} r_{i_1,i_j} + ... + p_{i,i_{m_i}} r_{i_{m_i},i_j}$ (für j = 1, ..., m_i)

Dies ist die übliche Regressionslösung, jetzt für den Fall, dass die Regression sich nur auf *die* kausal vorangehenden Variablen erstreckt, die einen kausalen Einfluss haben.
Schreibt man den nicht erklärten Rest $x_i - \hat{x}_i$ wieder einer Restvariablen in der Form $p_{i,u} \cdot u_i$ zu, so gilt:

$$p_{i,u} = \sqrt{1 - R^2 x_i . x_{i_1},...,x_{i_{m_i}}}$$

Ferner: Die Restvariable ist nur orthogonal zu (unabhängig von) den Prädiktoren von x_i, d.h. zu den Variablen, die einen kausalen Einfluss auf x_i haben. Also kann die Restvariable mit kausal vorangehenden Variablen korrelieren und mit den Restvariablen von kausal vorangehenden Variablen.

3.2.3.4 Standardisierte oder unstandardisierte Koeffizienten?

In der Pfadanalyse gibt es Fragestellungen, für die unstandardisierte Koeffizienten besser geeignet sind: Wenn man an Kausalgesetzen oder Kausalprozessen interessiert ist (vgl. z.B. Blalock) und/oder verschiedene Populationen vergleichen will. Dies liegt daran, dass standardisierte Koeffizienten zwei Aspekte zusammenfassen: Den direkten Effekt in absoluten Zahlen und die Streuung der Variablen in der Population. Will man ausschließlich den direkten Kausaleffekt betrachten, so muss man mit den unstandardisierten Koeffizienten arbeiten, die in folgendem Zusammenhang zu den standardisierten Koeffizienten stehen:

$$b_j \cdot \frac{s_{x_j}}{s_y} = \beta_j$$

Für den Vergleich von Kausalmodellen unterschiedlicher Populationen sollte man also die unstandardisierten Koeffizienten verwenden.

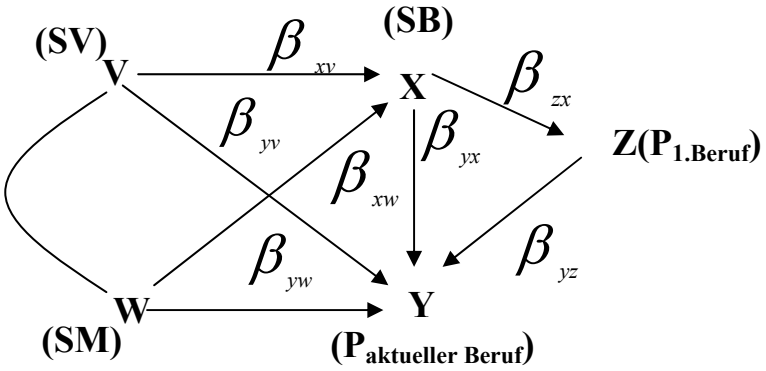
3.2.4 Die vier Mechanismen zur Erklärung einer Korrelation

Wie die Variation von y erklärt werden kann, wurde anlässlich der multiplen Regression dargestellt.

Im Folgenden wird nun gezeigt, dass sich die Korrelation zwischen x und y dadurch erklären lässt, dass sie aufgrund von vier gleichzeitig wirkenden Mechanismen zustande kommt.

Mit Hilfe der Pfadanalyse lassen sich die vier Mechanismen angeben, auf Grund derer ein beliebiger statistischer Zusammenhang zustande kommt. (Nach Finney 1972; bei Finney werden zwei Beispiele verwendet für die vier Mechanismen, es reicht aber ein Beispiel.)

Abbildung 3-14: Beispiel



Wie kommt die Korrelation r_{yx} zustande?
Mögliche Komponenten:
Direkter Kausaleffekt
Indirekte Kausaleffekte
Scheinkomponenten
Korrelierte Effekte (bzw. Assoziationseffekte)

(In dem Beispiel sind V und W exogene Variablen, d.h. sie werden im Modell nicht erklärt.)

(Endogene Variable dagegen werden im Modell zu erklären versucht.)

Pfadtheorem: $r_{yx} = \sum_j r_{x,x_j} \beta_{y,x_j}$

y ist die abhängige Variable.

Die Summation bezieht sich auf alle Variablen x_j, die im Modell eine direkte Ursache der zu erklärenden Variablen sind.

Den Zusammenhang des Prädiktors x mit der zu erklärenden Variablen y erhält man, indem man von x per Korrelation jeweils zu einem der übrigen Prädiktoren x_j geht, und von dort per direktem Effekt β_{y,x_j} zur abhängigen Variablen y.

Dies ist nichts anderes als die übliche Lösungsgleichung der multiplen Regression, die ja hier als Schätzverfahren für die Pfadkoeffizienten verwendet wird.

In kurzer Matrixschreibweise: $r = R \cdot \beta$

(Anzahl der Zeilen, Anzahl der Spalten): (1, 1) (1, k) (k, 1)

$$\begin{pmatrix} r_{y,x_1} \\ \vdots \\ r_{y,x_k} \end{pmatrix} = \begin{pmatrix} r_{x_1,x_1} & \cdots & r_{x_1,x_k} \\ \vdots & & \vdots \\ r_{x_k,x_1} & \cdots & r_{x_k,x_k} \end{pmatrix} \cdot \begin{pmatrix} \beta_{y,x_1} \\ \vdots \\ \beta_{y,x_k} \end{pmatrix}$$

Für das oben genannte Modell gilt nach dem Pfadtheorem:

$$r_{yx} = \underline{\beta_{yx}} + r_{xv}\underline{\beta_{yv}} + r_{xw}\underline{\beta_{yw}} + r_{xz}\underline{\beta_{yz}}$$

$r_{xz}(=r_{zx}) = \beta_{zx}$ (nach Pfadtheorem)

$r_{xv} = \beta_{xv} + \beta_{xw}r_{vw}$ (nach Pfadtheorem)

$r_{xw} = \beta_{xw} + \beta_{xv}r_{vw}$ (nach Pfadtheorem)

$$r_{yx} = \beta_{yx} + \beta_{xw}\beta_{yw} + \beta_{xv}\beta_{yw}r_{vw} + \beta_{xv}\beta_{yv} + \beta_{xw}\beta_{yv}r_{vw} + \beta_{yz}\beta_{zx}$$

Der Gesamtzusammenhang r_{yx} setzt sich zusammen aus folgenden 4 Komponenten:

1) β_{yx} (**direkter Kausaleffekt**)
2) $\beta_{yz}\beta_{zx}$ (**indirekter Kausaleffekt**)
3) Scheinkomponente der Korrelation auf Grund des verursachenden Faktors W: $\beta_{xw}\beta_{yw}$

   Scheinkomponente auf Grund von V: $\beta_{xv}\beta_{yv}$

   Insgesamt: Scheinkomponente: $\beta_{xw}\beta_{yw} + \beta_{xv}\beta_{yv}$

4) Komponente auf Grund der Korrelation von exogenen Variablen (**Korrelierte Effekte** bzw. **Assoziationseffekte**)

   $\beta_{xv}\beta_{yw}r_{vw} + \beta_{xw}\beta_{yv}r_{vw}$

   (Exogen: im Modell nicht erklärt; im Modell nur als unabhängige Variable)

   (Im Gegensatz dazu:
   Endogen: im Modell erklärt;
   im Modell abhängige Variable)

Abbildung 3-15: Beispiel: Mehr indirekte Effekte von x auf y

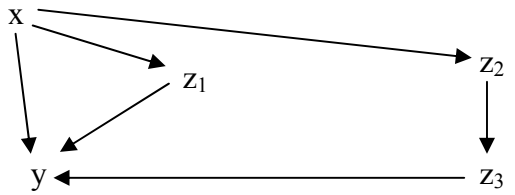
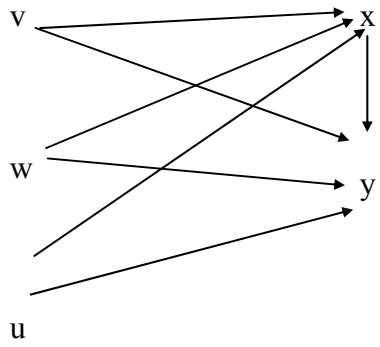
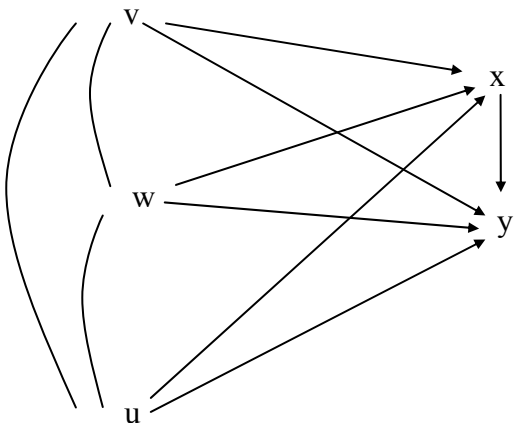


Abbildung 3-16: Beispiel: Mehr Scheinkomponenten in der Beziehung zwischen x und y



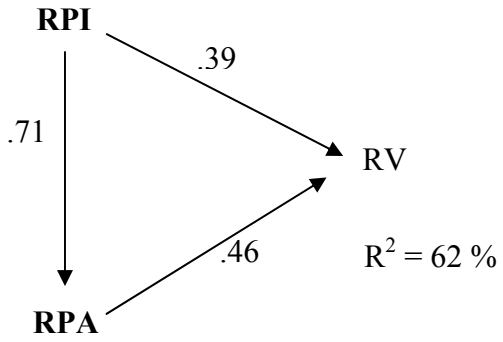
Scheinkomponenten könnten auch über Kausalketten entstehen.

Abbildung 3-17: Beispiel: Mehr „korrelierte Effekte“ in der Beziehung zwischen x und y



3.2.5 Anwendungsbeispiel: Parteienwahl in Abhängigkeit von Parteiidentifikation und Einstellungen

Abbildung 3-18: Beispiel nach Herbert B. Asher (1987)



RPI - respondent`s partisan identification

RPA - respondent`s partisan attitudes

RV - respondent`s vote

Effekte	Kausaleffekt	Kausaleffekt	
	Direkter Effekt	Indirekter Effekt	Gesamter Kausaleffekt
Party identification	.39	.33 (=.71 x .46)	.72
Partisan attitudes	.46	-	.46

Größerer direkter Effekt: Einstellungen, und nicht Parteiidentifikation.

Wegen des zusätzlichen indirekten Effekts über die Einstellungen hat die Parteiidentifikation den größeren Gesamteffekt.

Beziehung zwischen RPI und RV

direkt .39

indirekt .71 · .46 = .33

$s_{RV, RPI}$ = .72 (alles kausal)

Beziehung zwischen RPA und RV

direkt .46

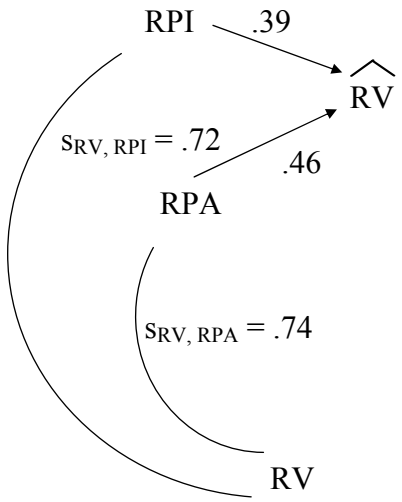
spurious .71 · .39 = .28

$s_{RV, RPA}$ = .74 (Komponente .28 nicht kausal)

Die Parteienwahl lässt sich in diesem Fall zu 62 % durch Parteiidentifikation und Einstellungen erklären.

Erklärte Varianz : Multiple $R^2 = s_{y,\hat{y}} = \sum_{i=1}^{k} r_{yx_i} \beta_{yx_i}$

Meine „pfadanalytische“ Darstellung der erklärten Varianz:



Multiple R^2 ist die Kovarianz von y und $\hat{y}$. Der Zusammenhang von Effekten und Erklärungskraft besteht darin, dass man die gesamte erklärte Varianz erhält, wenn man von der zu erklärenden Variablen per Kovarianz jeweils zu einer der „Ursachen“ geht, die jeweils die angegebenen „Kausaleffekte“ haben.

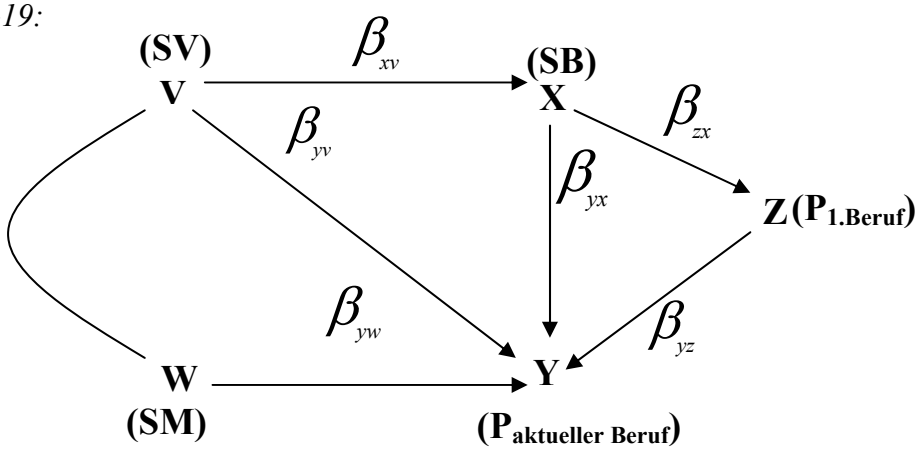
$$\underbrace{.72 \cdot .39}_{0,28} + \underbrace{.74 \cdot .46}_{0,34} = 0,62 = R^2$$

Der Bestandteil $.28 \cdot .46 = .13$ ist nicht kausal, sodass die „kausale“ Erklärung nur die Größenordnung $R^2 - 0,13 = 0.49$ hat. In diesem Sinne sind Kausaleffekte noch wichtiger als die Erklärungskraft.

3.2.6 Multiple R^2 in der Pfadanalyse

Die Grundelemente der Zerlegung sollen exemplarisch dargestellt werden, wobei das Beispiel aus Gründen der Übersichtlichkeit um den direkten Effekt von w auf x reduziert wird.

Abbildung 3-19:



Multiple $R^2 = r_{yx} \cdot \beta_{yx} + r_{yz} \cdot \beta_{yz} + r_{yv} \cdot \beta_{yv} + r_{yw}\, \beta_{yw}$

$r_{yx} \cdot \beta_{yx} = \underbrace{\beta_{yx} \cdot \beta_{yx}}_{\text{direkt von x • direkt von x}} + \underbrace{\beta_{yz}\, \beta_{zx} \cdot \beta_{yx}}_{\text{indirekt von x • direkt von x}} + \underbrace{\beta_{xv}\, \beta_{yv} \cdot \beta_{yx}}_{\text{spurious durch v • direkt von x}} + \underbrace{r_{vw}\, \beta_{yw}\, \beta_{xv} \cdot \beta_{yx}}_{\text{korreliert •direkt von v • direkt von x}}$

in (yx) in (yx) in (yx) in (yx)

$r_{yz} \cdot \beta_{yz} = \underbrace{\beta_{yz} \cdot \beta_{yz}}_{\text{direkt von z • direkt von z}} + \underbrace{\beta_{zx}\, \beta_{yx} \cdot \beta_{yz}}_{\text{spurious durch x • direkt von z}} + \underbrace{\beta_{xv}\, \beta_{zx}\, \beta_{yv} \cdot \beta_{yz}}_{\text{spurious durch v • direkt von z}} + \underbrace{r_{wv}\, \beta_{xv}\, \beta_{zx} \cdot \beta_{yw} \cdot \beta_{yz}}_{\text{korreliert •direkt von w • direkt von z}}$

in (yz) in (yz) in (yz) in (yz)

$r_{yv} \cdot \beta_{yv} = \underbrace{\beta_{yv} \cdot \beta_{yv}}_{\text{direkt von v • direkt von v}} + \underbrace{\beta_{yx}\, \beta_{xv} \cdot \beta_{yv}}_{\text{indirekt von v • direkt von v}} + \underbrace{r_{vw}\, \beta_{yw} \cdot \beta_{yv}}_{\text{korreliert von v • direkt von v}} + \underbrace{\beta_{xv}\, \beta_{zx}\, \beta_{yz} \cdot \beta_{yv}}_{\text{indirekt von v • direkt von v}}$

in (yv) in (yv) in (yv) in (yv)

$r_{yw} \cdot \beta_{yw} = \underbrace{\beta_{yw} \cdot \beta_{yw}}_{\text{direkt von w • direkt von w}} + \underbrace{r_{wv}\, \beta_{yv} \cdot \beta_{yw}}_{\text{korreliert von w • direkt von w}} + \underbrace{r_{wv}\, \beta_{xv}\, \beta_{yx} \cdot \beta_{yw}}_{\text{korreliert von w • direkt von w}} + \underbrace{r_{wv}\, \beta_{xv}\, \beta_{zx}\, \beta_{yz} \cdot \beta_{yw}}_{\text{korreliert von w • direkt von w}}$

in (yw) in (yw) in (yw) in (yw)

Insgesamt setzt sich Multiple R^2 also zusammen aus:

4 direkte Beiträge von x, z, v, w
3 Kombinationen von direktem und indirektem Effekt (von x und v)
3 Kombinationen von direktem und spurious (durch v und x) Effekt
4 Kombinationen von direktem und korreliertem (von v und w) Effekt
2 Kombinationen von direktem und korreliert • direktem (von v und w) Effekt

3.2.6.1 Zentrale Konzepte der statistischen Analyse gemäß der Pfadanalyse

Direkte Effekte
Indirekte Effekte
Korrelierte Effekte

Ferner: Die „spurious"-Effekte müssen bei der erklärten Varianz relativierend berücksichtigt werden.

Die relative Wichtigkeit von Erklärungsfaktoren für ein zu erklärendes Phänomen (y) kann einerseits anhand der direkten Beiträge diskutiert werden, andererseits anhand der Gesamtbeiträge.

Durch eine Korrelation r_{yx} allein kann man insofern getäuscht werden, als es „scheinkausale Korrelationen", Suppressor- und Distorter-Phänomene etc. gibt, wie sich im Vergleich zum eigentlichen direkten Kausalmechanismus (direkter Effekt) ergibt. D.h. die Bearbeitung des Problems von Korrelation und Kausalität sollte in der schrittweisen Elaborierung „pfadanalytischer" Modellierungen geschehen.

3.2.7 Effekte und Erklärungskraft in der Pfadanalyse

Das Argumentieren mit direkten und indirekten Effekten soll nun im Vergleich von multipler Regression und Pfadanalyse diskutiert werden.

3.2.7.1 Korrelierte Effekte (Multiple Regression)

Wegen der größeren Übersichtlichkeit der Formeln werden standardisierte Variablen y, x_1, ..., x_k vorausgesetzt, zumal die Zurückrechnung auf nicht standardisierte Variablen ja immer möglich ist.

Das Problem der multiplen Regression besteht bekanntlich darin, eine abhängige Variable y aufgrund einer Linearkombination $\hat{y} = \sum_{i-1}^{k} \beta_i x_i$ der Prädiktoren möglichst gut zu approximieren, indem der Fehler $\sum_{j-1}^{n} \left(y_j - \hat{y}_j\right)^2$ (über die Einheiten j) minimiert wird.

Die Lösung der multiplen Regression lautet bekanntlich: $\beta = R^{-1}$ r. Hierbei ist $\beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}$ der Koeffizientenvektor, $r = \begin{pmatrix} r_{y,x_1} \\ \vdots \\ r_{y,x_k} \end{pmatrix}$ der Vektor der Korrelationen der abhängigen Variablen mit den Prädiktoren und R die Korrelationsmatrix der Prädiktoren.

Also: $R\beta = r$, ausführlicher: $r_{y,x_i} = \sum_{j=1}^{k} r_{x_i,x_j} \beta_j$

Mit Hilfe des Skalarprodukts <a,b> : = $\sum_{i01}^{n} a_i b_i$ für 2 beliebige Vektoren a = (a_1, ..., a_n), b = (b_1, ..., b_n) (also Kovarianz $s_{a,b} = \frac{< a,b >}{n}$) erhält man: $< y, x_i >= \sum_{j=1}^{k} < \beta_j x_j, x_i >=< \hat{y}, x_i >$ oder: $< y - \hat{y}, x_i >= 0$, d.h. das Residuum $y - \hat{y}$ ist orthogonal zu allen Prädiktoren x_i, korreliert also nicht mit ihnen.

Es folgt: $r_{y,x_i} = \beta_i + \sum_{\substack{j \\ j \neq i}} r_{x_i,x_j} \beta_j$

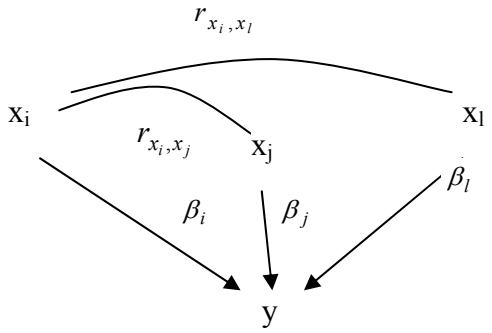
Rein algebraisch ist also die Korrelation zerlegt in verschiedene Bestandteile. Dies belegt noch nicht, dass sich dies besonders gut interpretieren lässt, dazu kann man aber folgendes berücksichtigen: r_{y,x_i} lässt sich als Gesamteffekt eines Prädiktors x_i auf die zu erklärende Variable y betrachten, denn eine Änderung von x_i um eine Einheit ergibt eine Änderung in y um r_{y,x_i}, weil die einfache Regression lautet (standardisierte Variablen unterstellt):

$y = r_{y,x_i} x_i + \varepsilon_i$

Also: $\Delta y = \underbrace{r_{y,x_i}}_{1} \Delta x_i$ (x_i korreliert nicht mit dem Fehler ε_i)

Die folgende graphische Darstellung illustriert die Unterscheidung in direkte und korrelierte Effekte:

Abbildung 3-20:



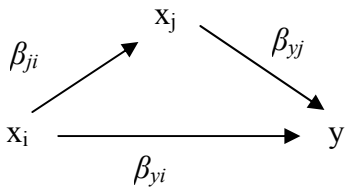
$y = \beta_i x_i + \sum_{\substack{j \\ j \neq i}} \beta_j x_j + \varepsilon$

Eine Änderung von x_1 um eine Einheit bewirkt direkt eine Änderung in y um β_i, in diesem Sinne ist β_i der direkte Effekt. Da x_i mit den x_j $(j \neq i)$ korreliert, ergibt die Änderung in x_i eine Änderung in x_j $(j \neq i)$ und dadurch schließlich eine weitere Änderung in y. (x_i korreliert nicht mit dem Fehler ε.) Alle Änderungen letzterer Art zusammen sollen als korrelierte Effekte verstanden werden.

$r_{y,x_i} = \beta_i + \sum_{\substack{j \\ j \neq i}} r_{x_i,x_j} \beta_j$

3.2.7.2 Indirekte Effekte (Pfadanalyse)

Abbildung 3-21:



Die Regressionsgleichungen eines solchen Pfadmodells lauten:

$$x_j = \beta_{ji} x_i + \varepsilon_j$$

$$y = \beta_{yi} x_i + \beta_{yj} x_j + \varepsilon$$

Eine Änderung in x_i um eine Einheit ergibt unmittelbar eine Änderung β_{yi} in y, dies soll wieder als direkter Effekt verstanden werden. Die Änderung in x_i ergibt außerdem eine Änderung in x_j um β_{ji} und deshalb eine Änderung $\beta_{yj}\ \beta_{ji}$ in y.

$$y = \beta_{yi} x_i + \beta_{yj} \beta_{ji} x_i + \beta_{yj} \varepsilon_j + \varepsilon$$

Die Residuen ε_j und ε korrelieren nicht mit x_i, also impliziert die Änderung in x_i keine weitere Änderung in y.

Die algebraische Zerlegung des Gesamteffektes in den direkten und indirekten Effekt ergibt sich aus: $r_{y,x_i} = s_{y,x_i} = s_{\hat{y},x_i} = \beta_{yi} + \beta_{yi}\ \beta_{ji}$

Da die Prädiktoren korrelieren dürfen, lässt sich der Effekt von x_i nicht von dem Einfluss der übrigen Prädiktoren isolieren. Gerade als Reaktion darauf ließe sich die Unterscheidung in einen direkten und einen indirekten Effekt interpretieren: Eine Änderung von x_i um eine Einheit induziert nicht nur die Änderung $\beta_i = \beta_i \Delta x_i$ in x, sondern – entsprechend der Grundgleichung – außerdem noch: $\sum_{\substack{j \\ j \neq i}} r_{x_i,x_j} \beta_j$

3.2.7.3 Problematisierung der Pfadanalyse

In der Pfadanalyse werden in einem Diagramm die direkten Effekte β_i eingetragen, um die relative Einflussstärke der Variablen auszudrücken. Genau so wichtig aber ist der Gesamteffekt r_{y,x_i} einer Variablen. Es soll nun im folgenden gezeigt werden, dass die Beschränkung der Interpretation auf die β_i völlig unangemessen ist, da die r_{y,x_i} in jeder Richtung von den β_i abweichen können. Eine Berücksichtigung beider Koeffizienten dagegen verhilft zur Klärung der Situation.

Im folgenden soll mit dem Konzept des Gesamteffekts $\left(r_{y,x_i}\right)$ des Prädiktors x_i und des bereinigten Effekts $\left(\beta_i\right)$, d.h. des Effekts des bereinigten Prädiktors $x_i - \hat{x}_i$, gearbeitet werden. Der Effekt ist die induzierte Änderung in der abhängigen Variablen, wenn der Prädiktor um eine Einheit geändert wird.

Also:
Gesamt-Effekt von x_i: r_{y,x_i}

Bereinigter Effekt von x_i, d.h. Effekt des bereinigten Prädiktors $x_i - \hat{x}_i$: β_i

Gesamt-Erklärungsbeitrag von x_i: r^2_{y,x_1}

Bereinigter Erklärungsbeitrag von x_i, d.h. Beitrag des bereinigten Prädiktors $x_i - \hat{x}_i$: $r^2_{y,x_1-\hat{x}_i} = R^2 - R^2_{(i)}$

Betrachtet man Zerlegung von Multiple R^2 von der Form $R^2 = \sum_{i=1}^{k} r^2_{y,x_i-\hat{x}_i} + \Delta_1$ und $R^2 = \sum_{i=1}^{k} r^2_{y,x_i} - \Delta_2$, so muss man berücksichtigen, dass Δ_1 und Δ_2 nicht positiv sein müssen, sie können auch Null oder kleiner Null sein. Solche Phänomene sollen nun genauer untersucht werden.

(Die Ursache des Problems besteht darin, dass $x_i - \hat{x}_i$ orthogonal zu allen x_j $(j \neq i)$ ist, nicht aber zu $\hat{x}_j$, da in $\hat{x}_j$ wieder der Prädiktor x_i auftritt. Rechnet man aus x_i die übrigen Prädiktoren heraus, so korreliert $x_i - \hat{x}_i$ nicht mehr mit x_j. Bereinigt man auch x_j um die übrigen Prädiktoren, so kommt dadurch wieder der Einfluss von x_i herein: $x_i - \hat{x}_i$ und $x_j - \hat{x}_j$ $(j \neq i)$ müssen nicht orthogonal sein. Die einzige Möglichkeit, dieses Problem zu vermeiden, ist die Vorgabe einer Hierarchie der Prädiktoren: Rechnet man aus x_i alle Prädiktoren $x_1 (1 < i)$ und aus x_j $(j > i)$ alle Prädiktoren $x_1 (1 < j)$ heraus, so tritt das Problem nicht mehr auf: $x_j - \hat{x}_j$ ist orthogonal zu allen $x_1 (1 < j)$, also auch zu $x_i - \hat{x}_i$. In $x_i - \hat{x}_i$ tritt x_j nicht auf, weil $\hat{x}_i$ die Regression von x_i auf x_1, ..., x_{i-1} ist und $j > i$.)

3.2.7.4 Vergleich von Gesamtzusammenhang und bereinigtem Zusammenhang: Typologie und Zerlegung von R^2

Der Gesamtzusammenhang (ges) lässt sich zerlegen in den bereinigten Zusammenhang (ber) und einen Rest (res = ges – ber).

(Die Typologie ist so allgemein formuliert, dass sie sich auch z.B. für ber = Part Correlation verwenden ließe. Die Verwendung von ber = β_i scheint mir aber fruchtbarer, da dann res (= korrelierte/ indirekte Effekte) besser zu analysieren ist, wie unten auch an einem Beispiel gezeigt wird.)

a) Ein **Suppressor-Phänomen** liegt vor, wenn der bereinigte Zusammenhang von x_i (dies sei β_i) größer ist als der Gesamtzusammenhang (r_{y,x_i}). D.h. die Störfaktoren supprimieren den „wahren" Einfluss von x_i.

Formal:

ges > 0, S0: ber > ges > 0 (Es folgt: res < 0)

ges < 0, S0: ber < ges < 0 (Es folgt: res > 0)

Daraus folgt jeweils: |ber| > |res|

sign (ber) ≠ sign (res), sign (ges) ≠ sign (res), sign (ges) = sign (ber)

1) Sei ges > 0.
   ber > ges = ber + res ⇨ res < 0, - res > 0
   0 < ges = ber + res ⇨ - res < - ber
   Also: |res| < |ber|

2) Sei ges < 0.
   ber < ges = ber + res ⇨ res > 0
   0 > ges = ber + res ⇨ res < - ber
   Also: |res| < |ber|

b) Ein **Distorter-Phänomen** (Verzerrung) liegt vor, wenn der bereinigte Zusammenhang von x_i (dies sei β_i) ein entgegengesetztes Vorzeichen zu dem Gesamtzusammenhang (r_{y,x_i}) hat. D.h. die Störfaktoren verzerren den „wahren" Einfluss von x_i.

Formal:
ges > 0, so: ges > 0 > ber (Es folgt: res > 0)
ges < 0, so: ges < 0 < ber (Es folgt: res < 0)

Daraus folgt jeweils: |ber| < |res|
sign (ber) ≠ sign (res), sign (ges) = sign (res), sign (ges) ≠ sign (ber)

1) Sei ges > 0.
   ber + ges = ges > ber ⇨ res < 0
   0 < ges = ber + res ⇨ res > - ber
   Also: |res| > |ber|

2) Sei ges < 0.
   ber + res = ges < ber ⇨ res < 0
   0 > ges = ber + res ⇨ ber < - res
   Also: |ber| < |res|

Dies kann man, wie in der Tabellenanalyse bekannt, in anschaulicher Form durch folgendes Schema darstellen:

	Gesamtzusammenhang	Bereinigter Zusammenhang	Rest
Suppressor Phänomen	+	++	-
	-	--	+
Distorter Phänomen	+	-	++
	-	+	--

c) Ein **„Schein-Effekt"** lässt sich als benachbart zum Distorter-Phänomen ansehen: Rechnet man die Störfaktoren heraus, so wechselt zwar nicht das Vorzeichen, aber der „wahre" Effekt (: = ber) erweist sich als verschwindend.

Formal:
ges > 0, so: ber = 0 (Es folgt: res > 0)
ges < 0, so: ber = 0 (Es folgt: res < 0)

Die Subsumtion unter das Distorter-Phänomen zeigt sich daran, dass:
|ber| = 0 < |res|, sign (ges) = sign (res)

Ferner gilt: ber = 0, ges = res

1) 0 < ges = ber + res = res ⇨ res = ges > 0.

   Also: |res| > |ber|
2) 0 > ges = ber + res = res ⇨ res = ges < 0 – res > 0.

   Also: |res| > |ber|

	Gesamtzu-sammenhang	Bereinigter Zusammenhang	Rest
Schein-Effekt	+	0	+
	-	0	-

d) Eine **„Überschneidung“** liegt vor, wenn der „wahre“ Einfluss von x_i (: = ber) als geringer sichtbar wird (aber nicht das Vorzeichen wechselt oder Null wird), wenn die anderen Prädiktoren als Störfaktoren herausgerechnet werden. D.h. es handelt sich um eine Abschwächung des „Schein-Effekts“.

Formal:

ges > 0, so: ges > ber > 0 (Es folgt: res > 0)
ges < 0, so: ges < ber < 0 (Es folgt: res < 0)

Die Subsumtion unter der Distorter Richtung zeigt sich daran, dass:

sign (ges) = sign (res)

Insgesamt: sign (ber) = sign (res)= sign (ges)

1) Sei ges > 0.
   ber <ges = ber + res ⇨ res > 0
2) Sei ges < 0.
   ber > ges = ber + res ⇨ res < 0

Schema:

Überschneidung	Gesamtzu-sammenhang	Bereinigter Zusammenhang	Rest
Überschneidung	++	+	+
	--	-	-

e) Eine **„scheinbare Nicht-Kausalität“** liegt vor, wenn die Existenz eines Einflusses erst sichtbar wird, wenn die übrigen Prädiktoren herausgerechnet werden. Dies ist ein Suppressor-Phänomen mit Gesamtzusammenhang Null.

Formal:

ges = 0, ber > ges (Es folgt: res < 0)
oder:
ges = 0, ber = ges = 0 (Es folgt: res > 0)

Insgesamt: ber = - res

0 = ges = ber + res ⇨ ber = - res

4) ber > 0 ⇨ res < 0
4) ber < 0 ⇨ res > 0

Schema:

	Gesamtzusammenhang	Bereinigter Zusammenhang	Rest
Scheinbare Nicht-Kausalität	0	+	-
	0	-	+

f) Schließlich können Gesamtzusammenhang und bereinigter Zusammenhang übereinstimmen, d.h. die anderen Prädiktoren wirken nicht als Störfaktoren (**Übereinstimmung**):

Formal:

ges = ber (Es folgt: res = 0)

Schema:

	Gesamtzusammenhang	Bereinigter Zusammenhang	Rest
Übereinstimmung	+	+	0
	-	-	0
	0	0	0

g) **Identifikation der Typen**

A) Falls ges = ber: Übereinstimmung

B) Falls ges ≠ ber

B1) Falls ges · ber < 0, so Distorter Phänomen

B2) Falls ges · ber = 0

a) Falls ges = 0 (wegen ber ≠ ges gilt ber ≠ 0), so: Scheinbare Nicht-Kausalität

b) Falls ber = 0 (wegen ges ≠ ber gilt ges ≠ 0), so: Schein-Kausalität

B3) Falls ges · ber > 0 (wegen ges ≠ ber nur |ges| > |ber| oder |ges| < |ber| möglich)

a) Falls |ges| > |ber|, so: Überschneidung

b) Falls |ber| > |ges|, so: Suppressor-Phänomen

h) **Zerlegung von R^2 gemäß der Typologie**

Wegen der Gleichung $R^2 = \sum_{i=1}^{k} \beta_i r_{y,x_i}$ lässt sich R^2 leicht zerlegen in:

$\sum_{r_{y,x_i} = \beta_i \neq 0} r^2_{y,x_i}$: Übereinstimmung

$\sum_{\beta_i r_{y,x_i} < 0} \beta_i r_{y,x_i}$: Distorter (negativer Summand)

$\sum_{\substack{\beta_i r_{y,x_i} > 0 \\ |\beta_i| < |r_{y,x_i}|}} \beta_i r_{y,x_i}$: Überschneidung

$\sum_{\substack{\beta_i r_{y,x_i} > 0 \\ |\beta_i| > |r_{y,x_i}|}} \beta_i r_{y,x_i}$: Suppressor

Diese Zerlegung zeigt rechnerisch, wie Multiple R^2 zustande kommt. Die korrelierten Effekte lassen sich also gemäß der Typologie gruppieren.

Beispiel 1: Juso-Präferenz versus Basisgruppen-Präferenz

Die dargestellten Argumente sollen nun auf ein Beispiel angewendet werden: Die Konzepte der „gewerkschaftlichen Orientierung“ (repräsentiert durch die Jusos) und der „Alternativ-Orientierung“ (repräsentiert durch die Basisgruppen (BG)) sollten durch verschiedene Indikatoren charakterisiert werden. Mit den Indikatoren, die sich im t-Test als diskriminierend erwiesen, als Prädiktoren wurde eine multiple Regression durchgeführt, wobei die Juso-BG-Wahldichotomie die abhängige Variable ist. (Die Problematik einer abhängigen Dichotomie soll hier ausgeklammert bleiben.)

Bei der Fragestellung, die (nominale) Wahl auf (metrische) Indikatoren zurückzuführen, wird man gleich an die *Diskriminanzanalyse* denken. Im Fall von zwei Gruppen sind die Koeffizienten a_i der (einzigen, weil min {k – 1, p} = 1 für k = 2) Diskriminanzfunktion $d = a_1 x_1 + ... + a_p x_p$ aber proportional zu den β_i der Regression der Dummy-Variablen der Wahlentscheidung auf die Prädiktoren:

$d = \frac{\hat{y}}{s_{\hat{y}}}, s_{\hat{y}}$ = Multiple R, also: $a_i = \frac{\beta_i}{R}$

Deshalb reicht es, eine multiple Regression durchzuführen. Die Anzahl der Prädiktoren einer schrittweisen Regression lässt sich mit Hilfe einer graphischen Darstellung des Erklärungszuwachses (wie in der Faktorenanalyse gebräuchlich) bestimmen. (Eine Alternative wäre, so viele Prädiktoren zu berücksichtigen, bis im nächsten Schritt ein Effekt nicht mehr signifikant ist.) Nach 8 Schritten ließ sich $R^2 = 85{,}4$ % der Varianz erklären. Der Gesamt-F-Wert lag mit 44,5 weit über dem kritischen Wert $F_{8,61} \approx 2{,}1$. (Die Stichprobe von 324 Studenten wird dadurch drastisch reduziert, dass nur die Teilgruppen BG- und Juso-Wähler berücksichtigt werden.) Die einzelnen Variablen liefern alle einen nach dem F-Test (kritischer Wert $F_{1,61} \approx 4$) signifikanten Erklärungszuwachs.

Tabelle 3-7: Juso-Präferenz (versus Basisgruppen-Präferenz)

(= Abhängige Variable)

Prädiktoren	Gesamtzu-sammen-hang r	Gesamter-klärungskraft r^2	Direkter Effekt bzw. Effekt von $x_i - \hat{x}_i$, d.h. Beta	F	Erklärungsbeitrag von $x_i - \hat{x}_i$, d.h. $r^2_{y,x_i-\hat{x}_i} = R^2 - R^2_{(i)}$	R^2 bei schrittweisem Vorgehen	ΔR^2
Bunte Liste positiv	- 0,56	31 %	- 0,38	49,3	12 %	31 %	31 %
SPD positiv	0,48	23 %	0,31	37,4	9 %	47 %	16 %
Für Frauenberufstätigkeit	0,24	6 %	0,37	53,3	13 %	57 %	10 %
Für Wohngemeinschaft	0,39	15 %	- 0,21	15,1	4 %	65 %	8 %
Verdienst wichtig bei Berufsentscheidung	0,29	8 %	0,27	28,6	7 %	71 %	6 %
Vater gewerkschaftsnah	0,28	8 %	0,23	22,3	5 %	77 %	6 %
Kommunistischer Bund positiv	- 0,46	21 %	- 0,22	17,2	4 %	82 %	5 %
Berufsaussichten positiv	0,40	16 %	0,20	14,7	4 %	85 %	3 %

Da $F = \dfrac{r^2_{y,x_i\hat{x}_i}}{\dfrac{(1-R^2)}{(n-k-1)}}$, berechnet man: $r^2_{y,x_i\hat{x}_i} = F \cdot \dfrac{1-0{,}85362}{61}$

1) Bunte Liste positiv

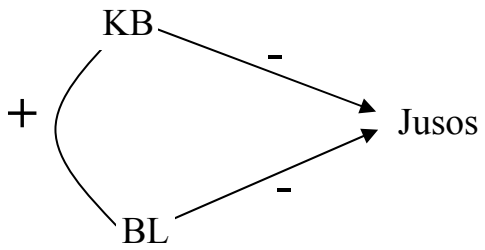
Typ B3a (*Überschneidung*), nämlich: ges ≠ ber, ges · ber > 0, |ges| > |ber|
Es soll nun genauer dargestellt werden, warum der bereinigte Zusammenhang kleiner ist.

r_{y,x_1} - 0,56

$\beta_1 =$ - 0,38	$r_{x_1x_2\beta_2}$ = -.15.31 = -.05	$r_{x_1x_3\beta_3}$ = .15.37 = .06	$r_{x_1x_4\beta_4}$ = .23 (-.21) = -.02	$r_{x_1x_5\beta_5}$ = .08.27 = -.02	$r_{x_1x_6\beta_6}$ = -.01.23 = -.002	$r_{x_1x_7\beta_7}$ = .36 (-.22) = -.08	$r_{x_1x_8\beta_8}$ = -.17.20 = -.03	$\sum =$ - 0,56
BL	SPD	FR.	WG	VE.	VG.	KB	BE.	

Die positiven „Pfade“ ergeben .06, die negativen -.25, weshalb aus dem bereinigten Zusammenhang – 0,38 ein Gesamtzusammenhang von – 0,56 (= - 0,38 + 0,056 – 0,232) wird. Am stärksten ins Gewicht fallen die negativen Pfade über KB, SPD, WG.

Abbildung 3-22:



Die KB-Variable ist ähnlich zu der BL-Variablen, so dass bei Kontrolle der KB-Variablen der Effekt der BL-Variablen geringer wird.

2) SPD positiv

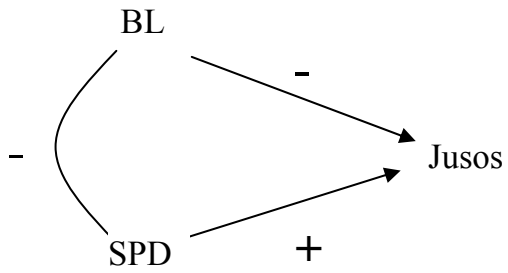
Typ B3a (*Überschneidung*)

r_{y,x_2} 0,48

$\beta_2 =$ 0,31	$r_{x_2x_1\beta_1}$ = -.15(-.38) = .06	$r_{x_2x_3\beta_3}$ = .03.37 = .01	$r_{x_2x_4\beta_4}$ = -.11 (-.21) = .02	$r_{x_2x_5\beta_5}$ = .02.27 = .005	$r_{x_2x_6\beta_6}$ = .06.23 = .01	$r_{x_2x_7\beta_7}$ = -.14 (-.22) = .03	$r_{x_2x_8\beta_8}$ = .17.20 = .03	$\sum =$ 0,48
SPD	BL	FR.	WG	VE.	VG.	KB	BE.	

Alle korrelierten Pfade sind positiv, so dass aus dem bereinigten Zusammenhang 0,31 ein Gesamtzusammenhang von 0,48 wird (= 0,31 + 0,17). Am stärksten ins Gewicht fällt der Pfad über die BL.

Abbildung 3-23:



Die Bunte Liste hängt negativ mit SPD und Jusos zusammen, so dass ein Herausrechnen der BL den Effekt der SPD auf die Jusos reduziert.

3) Für Frauenberufstätigkeit

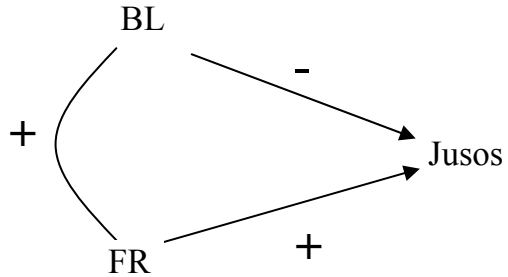
Typ B3b, nämlich: ges ≠ ber, ges · ber > 0, |ber| > |ges|

Also: *Suppressor-Phänomen*

r_{y,x_3}								0,24
$\beta_3 =$ 0,37	$r_{x_3x_1\beta_1}$ = -.15·(-.38) = -.06	$r_{x_3x_2\beta_2}$ = .03·.31 = .009	$r_{x_3x_4\beta_4}$ = .18·(-.21) = -.04	$r_{x_3x_5\beta_5}$ = -.17·.27 = -.05	$r_{x_3x_6\beta_6}$ = .09·.23 = .02	$r_{x_3x_7\beta_7}$ = .11·(-.22) = -.02	$r_{x_3x_8\beta_8}$ = .0·.20 = 0	$\sum =$ 0,24
FR.	BL	SPD	WG	VE.	VG.	KB	BE.	

Die positiven Pfade ergeben nur 0,0300, die negativen dagegen -0,1649, so dass sich aus dem bereinigten Effekt 0,37 ein geringerer Gesamteffekt 0,24 (= 0,37 + 0,0300 – 0,1649) ergibt. Am stärksten ins Gewicht fallen die negativen Pfade über BL, VE, WG.

Abbildung 3-24:



Anhänger der Bunten Liste befürworten die Frauenberufstätigkeit, lehnen die Jusos aber eher ab. Kontrolliert man die Einschätzung der BL, so wird ein stärkerer Zusammenhang zwischen der Befürwortung der Jusos und der Frauenberufstätigkeit sichtbar.

4) Alle anderen Effekte sind vom Typ der Überschneidung, so dass dies hier nicht genauer dargestellt werden soll.

Wie das Beispiel zeigt, dürften Überschneidung am häufigsten sein, ferner Suppressor-Phänomene, schließlich Distorter-Phänomene, wenn sie auch in diesem Beispiel nicht auftraten. Perfekte Übereinstimmung oder perfektes Verschwinden von ges oder ber dürfte selten sein. (Die Abweichung von den perfekten Typen wäre mit einem Signifikanztest zu behandeln.)

Da in dem Beispiel *Überschneidungen* überwiegen, muss gelten:

$$R^2 < \sum r^2_{y,x_i}, \; R^2 > \sum r^2_{y,x_i-\hat{x}_i}$$

	(a)		(b)		(c)
		a < > b		a < > c	
	$r_{y,x_i\beta_i}$		r^2_{y,x_i}		$r^2_{y,x_i-\hat{x}_i}$
Überschneidung	.21	<	.31	>	.12
Überschneidung	.15	<	.23	>	.09
Suppressor	.09	>	.06	<	.13
Überschneidung	.08	<	.15	>	.04
Überschneidung	.078	<	.084	>	.07
Überschneidung	.06	<	.08	>	.05
Überschneidung	.10	<	.21	>	.04
Überschneidung	.08	<	.16	>	.04
	$\sum = R^2 = 0{,}85$		$\sum = 1{,}28$		$\sum = 0{,}58$

$$\sum \left(r^2_{y,x_i} r_{y,x_i\beta_i}\right) = 0{,}46$$

Überschneidung

$$\sum \left(r_{y,x_i\beta_i} - r^2_{y,x_i}\right) = 0{,}03$$

Suppressor

Also: Multiple $R^2 = 0{,}85 < \sum r^2_{y,x_i} = 0{,}85 + 0{,}46 - 0{,}03 = 1{,}28$

$$\sum \left(r_{y,x_i\beta_i} - r^2_{y,x_i-\hat{x}_i}\right) = 0{,}31$$

Überschneidung

$$\sum \left(r^2_{y,x_i-\hat{x}_i} - r_{y,x_i\beta_i}\right) = 0{,}04$$

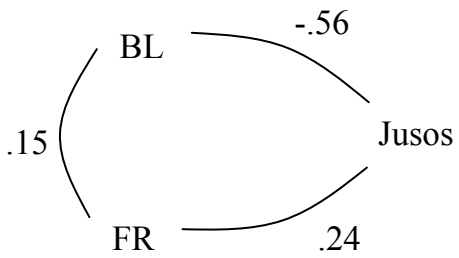
Suppressor

Also: Multiple $R^2 = 0{,}85 > \sum r^2_{y,x_i-\hat{x}_i} = 0{,}85 - 0{,}31 + 0{,}04 = 0{,}58$

Beispiel 2: **Supressor-Phänomen**

Es soll nun gezeigt werden, dass: $\sum r^2_{y,x_i} < R^2 < \sum r^2_{y,x_i-\hat{x}_i}$ möglich ist. Dazu beschränke ich mich einfach auf das Suppressor-Phänomen im vorigen Beispiel.

Abbildung 3-25:



Die Beta-Koeffizienten ergeben sich dann wegen $\beta_{yx} = \frac{r_{yx} - r_{xz}r_{yz}}{1 - r^2_{xz}}$ als:

$\beta_1 = -0{,}61 \qquad \beta_2 = 0{,}33$

Multiple $R^2 = \sum \beta_i r_{yx_i} = 0{,}42$

$\sum r^2_{y,x_i} = 0{,}37$

$r^2_{y,x_1-\hat{x}_1} = R^2 - r^2_{y,x_2} = 0{,}36$

$r^2_{y,x_2-\hat{x}_2} = R^2 - r^2_{y,x_1} = 0{,}11$

Also: $\sum r^2_{y,x_i-\hat{x}_i} = 0{,}47$

Insgesamt: $\sum r^2_{y,x_i} = 0{,}37 < R^2 = 0{,}42 < \sum r^2_{y,x_i-\hat{x}_i} = 0{,}47$

Ferner gilt im Fall von 3 Variablen (nicht aber allgemein, wie etwa Beispiel 1 zeigt): $\Delta_1 = \Delta_2$

Denn: $\sum \beta_i s_{\hat{y}(i),x_i} \left(= \sum \beta_i s_{y,\hat{x}_i}\right) = \beta_1 r_{yz} r_{xz} + \beta_2 r_{yx} r_{xz}$

(Zur 1. Gleichung: $\beta_i r_{y,x_i} = \beta_i s_{y,x_i-\hat{x}_i} + \beta_i s_{y,\hat{x}_i} = r^2_{y,x_i-\hat{x}_i} + \beta_i s_{y,\hat{x}_i}$)

Und: $\Delta_2 = r_{yz} r_{xz} \beta_2 + r_{yx} r_{xz} \beta_1$

Hier: $\Delta_1 = \Delta_2 = 0{,}05$

Beispiel 3: **Disorter-Phänomen**

Es soll nun ein Beispiel dargestellt werden, bei dem Distorter-Phänomene auftreten.

Die Einschätzung der Bunten Liste soll statistisch auf die Einschätzungen von Hochschulgruppen zurückgeführt werden.

Abhängige Variable Bunte Liste (Multiple $R^2 = 0{,}22807$)

Prädiktoren	Gesamteffekt r	Effekt von $x_i - \hat{x}_i$, d.h. Beta	F
SB	.37	.26	8.52
KB	.36	.23	7.95
BG	.32	.13	2.74
MSB	.08	-.15	2.96
RCDS	-.21	-.09	1.15
Jusos	.08	-.07	.67
SHB	.15	.05	.24
LHV	-.01	.03	.17
SLH	-.05	.01	.03

Daraus lassen sich (mit n – k – 1 =171) berechnen:

$$r^2_{y,x_i-\hat{x}_i} = F \cdot \frac{(1-R^2)}{n-k-1} = F \cdot \frac{1-0{,}22807}{171}$$

	(a)		(b)		(c)
	$r_{y,x_i\beta_i}$	$\gtrless$ b	r^2_{y,x_i}	$\gtrless$ c	$r^2_{y,x_i-\hat{x}_i}$
Überschneidung	.10	<	.14	>	.04
Überschneidung	.08	<	.13	>	.04
Überschneidung	.04	<	.10	>	.01
Distorter	-.01	$\not<$	.01	$\not<$	.01
Überschneidung	.02	<	.04	>	.01
Distorter	-.01	$\not<$	.01	$\not<$	.00
Überschneidung	.01	<	.02	>	.00
Distorter	-.00	$\not<$	.00	$\not<$	.00
Distorter	-.00	$\not<$	.00	$\not<$	.00
	$\sum = R^2 = .23$		$\sum = .45$		$\sum = .11$

$\sum$ $\left(r^2_{y,x_i} - r_{y,x_i\beta_i}\right) = 0{,}18$
Überschneidung

$\sum$ $\left(r^2_{y,x_i} - r_{y,x_i\beta_i}\right) = 0{,}04$
Distorter

Also: Multiple $R^2 = 0{,}23 < \sum r^2_{y,x_i} = 0{,}23 + 0{,}18 - 0{,}04 = .45$

$\sum$ $\left(r_{y,x_i}\beta_i - r^2_{y,x_i-\hat{x}_i}\right) = 0{,}15$
Überschneidung

$\sum$ $\left(r^2_{y,x_i-\hat{x}_i} - r_{y,x_i}\beta_i\right) = 0{,}03$
Distorter

Also: Multiple $R^2 = 0{,}23 > \sum r^2_{y,x_i-\hat{x}_i} = 0{,}11 \quad (= 0{,}23 - 0{,}15 + 0{,}03)$

Distorter-Phänomene reduzieren Multiple R^2 sowohl verglichen mit $\sum r^2_{y,x_i}$ als auch verglichen mit $\sum r^2_{y,x_i-\hat{x}_i}$, weil sie zu negativen Termen $r_{y,x_i\beta_i}$ führen. In dem Beispiel dominieren aber die Überschneidungen.

i) **<u>Effekte</u>** versus **<u>bereinigte Erklärungskraft</u>**

Für Überschneidungen (d.h. o.B.d.A. $r_{y,x_i>\beta_i} > \beta_i > 0$) gilt: $r_{y,x_i} > \beta_i > r^2_{y,x_i-\hat{x}_i}$

(Selbstverständlich gilt: $r^2_{y,x_i} > r_{y,x_i\beta_i}$)

Denn: $s_{y,x_i} > \frac{s_{y,x_i-\hat{x}_i}}{s^2_{x_i-\hat{x}_i}} \Rightarrow$

(*) $s_{y,x_i\hat{x}_i} < s_{y,x_i} \quad s^2_{x_i-\hat{x}_i} \le s_{y,x_i}$,

weil: $s^2_{x_i-\hat{x}_i} = \frac{1}{n} < x_i - \hat{x}_i, x_i - \hat{x}_i >= \frac{1}{n} < x_i - \hat{x}_i, x_i >= 1 - R^2_{x_i \cdot x_j(j\neq i)} \le 1$

Also: $r^2_{y,x_i-\hat{x}} = \frac{s^2_{y,x_i-\hat{x}_i}}{s^2_{x_i-\hat{x}_i}} = \beta_i \; s_{y,x_i-\hat{x}_i} < \beta_i s_{y,x_i}$ (nach (*) und weil $\beta_i > 0$)

Für Suppressor-Phänomene (o.B.d.A. $\beta_i > r_{y,x_i} > 0$) gilt das Entsprechende nicht:

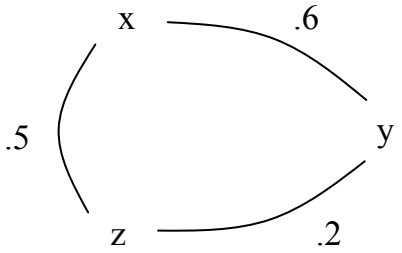
I.a. nicht: $r_{y,x_i\beta_i} < r^2_{y,x_i-\hat{x}_i}$

Noch nicht mal: $r^2_{y,x_i} < r^2_{y,x_i-\hat{x}_i}$

(Selbstverständlich gilt: $r^2_{y,x_i} < r_{y,x_i\beta_i}$)

Beispiel:

Abbildung 3-26:



$$\beta_1 = \frac{.6 - .5 \cdot .2}{1 - .5^2} = 0{,}67$$

$$\beta_2 = \frac{.2 - .5 \cdot .6}{1 - .5^2} = -0{,}13$$

(a) $\gtrless$ (b) $\gtrless$ (c)

a $\lessgtr$ b a $\lessgtr$ c

	r	β	r β		r^2		$r^2_{y,x_i-\hat{x}_i}$
(Suppressor)	.6	.67	.402	>	.36	>	.336
(Distorter)	.2	-.13	-.026		.04		.016

$$\sum = R^2 = .376$$

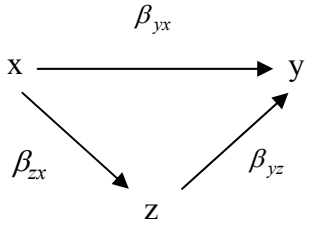
Dies zeigt, dass Aussagen über **Effekte** und Aussagen über **bereinigte Erklärungskraft** unabhängig voneinander variieren können:

Obwohl ein Effekt bei Kontrolle größer wird, muss der bereinigte Erklärungsanteil bei Kontrolle nicht wachsen.

j) **Vorzeichenregel**

Im Fall von 3 Variablen gibt es eine Vorzeichenregel analog zu Davis' Regel für die Tabellenanalyse:

Abbildung 3-27:



Gesamtzusammenhang von x auf y: r_{yx}

Bereinigter Zusammenhang von x auf y: β_{yx}

Rest-Zusammenhang: $\beta_{yz}\beta_{zx}$

Die Vorzeichenregel bezieht sich auf die Beziehung der Störvariablen z zu den Variablen x und y.

a) Falls der Rest- Zusammenhang kleiner Null ist, so haben die Beta-Koeffizienten von z zu x bzw. y unterschiedliche Vorzeichen.
Formal: $0 > \beta_{yz}\beta_{zx} \Rightarrow \text{sign}\left(\beta_{yz}\right) \neq \text{sign}\left(\beta_{zx}\right)$

b) Falls der Rest-Zusammenhang größer Null ist, so haben die Beta-Koeffizienten von z zu x bzw. y das gleiche Vorzeichen.
Formal: $0 < \text{res} = \beta_{yz}\beta_{zx} \Rightarrow \text{sign}\left(\beta_{yz}\right) = \text{sign}\left(\beta_{zx}\right)$

Bei der Darstellung der Typen ist jeweils angegeben worden, ob res größer oder kleiner als 0 ist, damit die Information zur Vorzeichenregel in der Darstellung der Typen enthalten ist.

Im Fall von mehr als 3 Variablen hat man zunächst die zusammenfassende Information, ob der Rest-Zusammenhang größer oder kleiner Null ist. Wie dieser Rest-Zusammenhang zustande kommt, kann dann in der Art analysiert werden, wie oben an Beispiel 1 demonstriert.

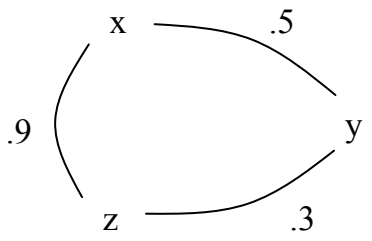
Hinweis: Die Vorzeichenregel gilt nicht für die einfachen Korrelationskoeffizienten statt der Beta-Koeffizienten:

Das Beispiel in (i) liefert:

Die Relation (x, y) ist Suppressor-Phänomen.
Also: $\beta_{yz}\beta_{zx} < 0$ (da $r_{yx} > 0$)
Aber: $r_{yz}r_{zx} > 0$

Würde man eine Typologie auf der Basis von ber = Part Correlation oder ber = Partial Correlation erstellen, so würde man auch keine entsprechende Vorzeichenregel für die einfachen Korrelationskoeffizienten erhalten:

Abbildung 3-28: Beispiel für Korrelationen



$$r_{xy.z} = \frac{.5 - .9 \cdot .3}{\sqrt{1 - .9^2}\sqrt{1 - .3^2}} = 0{,}55$$

$$r_{y,x-\hat{x}(z)} = \frac{.5 - .9 \cdot .3}{\sqrt{1 - .9^2}} = 0{,}53$$

Unter beiden Voraussetzungen handelt es sich um ein Suppressor-Phänomen mit ges > 0.

Also res < 0.
Aber: $r_{xz} \cdot r_{zy} > 0$

(Dies liegt daran, dass:
$r_{xy} = \sqrt{1-r_{xz}^2}\sqrt{1-r_{yz}^2}r_{xy.z} + r_{xz}r_{yz}$, $r_{xy} = \sqrt{1-r_{xz}^2}r_{y,x-\hat{x}(z)} + r_{xz}r_{yz}$,
d.h. die Gewichtung von ber ist jeweils zu berücksichtigen.)

D.h. von den naheliegenden Koeffizienten liefert keiner eine Vorzeichenregel für die einfachen Korrelationskoeffizienten, wie sie der Intuition in Analogie zur Tabellenanalyse entspricht. Die auf der Basis von Beta erstellte Typologie leistet dies für die Beta.

Die **Kovarianz** liefert eine Vorzeichenregel für die einfachen Korrelationskoeffizienten:
$r_{xy} = s_{xy} = s_{x-\hat{x}(z),\, y-\hat{y}(z)} + s_{\hat{x}(z),\hat{y}(z)} = s_{xy.z} + s_{xz}s_{yz} = s_{xy.z} + r_{xz}r_{yz}$

Oder: $s_{y,x-\hat{x}(z)} + r_{xz}r_{yz}$

$s_{xy.z}$ soll als Partial Covariance, $s_{y,x-\hat{x}(z)}$ als Part Covariance bezeichnet werden. Wählt man als bereinigten Effekt die Partial oder die Part Covariance, so folgt unmittelbar:
(res > 0) $\Leftrightarrow$ ($r_{xz}r_{yz} > 0$)

In der Vierfeldertafel ist die Kovarianz sogar eine normierte Maßzahl, sodass man dort nur mit der **Kovarianz** zu argumentieren braucht, vgl. S. 40-42.

Résumé:

Die Pfadanalyse wird häufig derart verwendet, dass nur die Beta-Koeffizienten als Einflussstärken interpretiert werden.

1) Interpretation der Beta-Koeffizienten

Dazu bedarf die Interpretation der Beta aber einer Begründung. Beta darf nicht interpretiert werden als Effekt von x_i, wobei die übrigen Prädiktoren „konstant gehalten" werden, denn letzteres widerspricht der möglichen Interkorrelation der Prädiktoren. Dagegen kann man in einer einfachen Regression Beta als Effekt des entsprechenden bereinigten Prädiktors $x_i - \hat{x}_i$ interpretieren. Dies hat den Vorteil, dass sich dem Prädiktor x_i die Gesamterklärungskraft r_{y,x_i}^2 und dem Prädiktor $x_i - \hat{x}_i$ die *bereinigte Erklärungskraft* $r_{y,x_i-\hat{x}_i}^2 = R^2 - R_{(i)}^2$ zuordnen lässt. Ein bereinigter Prädiktor $x_i - \hat{x}_i$ ist aber nicht standardisiert, weshalb die Effekte von $x_i - \hat{x}_i$ und $x_j - \hat{x}_j \left(j \neq j\right)$ nicht ohne weiteres vergleichbar sind. Die standardisierte bereinigte Variable $\frac{x_i - \hat{x}_i}{s_{x_i-\hat{x}_i}}$ hat den Effekt $r_{y,x_i-\hat{x}_i}$. (Der Erklärungsanteil von $x_i - \hat{x}_i$ und $\frac{x_i - \hat{x}_i}{s_{x_i-\hat{x}_i}}$ an y ist gleich, da Korrelationen invariant sind unter Linearkombinationen (also $r_{y,x_i-\hat{x}_i} = r_{y,\frac{x_i-\hat{x}_i}{s_{x_i-\hat{x}_i}}}$), nämlich gleich $r_{y,x_i-\hat{x}_i}^2 = R^2 - R_{(i)}^2$.)
Aber eine günstige Zerlegung des Gesamtzusammenhangs $\left(r_{y,x_i}\right)$ in bereinigten Zusammenhang und Restzusammenhang lässt sich in diesem Fall gerade wegen der Standardisierung wohl nicht

angeben. Beta² ist kein Anteil erklärter Varianz. Ferner gibt es keine standardisierte Variable, deren induzierte Änderung Beta ergäbe. Das Problem der Interkorrelation der Prädiktoren lässt sich aber auch so angehen, dass man zwischen direktem Effekt Beta und den gerade wegen der Interkorrelation vorliegenden korrelierten Zusammenhängen (Regression) bzw. indirekten Effekten (Pfadanalyse) unterscheidet. Betrachtet man r_{y,x_i} als Gesamtzusammenhang, so ist Beta eine mögliche Version der Bereinigung eines Zusammenhangs um den Einfluss der übrigen Prädiktoren (aus x_i). Weitere Möglichkeiten der Bereinigung wären Part und Partial Correlation, Part und Partial Covariance.

2) Typologie von Kausalstrukturen

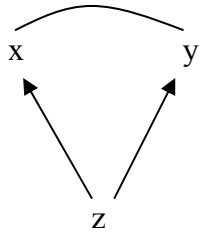
Es ist unangemessen, in einer Pfadanalyse nur mit den Beta zu argumentieren, da der Gesamtzusammenhang $\left(r_{y,x_i}\right)$ in jeder Richtung davon abweichen kann. Um nicht zu präjudizieren, ob Beta die günstigste Version der Bereinigung eines Zusammenhangs um den Einfluss der übrigen Prädiktoren ist, wird mit einer allgemeinen Zerlegung des Gesamtzusammenhangs in einen bereinigten Zusammenhang und einen Restzusammenhang gearbeitet. Die möglichen Abweichungen zwischen einem Gesamtzusammenhang und einem bereinigten Zusammenhang werden in einer *Typologie* vollständig erfasst (Übereinstimmung, Distorter, scheinbare Nicht-Kausalität, Schein-Kausalität, Überschneidung, Suppressor). Mit diesem Instrumentarium kann dann (für Bereinigung = Beta) geklärt werden, unter welchen Bedingungen die Summe der *bereinigten Erklärungskraft* der Prädiktoren (Part Correlation² $= R^2 - R^2_{(i)}$) und die Summe der Gesamterklärungskraft $\left(r^2_{y,x_i}\right)$ größer oder kleiner als Multiple R² sind. Die Bedingungen werden an drei Beispielen demonstriert.

3) *Effekt* versus *bereinigte Erklärungskraft*

Anschließend wird (in Punkt (i)) an einem Beispiel (für bereinigter Effekt = Beta) gezeigt, dass Aussagen über Effekte und Aussagen über Erklärungskraft unabhängig voneinander variieren können. Dies zeigt, dass man sich bei der Interpretation nicht ohne Informationsverlust entweder auf Beta oder auf Part Correlation beschränken kann.
Im Fall von 3 Variablen liefert die Wahl von Beta als bereinigtem Zusammenhang eine Vorzeichenregel für die Beta ähnlich wie in der Tabellenanalyse. Diese Vorzeichenregel existiert jedoch allein für die Beta, nicht für die einfachen Korrelationskoeffizienten. Die Verwendung von Part oder Partial Correlation als Bereinigung liefert nicht die Vorzeichenregel für die einfachen Korrelationskoeffizienten. Die Verwendung von Part oder Partial Covariance liefert die Vorzeichenregel für die einfachen Korrelationskoeffizienten genau wie in der Tabellenanalyse. Dazu muss aber kritisch angemerkt werden, dass man die Gesamtkorrelation mit der partiellen Kovarianz vergleicht, um zu entscheiden, ob die einfachen Korrelationen mit dem Testfaktor gleiches oder ungleiches Vorzeichen haben. D.h. hier wird gleichzeitig mit Kovarianz und Korrelation argumentiert, was inkonsistent ist. Die Vorzeichenregel der Tabellenanalyse ist nach meiner Auffassung nur konsistent, wenn man nur mit Kovarianzen argumentiert, was ja – wie gezeigt – möglich ist. Die Vorzeichenregel für die Beta weist die gleiche Konsistenz auf. Ferner ist die Zerlegung des Gesamtzusammenhangs in bereinigten und Rest-Zusammenhang bei mehr als zwei Prädiktoren für Part oder Partial Covariance weniger günstig als für Beta. Die Interpretation des Beta-Koeffizienten als *direktem Effekt* scheint mir von den betrachteten Möglichkeiten noch der günstigste Lösungsversuch der Bereinigung einer Variablen um den Einfluss der übrigen Prädiktoren. Davon unabhängig variiert noch die *bereinigte Erklärungskraft* $\left(r^2_{y,x_i-\hat{x}_i} = R^2 - R^2_{(i)}\right)$ eines Prädiktors. Es ist demnach nicht möglich, den Problemkreis nur mit einem der beiden Konzepte adäquat zu behandeln.

3.2.8 Partielle Korrelation oder Pfadkoeffizient?

Abbildung 3-29:



Die Korrelation zwischen x und y ist ungleich Null.

a) Beispiel einer **nur scheinbaren Kausalität**: $r_{xy.z} = 0$

Abbildung 3-29*:* zeigt die graphische Darstellung einer reinen „Scheinkorrelation", d.h. einer nur scheinbaren Kausalität. Das Vorliegen einer reinen „Scheinkorrelation" lässt sich in bestimmtem Ausmaß untersuchen durch den partiellen Korrelationskoeffizienten: Im Fall einer reinen „Scheinkorrelation" muss notwendig gelten:

(Seien ohne Beschränkung der Allgemeinheit x, y, z standardisiert.)

$\left(0 = r_{xy.z} = \frac{s_{x-\hat{x}(z),\,y-\hat{y}(z)}}{s_{x-\hat{x}(z)} s_{y-\hat{y}(z)}} = \frac{s_{x-\hat{x}(z),\,y}}{s_{x-\hat{x}(z)} s_{y-\hat{y}(z)}}\right)$ genau dann, wenn $\left(s_{x-\hat{x}(z),y} = 0\right)$

Also ist es äquivalent, dass: $\beta_{yx} = \frac{s_{y,\,x-\hat{x}(z)}}{s^2_{x-\hat{x}(z)}} = 0$

Warnung: Diese Bedingungen sind nicht hinreichend für das Vorliegen einer echten Scheinkorrelation.

Beispiel (nach Wright 1934: 190, Li 1977: 171):

Zwischen beobachteten Variablenwerten möge als Ergebnis vorliegen: $r_{xz} = 0{,}50 \;\; r_{zy} = 0{,}50 \;\; r_{xy} = 0{,}25$

Dann ergibt sich: $r_{xy.z} = \frac{r_{xy} - r_{xz}\, r_{yz}}{\sqrt{1 - r_{xz}^2}\sqrt{1 - r_{yz}^2}} = \frac{0{,}25 - 0{,}50 \cdot 050}{Nenner} = 0$

Man ist nun versucht zu schließen, dass z für die Korrelation zwischen x und y verantwortlich ist. Da Scheinkorrelation und intervenierende Variable bekanntlich statistisch nicht zu unterscheiden sind, könnte man an folgende Modelle denken:

Abbildung 3-30: Modell A

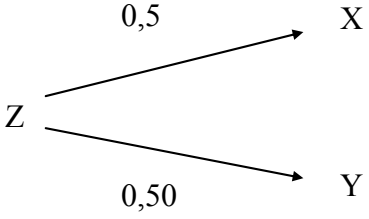
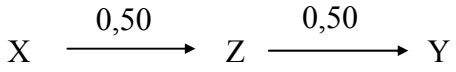
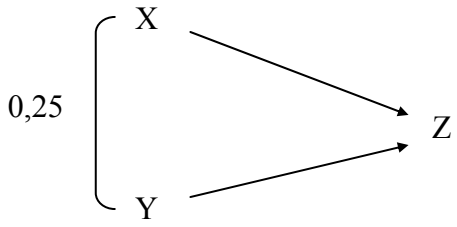


Abbildung 3-31: Modell B



So könnte man zu einem Fehlschluss kommen, wenn nämlich das richtige Modell die Form des Modells C hätte:

Abbildung 3-32: Modell C



(Denn nach dem Pfadtheorem: $r_{zx} = \beta_{zx} + r_{xy}\,\beta_{zy} = 0{,}40 + 0{,}25 \cdot 0{,}40 = 0{,}50$

r_{zy} analog)

In diesem Modell ist z überhaupt nicht verantwortlich für die Korrelation zwischen x und y.

Diese Schwäche der Argumentation mit dem symmetrischen Maß der partiellen Korrelation war für Wright eine der Anlässe, die Pfadanalyse mit den gerichteten (asymmetrischen) Pfadkoeffizienten zu entwickeln.

Wäre nun das Argumentieren mit Beta vorteilhafter?

Modell A: Geht x der Variablen y kausal voran, so ist bei der Datenlage $\beta_{yx} = 0$ und Modell A verträglich mit den Daten.

Modell B: Bei der Datenlage ist $\beta_{yx} = 0$ und Modell B mit den Daten verträglich.

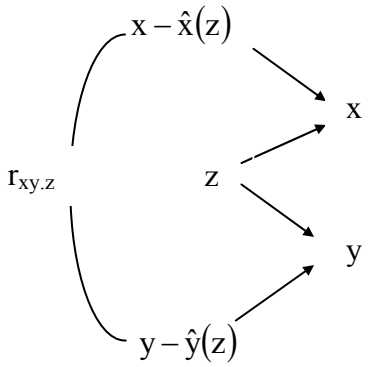
Modell C: In Modell C ist die Kausalstruktur zwischen x und y offen gelassen bzw. es gibt keinen kausalen Einfluss von x auf y oder von y auf x. (Ein realistisches Beispiel: z = Schulbildung eines Befragten, x = Schulbildung des Vaters, y = Schulbildung der Mutter). In diesem Fall ist ein Pfadkoeffizient von x nach y nicht definiert, da in dem Modell gar kein Pfad von x nach y vorgesehen ist.

Es ist also nicht der Pfadkoeffizient von x nach y (berechnet als β_{yx}), der eine Verbesserung gegenüber der Berechnung von $r_{xy.z}$ bringt, sondern die Pfadanalyse insgesamt als Untersuchung, welche Kausalmodelle mit beobachteten Daten verträglich sind. Dadurch ist die Situation besser einzuschätzen, auch wenn eventuell bei gegebener Datenlage noch nicht zwischen verschiedenen möglichen Modellen zu entscheiden ist.

b) Weiteres Beispiel für die Vorzüge der Pfadanalyse

Dass die Pfadanalyse ein flexibles Instrumentarium ist, lässt sich auch in Abbildung 3-33 ablesen, in der die *partielle Korrelation* zwischen x und y unter Kontrolle eines verursachenden Störfaktors z *pfadanalytisch* dargestellt ist als Korrelation von Residualfaktoren. Dieses Diagramm ließe sich leicht erweitern auf die Situation, dass mehrere Störfaktoren z_1, ..., z_m herauszurechnen sind. Solche Korrelations- oder Kovarianzzerlegungen bilden die analytische Grundlage der Mehrebenenanalyse.

Abbildung 3-33:



3.2.9 Relative Bedeutung von Multiple R² und den Effekten für die Erklärung

Multiple R² kann im Sinne der Kausalanalyse auch „Scheinkomponenten“ der Erklärung enthalten. Insofern ist für die Erklärung die Größenordnung der Effekte noch wichtiger als die „Gesamterklärungskraft“ Multiple R².

Literaturverzeichnis

Alwin, D.F., Hauser, R.M., 1975: *The decomposition of effects in path analysis*. In: American Sociological Review 40: 37-47.

Asher, H.B., 1987: *Causal modeling*. Beverly Hills: Sage.

Blalock, H.M., 1964: *Causal Inferences in Nonexperimental Research*. Chapel Hill: University of North Carolina Press.

Blalock, H.M., 1969: *Theory construction*: *From Verbal to Mathematical Formulations*. New York: Englewood Cliffs.

Blalock, H.M. (Hg.), 1985[2]: *Causal models in social sciences*. Chicago: Aldine.

Blau, P.M., Duncan, O.D., 1967: *The American occupational structure*. New York: Wiley.

Bortz, J., 2005[6]: *Statistik für Human- und Sozialwissenschaftler*. Berlin: Springer.

Davis, J.A., 1971: *Elementary survey analysis*. Englewood Cliffs, New Jersey: Prentice Hall.

Duncan, O.D., 1966: *Path analysis: Sociological examples*. In: American Journal of Sociology 72: 1-16.

Duncan, O.D., 1975: *Introduction to structural equation models*. New York: Academic Press.

Finney, J.M., 1972: *Indirect effects in path analysis*. In: Sociological Methods & Research 1: 175-186.

Goldberger, A.S., 1964: *Econometric theory*. New York: Wiley.

Goldberger, A.S., Duncan, O.C., 1973: *Structural equation models in the social sciences*. New York: Seminar Press.

Heise, D.R., 1975: *Causal analysis*. New York: Wiley.

Holm, K., 1977: *Lineare multiple Regression und Pfadanalyse*. In: Ders. (Hg.): Die Befragung. Band 5. München: UTB.

Hummell, H.J., Ziegler, R. (Hg.), 1982: *Korrelation und Kausalität*. 3 Bände. Stuttgart: Enke.

Johnston, J., 1996[4]: *Econometric Methods*. New York: McGraw-Hill/Irwin.

Land, K.C., 1969: *Principles of path analysis*. In: Borgatta, E.F. (Hg.): *Sociological Methodology*. San Francisco: Jossey-Bass.

Lazarsfeld, P.F., 1961: *The Algebra of Dichotomous Systems*. In: Solomon, H. (Hg.): *Item Analysis and Prediction*. Stanford: Stanford University Press, 111-157.

Li, C.C., 1977[2]: *Path Analysis: A Primer*. Pacific Grove, Californa: The Boxwood Press.

Namboodiri, N.K. et al., 1975: *Applied multivariate analysis and experimental designs*. New York: McGraw Hill.

Nie, N.H. et al., 1975[2]: *Statistical package for the social sciences* (SPSS). New York: McGraw-Hill.

Opp, K.-D., Schmidt, P., 1976: *Einführung in die Mehrvariablenanalyse*. Grundlagen der Formulierung und Prüfung komplexer sozialwissenschaftlicher Aussagen. Reinbek bei Hamburg: Rowohlt.

Rosenberg, M., 1968: *The Logic of Survey Analysis*. New York: Basic Books.

Simon, H., 1954: *Spurious correlation: A causal interpretation*. In: Journal of American Statistical Association 49: 467-479.

Theil, H., 1971: *Principles of econometrics*. New York: Wiley.

Theil, H., 1978: *Introduction to econometrics*. Englewood Cliffs, New Jersey: Prentice Hall.

Van de Geer, J.P., 1971: *Introduction to Multivariate Analysis for the Social Sciences*. San Francisco: Freeman.

Wonnacott, R.J., Wonnacott, T.H., 1979[2]: *Econometrics*. New York: Wiley.

Wright, S., 1934: *The method of path coefficients*. In: Annals of Math. Stat. 5: 161-215.

Wright, S., 1960: *Path coefficients and path regressions: alternative or complementary concepts?* In: Biometrics 16: 189-202.

4. Varianzanalyse und Kovarianzanalyse

In der von Fisher entwickelten Varianzanalyse wird versucht, die Varianz einer abhängigen metrischen Variablen y durch unabhängige, nominale Variablen zu erklären. Handelt es sich nur um eine unabhängige Variable, so spricht man von einfacher Varianzanalyse, bei zwei unabhängigen Variablen von zweifacher Varianzanalyse etc.

Ein **Beispiel** soll die Bedeutung der Varianzanalyse veranschaulichen:
Das Nettoeinkommen in der Bundesrepublik beträgt gemäß im Durchschnitt 1724,- €. Hat die Tatsache, dass die Personen unterschiedliche Stellungen im Beruf haben, einen Einfluss? Lassen sich dadurch Unterschiede in der Vergütung in einem statistischen Sinne erklären? Lässt sich der Fehler in der Vorhersage des Einkommens (abhängige Variable) durch die Kenntnis der Stellung im Beruf (unabhängige Variable) reduzieren?

Zur Notation:

Die erklärende Gruppeneinteilung umfasst k = 7 Gruppen.
Der laufende Index der Gruppen ist i = 1, ..., k.
Der Umfang der Gruppe i beträgt n_i.
Die Gesamtzahl der Untersuchungseinheiten umfasst: $\sum_{i=1}^{k} n_i = n$
Der laufende Index innerhalb der Gruppen ist j = 1, ..., n_i.
Der Wert der zu erklärenden Variablen y für Personen j in Gruppe i wird als y_{ij} bezeichnet.

Die beste Schätzung der Beobachtungswerte y_{ij} im Sinne der Minimierung des Fehlers $\sum_i \sum_j \left(y_{ij} - \hat{y}\right)^2$ ist der Mittelwert $\hat{y} = \bar{y}$,weil die quadratische Abweichung bezogen auf einen beliebigen Bezugspunkt nie kleiner sein kann als die quadratische Abweichung bezogen auf den Mittelwert.

Entsprechend ist in den verschiedenen Bereichen der Mittelwert des Bereichs die beste Schätzung:

Tabelle 4-1: Einkommen erklären durch Stellung im Beruf

BFR.: NETTOEINKOMMEN <OFFEN+LISTENANGABE> (Allbus 2002)

BEFR.:JETZIGE BERUFLICHE	Mittelwert	N	Standardabweichung
LANDWIRT	1586,00	5	879,157
AKADEM. FREIER BERUF	4530,18	17	2907,772
SONST. SELBSTSTÄNDIGE	2469,74	109	2032,131
BEAMT; RICHTER, SOLDAT	2425,13	85	1200,451
ANGESTELLTE	1730,53	659	1300,663
ARBEITER	1297,61	340	476,849
IN AUSBILDUNG	535,50	40	284,284
Insgesamt	1723,82	1255	1340,490

Der Fehler dieser Schätzung entsteht durch die Abweichung der Beobachtungswerte vom jeweiligen Gruppenmittelwert. Der Gesamtfehler dieser Schätzung wird durch die Summe der Abweichungsquadrate zwischen den y-Werten der i-ten Gruppe und dem Mittelwert $\bar{y}_i$ dieser Gruppe erfasst, summiert über alle k Gruppen: $\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_i\right)^2$. Der erste Index i repräsentiert die Gruppe, der zweite Index j den Messwert innerhalb der i-ten Gruppe. Der Fehler dieser Schätzung ist geringer als der Fehler der Schätzung $\bar{y}$, weil:

$$\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_i\right)^2 \le \sum_{j=1}^{n_i}\left(y_{ij}-z\right)^2 \quad \text{(für alle z)}$$

Der multiple Regressionskoeffizient R^2 = Multiple $\eta^2 = \dfrac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}\right)^2 - \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_i\right)^2}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}\right)^2}$ ist die proportionale Reduktion des Fehlers der Schätzung der abhängigen Variablen y durch Kenntnis der nominalen unabhängigen Variablen mit den Ausprägungen i = 1, ..., k, nach denen die Gruppeneinteilung vorgenommen wurde. Dies ist der Anteil der Varianz von y, der durch die Gruppeneinteilung erklärt wird (s.u.):

$$\text{Multiple } R^2 = \text{Multiple } \eta^2 = \frac{\sum_{i=1}^{k} n_i\left(\bar{y}_i-\bar{y}\right)^2}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}\right)^2} = \frac{\textit{Erklärte Variation}}{\textit{Gesamtvariation}}$$

In dem Beispiel beträgt Multiple R^2 = Multiple η^2 = 0,157, d.h. 15,7 % der Variation des Einkommens lässt sich durch die Stellung im Beruf statistisch erklären. Oder: Der Fehler in der Vorhersage des Einkommens wird durch die Kenntnis der Stellung im Beruf um 15,7 % reduziert.

4.1 Einfache Varianzanalyse als Verallgemeinerung des t-Tests

Die einfaktorielle Varianzanalyse, d.h. die Varianzanalyse auf Basis eines Erklärungsfaktors, ist eine Verallgemeinerung des t-Tests zur Prüfung des Mittelwertunterschieds zweier unabhängiger Stichproben auf k Stichproben (k > 2). Der Vorteil der Varianzanalyse gegenüber den Einzelvergleichen des t-Tests besteht darin, dass beliebig viele Mittelwerte gleichzeitig miteinander verglichen werden.

4.1.1 Varianzzerlegung

In der Varianzanalyse wird vorausgesetzt, dass k unabhängige Stichproben aus Grundgesamtheiten vorliegen, die nach den Normalverteilungen N (μ_i, σ_i^2) verteilt sind, wobei alle Varianzen gleich sind ($\sigma_1^2 = ... = \sigma_k^2 =: \sigma^2$). Ob die Homogenität der Varianzen gegeben ist, kann mit dem Bartlett-Test geprüft werden (vgl. Clauß/Finze/Partzsch 1994: 272-274).

Die Anwendung besteht darin, dass die Gruppeneinteilung aufgrund der k Ausprägungen eines nominalen Merkmals A vorgenommen wird.

Das Modell lässt sich allgemein wie folgt charakterisieren: $y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$

Hierbei sind: μ der Mittelwert der Gesamt-Grundgesamtheit, $\alpha_i = \mu_i - \mu$ die Wirkungen des Faktors A, ε_{ij} die Fehlerterme.

Für inferenzstatistische Überlegungen muss die Voraussetzung erfüllt sein, dass ε_{ij} unabhängige Zufallsvariablen sind, die nach N $(0, \sigma^2)$ verteilt sind ($\Leftrightarrow$ y_{ij} nach N (μ_i, σ^2) verteilt).

Fasst man die k Stichproben mit ihren Mittelwerten $\bar{y}_i$ und Varianzen s_i^2 zusammen, so erhält man als Mittelwert und Varianz der Gesamtstichprobe:

$$\sum_{i=1}^{k} n_i \frac{\bar{y}_i}{n} = \sum_{i=1}^{k}\sum_{j=1}^{n_i} \frac{y_{ij}}{n} = \bar{y} \qquad \left(\text{wobei } \bar{y}_i = \frac{\sum_{j=1}^{n_i} y_{ij}}{n_i} \right)$$

$$\sum_{i=1}^{k} \frac{(n_i - 1) s_i^2}{(n-1)} = \sum_{i=1}^{k}\sum_{j=1}^{n_i} \frac{(y_{ij} - \bar{y})^2}{(n-1)} = s^2$$

Basis der Varianzanalyse ist folgende Streuungszerlegung:

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 \quad = \quad \sum_{i=1}^{k}\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 \quad + \quad \sum_{i=1}^{k} n_i (\bar{y}_i - \bar{y})^2$$

Gesamtstreuung	Streuung innerhalb der Gruppen (= durch die Einteilung in Gruppen nicht erklärte Streuung)	Streuung zwischen den Gruppen (= durch die Einteilung in Gruppen erklärte Streuung)
SS_y	SS_{Fehler}	SS_A
SS_{gesamt}	$SS_{innerhalb\ der\ Gruppen}$	$SS_{zwischen\ den\ Gruppen}$

Die Gesamtstreuung des unabhängigen Merkmals wird in die Streuung innerhalb der Gruppen (Ausprägungen des Faktors) und zwischen den Gruppen zerlegt. Wenn die Streuung der abhängigen Variablen völlig auf die Unterschiede der unabhängigen Variablen zurückgeführt werden könnte und demzufolge keine Varianzunterschiede auf die Streuung innerhalb der Gruppen zurückzuführen wäre, würde die erklärte Varianz 100 % betragen. Die erklärte Varianz ist ein anschaulicher Maßstab für die Erklärungskraft.

Γ

Beweis der Gleichung:

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i} (y_{ij} - \bar{y}) = \sum_{i=1}^{k}\sum_{j=1}^{n_i} \left[(y_{ij} - \bar{y}_i) + (\bar{y}_i - \bar{y}) \right]^2$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 + 2\sum_{i=1}^{k}\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)(\bar{y}_i - \bar{y}) + \sum_{i=1}^{k}\sum_{j=1}^{n_i} (\bar{y}_i - \bar{y})^2 \quad \text{, wobei :}$$

$$2\sum_i \sum_j (y_{ij} - \bar{y}_i)(\bar{y}_i - \bar{y}) = 2\sum_i (\bar{y}_i - \bar{y}) \underbrace{\sum_j (y_{ij} - \bar{y}_i)}_{0} = 0$$

Deshalb:

$$\underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}\right)^2}_{SS_y} = \underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_i\right)^2}_{SS_{Fehler}} + \underbrace{\sum_{i=1}^{k} n_i\left(\bar{y}_i-\bar{y}\right)^2}_{SS_A}$$

L

Tabelle 4-2: Varianzanalyse-Tabelle

	Quadratsummen (SS) "Sum of Squares"	Freiheitsgrade (df)	Varianz ("Mean Square")
Erklärt durch die Gruppeneinteilung (SS_A)	$\sum_{i=1}^{k} n_i\left(\bar{y}_i-\bar{y}\right)^2$	k - 1	$\sum_{i=1}^{k} n_i \frac{\left(\bar{y}_i-\bar{y}\right)^2}{\left(k-1\right)}$
Nicht erklärt durch die Gruppeneinteilung (SS_{Fehler})	$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_i\right)^2$	n - k	$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\frac{\left(y_{ij}-\bar{y}_i\right)^2}{\left(n-k\right)}$
Insgesamt (SS_y)	$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}\right)^2$	n - 1	$\frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}\right)^2}{\left(n-1\right)}$

Die einzelnen Varianzen erhält man, indem man die Quadratsummen durch ihre Freiheitsgrade (df) dividiert.

Für SS_y ist die Zahl der Freiheitsgrade durch folgende Beziehung festgelegt:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}$$

Bei Kenntnis von können n - 1 y-Werte frei variieren, der n-te Wert aber ist dann durch die Beziehung bestimmt und ermittelbar.

$df\,(SS_y) = n - 1$

Für SS_{Fehler} ergeben sich aufgrund der Beziehung $\bar{y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$ innerhalb jeder Gruppe n_i - 1 Freiheitsgrade. Über alle Gruppen addiert erhält man:

$$df\,(SS_{Fehler}) = \sum_{i=1}^{k}\left(n_i - 1\right) = n - k$$

Die Zahl der Freiheitsgrade von SS_A wird durch die Beziehung $\bar{y} = \frac{1}{k}\sum_{i=1}^{k}\bar{y}_i$ bestimmt; bei Kenntnis von $\bar{y}$ können k - 1 Werte frei variieren: $df\,(SS_A) = k - 1$

Analog zu den Quadratsummen ergibt sich für die Freiheitsgrade folgende Beziehung:

$df\,(SS_y) = df\,(SS_{Fehler}) + df\,(SS_A)$

Das Verhältnis von erklärter und nichterklärter Varianz ist $F_{k-1,\,n-k}$-verteilt:

$$F = \frac{\sum_{i=1}^{k} n_i \frac{(\bar{y}_i - \bar{y})^2}{(k-1)}}{\sum_{i=1}^{k}\sum_{j=1}^{n_i} \frac{(y_{ij} - \bar{y}_i)^2}{(n-k)}} = \frac{Mean\ Square_A}{Mean\ Square_{Fehler}}$$

4.1.2 Signifikanztest

Nach der Bestimmung des Anteils erklärter Varianz bleibt zu fragen, ob die Varianzerklärung zufällig oder aufgrund der Unterschiede der Gruppen zustande gekommen ist. Die Nullhypothese geht davon aus, dass die Gruppenmittelwerte der Grundgesamtheiten gleich sind bzw. dass der Faktor A (nominalskalierte Größe) keinen Effekt auf die zu erklärende Variable (abhängige metrisch skalierte Größe) hat.

H_0: $\mu_1 = \mu_2 = ... \mu_k = \mu$ bzw. $\alpha_1 = \alpha_2 = ... \alpha_k = 0$

Die Nullhypothese (H_0) wird gegen die Alternativhypothese (H_1) getestet:

H_1: $\mu_i \neq \mu_j$ für mindestens zwei der Mittelwerte
$\alpha_i \neq 0$ für wenigstens ein α_i

Ausgehend von den ermittelten Varianzen (siehe Varianztabelle) wird ein Quotient gebildet. Dessen Zähler ist die Varianz zwischen den Gruppen. Sie ist immer dann ein erwartungstreuer Schätzwert der Varianz der Grundgesamtheit (d.h. Erwartungswert des Stichprobenmittelwerts = Mittelwert der Grundgesamtheit), wenn die Nullhypothese zutrifft. Im Nenner des Quotienten steht die Varianz innerhalb der Gruppen. Unabhängig von der Gültigkeit der Nullhypothese schätzt sie die Varianz der Grundgesamtheit immer erwartungstreu. Bei Gültigkeit der Nullhypothese werden die beiden unabhängigen Varianzen – von Zufallsabweichungen abgesehen – gleich sein. Durch den F-Test wird geprüft, ob sich die beiden Schätzwerte signifikant unterscheiden. Übersteigt der Quotient den für das zuvor festgelegte Signifikanzniveau kritischen Wert, wird die Nullhypothese abgelehnt.

$$F = \frac{Mean\ Square_A}{Mean\ Square_{Fehler}} = \frac{\sum_{i=1}^{k} n_i \frac{(\bar{y}_i - \bar{y})^2}{k-1}}{\sum_{i=1}^{k}\sum_{j=1}^{n_i} \frac{(y_{ij} - \bar{y}_i)^2}{n-k}}$$ vergleichen mit $F_{k-1,\,n-k,\,1-\alpha}$.

Folgende drei Punkte sollen die inferenzstatistischen Grundlagen bei einer einfaktoriellen Varianzanalyse erläutern.

1) Aus den Schätzungen $s_i^2,...,s_k^2$ für σ^2 (wobei die Schätzungen erwartungstreu sind: E $\left(s_i^2\right) = \sigma^2$) kann man eine verbundene Schätzung bilden:

$$\hat{s}_a^2 := \frac{\sum_{i=1}^{k}(n_i - 1)s_i^2}{\sum_{i=1}^{k}(n_i - 1)} = \frac{\sum_{i=1}^{k}(n_i - 1)s_i^2}{n-k} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2}{n-k}$$

Diese Schätzung ist ebenfalls erwartungstreu: $E\left(\hat{s}_a^2\right) = \frac{(n-k)\sigma^2}{n-k} = \sigma^2$

Die verbundene Schätzung hat $\sum_{i=1}^{k}(n_i - 1) = n - k$ Freiheitsgrade, da ein Freiheitsgrad jeweils durch Kenntnis des Mittelwerts verloren geht.[9]

2) Nun soll der Erwartungswert von $\sum_{i=1}^{k} n_i(\bar{y}_i - \bar{y})^2$ berechnet werden.

Mit E (f) = μ gilt nach dem Verschiebungssatz: E $[(f - \mu)^2] = E(f^2) - \mu^2$

Also: $\sum_{i=1}^{k} \frac{n_i}{n}(\bar{y}_i - \bar{y})^2 = \sum_{i=1}^{k} \frac{n_i}{n}\bar{y}_i^2 - \bar{y}^2$ (*)

Nach der Stichprobentheorie gilt: $\sigma_{\bar{y}}^2 = \frac{\sigma^2}{n}, \sigma_{\bar{y}_i}^2 = \frac{\sigma^2}{n_i}$

Der Verschiebungssatz liefert also für $\bar{y}$ und $\bar{y}_i$:

$$\frac{\sigma^2}{n} = \sigma_{\bar{y}}^2 = E(\bar{y}^2) - \left(\sum_{i=1}^{k} \frac{n_i}{n}\mu_i\right)^2$$

$$\frac{\sigma^2}{n_i} = \sigma_{\bar{y}_i}^2 = E(\bar{y}_i^2) - \mu_i^2$$

Daraus folgt:

$$E\left(\sum_{i=1}^{k} n_i(\bar{y}_i - \bar{y})^2\right) \overset{*}{=} \sum_{i=1}^{k} n_i E(\bar{y}_i^2) - n \cdot E(\bar{y})^2$$

$$= \sum_{i=1}^{k} n_i\left(\frac{\sigma^2}{n_i} + \mu_i^2\right) - \left(\frac{\sigma^2}{n} + \left(\sum_{i=1}^{k} \frac{n_i}{n}\mu_i\right)^2\right)$$

$$= (k-1)\sigma^2 + \sum_{i=1}^{k} n_i\,\mu_i^2 - n\left(\sum_{i=1}^{k} \frac{n_i}{n}\mu_i\right)^2$$

$$\overset{*}{=} (k-1)\sigma^2 + \sum_{i=1}^{k} n_i\left(\mu_i - \sum_{j=1}^{k} \frac{n_j}{n}\mu_j\right)^2$$

[9] Allgemein gilt: $\sum (x_i - \bar{x})^2 = \sum x_i^2 - n \cdot \bar{x}^2$

Also: $\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 + n\bar{x}^2$

df = n df = n - 1 df = 1

Unter der Annahme der Nullhypothese: H_0: $\mu_1 = ... = \mu_k$ entfällt der zweite Summand in der berechneten Formel, so dass man erhält:

$$E\left(\sum_{i=1}^{k} n_i \frac{(\bar{y}_i - \bar{y})^2}{(k-1)}\right) = \sigma^2$$

Die Anzahl der Freiheitsgrade dieser Schätzung von σ^2 beträgt k - 1, da ein Freiheitsgrad durch Kenntnis des Gesamtmittelwerts verloren geht.

3) Der Quotient der beiden Schätzungen von σ^2 ist verteilt nach $F_{k-1,\,n-k}$.

a) $\frac{\bar{y}_i - \mu}{\frac{\sigma}{\sqrt{n_i}}}$ ist unter Annahme der Nullhypothese N (0, 1)-verteilt; weil es sich um unabhängige Stichproben handelt, ist $\sum_{i=1}^{k} \frac{(\bar{y}_i - \mu)^2}{\frac{\sigma^2}{n_i}} \chi_k^2$ -verteilt.

Ersetzt man μ durch die Schätzung $\bar{y}$, so geht ein Freiheitsgrad verloren, da sich z.B. $\bar{y}_k$ aus $\bar{y}_1, ..., \bar{y}_{k-1}$ und $\bar{y}$ berechnen lässt, d.h. nicht mehr frei variieren kann.

Also: $\sum_{i=1}^{k} n_i \frac{(\bar{y}_i - \bar{y})^2}{\sigma^2} ist \chi_{k-1}^2$ -verteilt.

b) $(n_i - 1)\frac{s_i^2}{\sigma^2} \sum_{j=1}^{n_i} \frac{(y_{ij} - \bar{y}_j)^2}{\sigma_2}$ ist $\chi_{n_i-1}^2$ -verteilt; weil es sich um unabhängige Stichproben handelt, ist $\sum_{i=1}^{k} (n_i - 1)\frac{s_i^2}{\sigma^2}$ verteilt nach χ_{n-k}^2.

c) Da für die Normalverteilung $\bar{y}_i$ und s_1^2 unabhängig sind, ist der Quotient

$$\frac{\frac{\sum_{i=1}^{k} n_i (\bar{y}_i - \bar{y})^2}{(k-1)}}{\frac{\sum_{i=1}^{k} (n_i - 1) s_i^2}{(n-k)}}$$

als Quotient einer χ_{k-1}^2 - und einer χ_{n-k}^2 -verteilten Größe selbst verteilt nach $F_{k-1,\,n-k}$.

Ist der beobachtete F-Wert größer als der kritische Wert aus der Tabelle bei gegebenem Signifikanzniveau und den Freiheitsgraden (k - 1, n - k), so wird die Nullhypothese abgelehnt. D.h., dass der Faktor A (Gruppeneinteilung) einen signifikanten Beitrag zur Erklärung der abhängigen Variablen y leistet. Anderenfalls wird die Nullhypothese, dass der Faktor A (Gruppeneinteilung) keinen Einfluss hat, beibehalten.

4.1.3 Einfache Varianzanalyse und t-Test

Speziell für k = 2 ist $\sum_{i=1}^{k} n_i(\bar{y}_i - \bar{y})^2$ bis auf eine Konstante gleich $(\bar{y}_1 - \bar{y}_2)^2$. Das zeigt, dass die Testgröße für k Stichproben eine Verallgemeinerung des t-Tests ist.

Denn: $n_1\left(\bar{y}_1 - \frac{n_1\bar{y}_1 + n_2\bar{y}_2}{n}\right)^2 = n_1\left(\frac{n_2(\bar{y}_1 - \bar{y}_2)}{n}\right)^2$

Entsprechend für den 2. Term, insgesamt also:

$$\sum = \frac{n_1 n_2^2 + n_2 n_1^2}{n^2} \cdot (\bar{y}_1 - \bar{y}_2)^2 = \frac{n_1 n_2}{n_1 + n_2}(\bar{y}_1 - \bar{y}_2)^2 = \frac{1}{\frac{1}{n_1} + \frac{1}{n_2}}(\bar{y}_1 - \bar{y}_2)^2$$

Aus der Inferenzstatistik ist bekannt, dass das Quadrat einer t_{n-2}-verteilten Größe $F_{1, n-2}$-verteilt ist.

Im SPSS-Output wird im Fall k = 2 die Prüfgröße T des t-Tests ausgedruckt, anstatt η^2. Es gilt dann der Zusammenhang: $T^2 = F$.

Beweis:

$$T^2 = \frac{(\bar{y}_1 - \bar{y}_2)}{\sum_{i=1}^{2} \frac{(n_i - 1)s_i^2}{(n-2)}} \cdot \frac{1}{\frac{1}{n_1} + \frac{1}{n_2}} = \frac{\sum_{i=1}^{2} n_i(\bar{y}_i - \bar{y})^2}{\sum_{i=1}^{2} \frac{(n_i - 1)s_i^2}{(n-2)}} = F \quad \text{für den Fall k = 2.}$$

4.1.4 Anteil erklärter Varianz als Deskription

Multiple $R^2 = \frac{SS_A}{SS_y} = \frac{SS_{between}}{SS_y}$ ist der Anteil der durch die Gruppeneinteilung nach Merkmal A erklärten Varianz von y. In diesem Fall einer nominalen unabhängigen Variablen spricht man auch von Multiple η^2 (**Eta-Quadrat**) statt **Multiple R^2**. Diese deskriptive Kennzeichnung der Erklärungskraft erfordert keine Voraussetzung über die Verteilung oder die Gleichheit der Varianzen.

4.2 Zweifache Varianzanalyse

In der zweifachen Varianzanalyse wird der Effekt zweier unabhängiger nominaler Variablen A und B auf die abhängige metrische Variable y untersucht. Das Modell lässt sich allgemein wie folgt charakterisieren:
$y_{iju} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{iju}$

Hierbei sind:
μ der Mittelwert der Gesamt-Grundgesamtheit,
$\alpha_i = \mu_i - \mu$ die Wirkungen des Faktors A,
$\beta_j = \mu_j - \mu$ die Wirkungen des Faktors B,
$(\alpha\beta)_{ij} = \mu_{ij} - \mu - \alpha_i - \beta_j$ die (um die Effekte von A und B bereinigten) Wirkungen der Interaktionen,
ε_{iju} die Fehlerterme.

Für inferenzstatistische Überlegungen muss die Voraussetzung erfüllt sein, dass ε_{iju} unabhängige Zufallsvariablen sind, die nach N $(0, \sigma^2)$ verteilt sind ($\Leftrightarrow$ y_{iju} verteilt nach N (μ_{ij}, σ^2).

Zweifaktorielle Varianzanalysen teilen je nach Anzahl der Ausprägungen der beiden Faktoren die Befragten in mindestens 2 x 2 Gruppen ein. Sind alle Gruppen gleich groß, wird anders verfahren als bei unterschiedlicher Gruppengröße, was die nachfolgenden beiden Abschnitte erläutern.

4.2.1 Gleiche Zellenhäufigkeiten

Bei experimentellen Versuchsanordnungen kann man dafür Sorge tragen, dass alle Bedingungskombinationen (A_i, B_j) auf eine gleiche Anzahl m von Untersuchungseinheiten angewendet werden. In diesem Fall sind alle Effekte orthogonal: Die Effekte von A und B sind unabhängig, der Interaktionseffekt ist unabhängig von den beiden Haupteffekten.

Die Streuungszerlegung soll mittels folgender Notation veranschaulicht werden, wobei A und B die beiden Faktoren (unabhängige Variablen) mit den Ausprägungen k und *l* sind. Es ergeben sich k x *l* Kombinationen.

k = Zahl der Ausprägungen von A
l = Zahl der Ausprägungen von B

		B			
		1	... j ...		*l*
A	1	y_{11u}	y_{1ju}		y_{1lu}
	⋮				
	i	y_{i1u}	y_{iju}		y_{ilu}
	⋮				
	k	y_{k1u}	y_{kju}		y_{klu}

y_{iju} ist verteilt nach N (μ_{ij}, σ^2). $\mu := E(\bar{y})$

Mittelwert bezüglich A_i: $\bar{y}_{i..} := \frac{1}{l \cdot m} \sum_{j=1}^{l} \sum_{u=1}^{m} y_{iju}$

Mittelwert bezüglich B_j: $\bar{y}_{.j.} := \frac{1}{k \cdot m} \sum_{i=1}^{k} \sum_{u=1}^{m} y_{iju}$

Zellenmittelwert: $\bar{y}_{ij} := \frac{1}{m} \sum_{u=1}^{m} y_{iju}$

Gesamtmittelwert: $\bar{y} := \frac{1}{k \cdot l \cdot m} \sum_{i=1}^{k} \sum_{j=1}^{l} \sum_{u=1}^{m} y_{iju}$

Die Gesamtstreuung des abhängigen Merkmals SS_y ergibt sich aus der Summe der durch die Faktoren A sowie B hervorgerufenen Streuung, aus ihren Wechselwirkungen (Interaktionen) erzeugten Streuung sowie der nicht erklärten Streuung (SS_{within}):

$SS_y = SS_A + SS_B + SS_{AB} + SS_{Fehler}$

$$\sum_{i=1}^{k} \sum_{j=1}^{l} \sum_{u=1}^{m} \left(y_{iju} - \bar{y}\right)^2 = l \cdot m \sum_{i=1}^{k} \left(\bar{y}_{i..} - \bar{y}\right)^2 + k \cdot m \sum_{j=1}^{l} \left(\bar{y}_{.j.} + \bar{y}\right)^2 + \sum_{i=1}^{k} \sum_{j=1}^{l} \left(\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}\right)^2 + \sum_{i=1}^{k} \sum_{j=1}^{l} \sum_{u=1}^{m} \left(y_{iju} - \bar{y}_{ij.}\right)^2$$

(Die nicht quadratischen Terme der Zerlegung sind alle gleich 0. Dies beweist die Orthogonalität der Erklärungsfaktoren[10].)

Die Summanden des Interaktionsterms lassen sich auch schreiben als: $\bar{y}_{ij.} - \left(\bar{y}_{i..} - \bar{y}\right) - \left(\bar{y}_{.j.} - \bar{y}\right) - \bar{y}$
In dieser Form wird deutlicher, dass der Gesamteffekt der Kombinationen von A und B um die Einzeleffekte von A und B bereinigt wird.

Die inferenzstatistischen Berechnungen folgen den nachstehenden Überlegungen.

1) df (SS_y) = n - 1, da ein Freiheitsgrad durch Kenntnis des Gesamtmittelwerts verloren geht.

2) Entsprechend: df (SS_A) = k - 1. Die Schätzung $\frac{SS_A}{k-1}$ hat den Erwartungswert:

   $$\sigma^2 + \frac{lm}{k-1} \sum_{i=1}^{k} \left(\mu_{i.} - \mu\right)^2$$

   Dies ist <u>unter Annahme von H_0</u> eine erwartungstreue Schätzung für σ^2.

---

[10] Wegen der Orthogonalität lässt sich der Gesamteffekt von A und B ($SS_{A, B, AB}$) berechnen als: $SS_{A, B, AB} = SS_A + SS_B + SS_{AB}$. Dies gilt nicht allgemein.

Bezeichnet $\eta_A^2 = \frac{SS_A}{SS_y}$ den Anteil der durch A erklärten Varianz von y, η_B^2 entsprechend, so gilt: Multiple R^2 =

Multiple $\eta^2 = \frac{SS_y - SS_{error}}{SS_y} = \eta_A^2 + \eta_B^2 + \frac{SS_{I(AB)}}{SS_y}$

Nur wenn die Interaktion gar nichts erklärt, setzt sich die Gesamterklärung additiv aus den Erklärungen der einzelnen Faktoren zusammen.

3) Analog zu SS_A: df $(SS_B) = l - 1$.

$$E\left(\frac{SS_B}{l-1}\right) = \sigma^2 + \frac{km}{l-1}\sum_{j=1}^{l}\left(\mu_{.j} - \mu\right)^2$$

Unter Annahme von H_0 ist dies ebenfalls eine erwartungstreue Schätzung für σ^2.

4) df $(SS_{AB}) = kl - (k - 1) - (l - 1) - 1 = (k - 1)(l - 1)$

$$E\left(\frac{SS_{AB}}{(k-l)(l-1)}\right) = \sigma^2 + \frac{m}{(k-1)(l-1)}\sum_{i=1}^{k}\sum_{j=1}^{l}\left(\mu_{ij} - \mu_{i.} - \mu_{.j} + \mu\right)^2$$

Unter Annahme von H_0 ist dies ebenfalls eine erwartungstreue Schätzung für σ^2.

5) df $(SS_{within}) = n - k \cdot l$

$$E\left(\frac{SS_{within}}{n - k \cdot l}\right) = \sigma^2$$

Dies ist immer eine erwartungstreue Schätzung für σ^2, unabhängig von H_0.

6) df (SS_A) + df (SS_B) + df $(SS_{I\,(AB)})$ + df (SS_{within})

= $(k - 1) + (l - 1) + (k - 1)(l - 1) + n - kl$

= $(k - 1)\ l + (l - 1) + n - kl = kl - 1 + n - kl = n - 1$

= df (SS_y)

Zu 2): H_0: $\mu_{1.} = ... = \mu_{k.}$

$$\frac{\frac{SS_A}{(k-1)}}{\frac{SS_{error}}{(n-kl)}}$$ ist $F_{k-1,\,n-kl}$- verteilt.

Falls der beobachtete F-Wert größer als der kritische Wert aus der F-Tabelle für gegebenes Signifikanzniveau ist, so wird die Nullhypothese zurückgewiesen, d.h. dass die Gruppeneinteilung gemäß Faktor A einen signifikanten Erklärungsbeitrag leistet.

Zu 3): H_0: $\mu_{.1} = ... = \mu_{.l}$

$$\frac{\frac{SS_B}{(l-1)}}{\frac{SS_{error}}{(n-kl)}}$$ ist $F_{l-1,\,n-kl}$-verteilt.

F-Test entsprechend zu (2).

Zu 4): H_0: $\mu_{ij} - \mu = (\mu_{i.} - \mu) + (\mu_{.j} - \mu)$
(für i = 1, ..., k; j = 1, ..., *l*)

Falls H_0 zutrifft, so wirken A und B nur einzeln, und zwar additiv, nicht aber in Interaktion miteinander.

$$\frac{\frac{SS_{(AB)}}{[(k-1)(l-1)]}}{\frac{SS_{error}}{(n-kl)}}$$ ist F (k - 1) (*l* - 1), n - k*l*-verteilt.

Wird H_0 zurückgewiesen, so heißt das, dass es einen signifikanten Interaktionseffekt der Faktoren A und B auf die abhängige Variable y gibt.

L

Tabelle 4-3: Zweifache Varianzanalyse mit gleicher Zellenhäufigkeit

Herkunft der Variation	Quadratsumme (SS)	DF	Mittlere Quadratsumme (MS, Mean Square)	F
Haupteffekte (Main Effects)	$(SS_A + SS_B)$			
Faktor A	SS_A	k - 1	$\frac{SS_A}{(k-1)}$	$\frac{MS_A}{\frac{SS_{error}}{(n-kl)}}$
Faktor B	SS_B	*l* - 1	$\frac{SS_B}{(l-1)}$	$\frac{MS_B}{\frac{SS_{error}}{(n-kl)}}$
Interaktion A B	SS_{AB}	(k - 1) (*l* - 1)	$\frac{SS_{AB}}{(k-1)(l-1)}$	$\frac{MS_{AB}}{\frac{SS_{error}}{(n-kl)}}$
Nicht erklärter Rest (Residuum)	SS_{error}	n - k*l*	$\frac{SS_{error}}{n-kl}$	
Total	SS_y	$\sum$ = n - 1		

4.2.2 Ungleiche Zellenhäufigkeiten

In den meisten Fällen, v.a. wenn es sich nicht um ein Experiment handelt, dürften die Zellenhäufigkeiten ungleich sein. In diesem Fall ist die Analyse erheblich komplizierter, da sowohl die Haupteffekte nicht unabhängig voneinander als auch der Interaktionseffekt nicht unabhängig von den Haupteffekten sein brauchen. Da die Orthogonalität nicht vorausgesetzt werden kann, wird

sie künstlich erzeugt, und zwar bei dem klassischen experimentellen Ansatz durch folgende Hierarchie:

1) Additive Effekte von A und B: $(\mu_{i.} - \mu) + (\mu_{.j} - \mu)$
   Bezeichnung der Streuung: $SS_{A, B}$

2) Interaktionseffekt: = Gemeinsamer Effekt von A und B minus der additiven Effekte
   $\alpha\beta_{ij} = (\mu_{ij} - \mu) - [(\mu_{i.} - \mu) + (\mu_{.j} - \mu)]$
   $= \mu_{ij} - \mu_{i.} - \mu_{.j} + \mu$

   Bezeichnungen der Streuungen:
   Aufgrund des gemeinsamen Effekts von A und B: $SS_{A, B, AB}$; aufgrund des Interaktionseffekts von A und B: SS_{AB}; also: $SS_{AB} = SS_{A, B, AB} - SS_{A, B}$

3) Nicht erklärter Rest (Residuum): $\varepsilon_{ij} = \mu_{ij} - \mu$
   Bezeichnung der Streuung: $SS_{error} = SS_y - SS_{A, B, AB}$

Die Streuungszerlegung lautet also:

$$SS_y = SS_{A, B} + \underbrace{(SS_{A, B, AB} - SS_{A, B})}_{SS_{AB}} + \underbrace{(SS_y - SS_{A, B, AB})}_{SS_{error}}$$

Die Erklärungskomponenten der Haupteffekte werden künstlich definiert als:
SS_A, adjustiert um B = $SS_{A, B} - SS_B$
SS_B, adjustiert um A = $SS_{A, B} - SS_A$

Im allgemeinen gilt nicht, dass die Erklärungskomponenten der beiden Haupteffekte sich zur Erklärungskomponente der additiven Effekte $SS_{A, B}$ aufaddieren lassen.

Präziser lassen sich die Zusammenhänge mit folgender Notation beschreiben:

		B: 1	B: j	B: *l*	
A	1	n_{11}	n_{ij}	n_{1l}	$n_{1.}$
	i	n_{i1}	n_{ij}	n_{il}	$n_{i.}$
	k	n_{kl}	n_{kj}	n_{kl}	$n_{k.}$
		$n_{.1}$	$n_{.j}$	$n_{.l}$	$\sum_{i=1}^{k} n_{i.} = \sum_{j=1}^{l} n_{.j} = n$

$$(n_{i.} = \sum_{j=1}^{l} n_{ij}, n_{.j} = \sum_{i=1}^{k} n_{ij}, \bar{y}_{i..} = \frac{1}{n_{i.}} \sum_{j=1}^{l} \sum_{u=1}^{n_{ij}} y_{iju} \text{ , etc.})$$

$$SS_y = \sum_{i=1}^{k} \sum_{j=1}^{l} \sum_{u=1}^{n_{ij}} \left(y_{iju} - \bar{y}\right)^2$$

$$SS_{error} = SS_{within} = \sum \sum \sum \left(y_{iju} - \bar{y}_{ij.}\right)^2$$

Streuungszerlegung:

$$SS_y = \underbrace{\sum\sum\sum\left(y_{iju} - \bar{y}_{ij.}\right)^2}_{SS_{error}} + \underbrace{\sum\sum\sum\left(\bar{y}_{ij.} - \bar{y}\right)}_{SS_{A,\,B,\,AB}}$$

Die Definition von SS_A, SS_B und $SS_{A,\,B}$ hängt von dem gewählten Ansatz ab.
df $(SS_{A,\,B,\,AB})$ = k*l* - 1, df $(SS_{A,\,B})$ = (k - 1) + (*l* - 1),
also: df (SS_{AB}) = k*l* - 1 - (k - 1) - (*l* - 1) = (k - 1) (*l* - 1).
df (SS_{error}) = n – k*l*

Insgesamt:
df (SS_y) = n - 1 = (k - 1) + (*l* - 1) + (k - 1) (*l* - 1) + n - k*l* = df $(SS_{A,\,B})$ + df (SS_{AB}) + df (SS_{error})

Für jede Komponente lautet der F-Test:

$$F = \frac{\frac{SS\,(Komponente)}{df\,(Komponente)}}{\frac{SS_{error}}{df\,(error)}}$$

Als Alternative zum klassischen Ansatz gibt es ferner den hierarchischen Ansatz: Hier wird z.B. in einer dreifachen Varianzanalyse vorausgesetzt , dass zunächst der Einfluss von A berücksichtigt werden soll, *dann* die zusätzliche Erklärung durch B, *dann* die zusätzliche Erklärung durch C.

Tabelle 4-4: Klassischer (experimenteller) und hierarchischer Ansatz bei zweifacher Varianzanalyse

	Klassisch		Hierarchisch	
Faktoren	SS	df	SS	df
Faktor A			SS_A	k - 1
Faktor B			$SS_{A,\,B}$ - SS_A	*l* - 1
Additive Effekte von A und B	$SS_{A,\,B}$	(k - 1) + (*l* - 1)		
Interaktionen AB	$SS_{A,\,B,\,AB}$ - $SS_{A,\,B}$	(k - 1) · (*l* - 1)	$SS_{A,\,B,\,AB}$ - $SS_{A,\,B}$	(k - 1) (*l* - 1)
Fehler	SS_y - $SS_{A,\,B,\,AB}$	n - k*l*	SS_y - $SS_{A,\,B,\,AB}$	n - k*l*
3	SS_y	n - 1	SS_y	n - 1
A B	Anschließend: $SS_{A,\,B}$ - SS_B $SS_{A,\,B}$ - SS_A (Summe ergibt nur dann $SS_{A,\,B}$, falls die Faktoren orthogonal sind. Dies gilt genau dann, falls die Zellenhäufigkeiten proportional sind zu den Randverteilungen.)	 k - 1 *l* – 1		

Klassischer und hierarchischer Ansatz unterscheiden sich nur in der Berechnung der Haupteffekte.

Eine dritte Möglichkeit bietet der Regressionsansatz (vgl. zur Regression Kap. 3.1), bei dem jeweils alle übrigen Effekte kontrolliert werden (auch „Unique – Methode“ bzw. „Eindeutige Methode“ genannt).

Tabelle 4-5: Regressionsansatz bei zweifacher Varianzanalyse

Quelle der Variation	Quadratsumme (SS)
A und B, additiv	$SS_{A, B, AB} - SS_{AB}$
A	$SS_{A, B, AB} - SS_{B, AB}$
B	$SS_{A, B, AB} - SS_{A, AB}$
Interaktion A B	$SS_{A, B, AB} - SS_{A, B}$

Es bestehen folgende Zusammenhänge zwischen dem klassischen, dem hierarchischen und dem Regressionsansatz:

- Bei gleichen Zellenhäufigkeiten führen alle drei Ansätze zum gleichen Ergebnis.

- Falls die Haupteffekte orthogonal sind (dies ist der Fall, wenn die Zellenhäufigkeiten proportional sind zu den Randverteilungen, d.h. statistische Unabhängigkeit der Faktoren), so führen die ersten beiden Ansätze zum gleichen Ergebnis. Bei der zweifachen Varianzanalyse gilt dann: $SS_{A, B} = SS_A + SS_B$

4.3 Dreifache Varianzanalyse

1) Im Regressionsansatz werden jeweils alle übrigen Faktoren kontrolliert.
2) Die Streuungszerlegung für den klassischen und den hierarchischen Ansatz lautet:

Faktoren	SS
A, B, C (Haupteffekte)	$SS_{A, B, C}$
2-fache Interaktionen	$SS_{A, B, C, AB, AC, BC} - SS_{A, B, C}$
3-fache Interaktionen	$SS_{A, B, C, AB, AC, BC, ABC} - SS_{A, B, C, AB, AC, BC}$
Fehler	$SS_y - SS_{A, B, C, AB, AC, BC, ABC}$
$\sum$	SS_y

Additive „Effekte“ bei dreifacher Varianzanalyse

Haupteffekte	Klassisch	Hierarchisch
A	$SS_{A, B, C} - SS_{B, C}$	SS_A
B	$SS_{A, B, C} - SS_{A, C}$	$SS_{A, B} - SS_A$
C	$SS_{A, B, C} - SS_{A, B}$	$SS_{A, B, C} - SS_{A, B}$

Klassischer und hierarchischer Ansatz unterscheiden sich nur in der Berechnung der Haupteffekte.

Welche Effekte werden kontrolliert?

Kontrolle von Effekten in der zweifachen Varianzanalyse:

Faktoren	Klassisch	Hierarchisch	Regression
A	B	keiner	B und AB
B	A	A	A und AB
Interaktion AB	A und B	A und B	A und B

In der dreifachen Varianzanalyse:

Faktoren	Klassisch	Hierarchisch	Regression
A	BC	keiner	Alle anderen
B	AC	A	Alle anderen
C	AB	A, B	Alle anderen
AB	A, B, C, AC, BC	A, B, C, AC, BC	Alle anderen
AC	A, B, C, AB, BC	A, B, C, AB, BC	Alle anderen
BC	A, B, C, AB, AC	A, B, C, AB, AC	Alle anderen
ABC	A, B, C, AB, AC, BC	A, B, C, AB, AC, BC	Alle anderen

Die ersten beiden Ansätze unterscheiden sich also nur in der Berechnung der Haupteffekte.

Präzisere Formulierung der dreifachen Varianzanalyse mit gleicher Zellenhäufigkeit

Faktor	SS	df
A	SS_A	k - 1
B	SS_B	l - 1
C	SS_C	m - 1
AB-Interaktion	SS_{AB}	(k - 1) (l - 1)
AC-Interaktion	SS_{AC}	(k - 1) (m - 1)
BC-Interaktion	SS_{BC}	(l - 1) (m - 1)
ABC-Interaktion	SS_{ABC}	(k - 1) (l - 1) (m - 1)
Fehler	$SS_e = \sum_i \sum_j \sum_u \sum_v \left(y_{ijuv} - \bar{y}_{iju.}\right)^2$	n - k · l · m
∑	SS_y	n - 1

Wegen der gleichen Zellenhäufigkeiten sind alle 8 Faktoren orthogonal, so dass sich die Effekte einfach aufaddieren lassen. Der Fall gleicher Zellenhäufigkeiten ist einfach und lässt sich leicht entsprechend auf eine beliebige Zahl von unabhängigen Variablen verallgemeinern.

4.4 Die einfache Varianzanalyse als Spezialfall der multiplen Regression

Durch Verwendung dichotomer Dummy-Variablen (Stellvertretervariablen für die Ausprägungen) für eine nominale unabhängige Variable lässt sich die einfache Varianzanalyse als Spezialfall der multiplen Regression darstellen.

Die Information über eine nominale Variable A mit k Ausprägungen lässt sich vollständig beschreiben durch k - 1 dichotome Variablen $1_{A_1},...,1_{A_{k-1}}$, die k-te Ausprägung wird dadurch vercodet, dass sie in allen dichotomen Variablen gleich Null gesetzt wird.

		Dichotome Variablen		
		1_{A_1}	1_{A_2}	$1_{A_{k-1}}$
A	1	1	0…	0
	2	0		⋮
	⋮	⋮		
			1…	⋮
			⋮	0
	⋮	⋮	⋮	1
	k	0	0…	0

Die Anzahl der Einheiten mit Ausprägungen i sei n_i, ferner: $\sum_{i=1}^{k} n_i = n$

Für Untersuchungseinheiten (i, j) (z.B. Personen) gilt (j = 1, ..., n_i; *l*, i = 1, ..., k):

$$1_{A_l}(i,j) = \begin{cases} 1, \textit{falls Person}\,(i,j)\,\textit{bei Merkmal A die Ausprägung l hat, d.h. falls } i = l; \\ 0\ \textit{sonst}. \end{cases}$$

Für die Regressionsschätzung benötigt man nur k - 1 dieser Dichotomien, da sich die k-te daraus berechnen lässt: Aus $\sum_{l=1}^{k} 1_{A_l} = 1$ folgt: $1_{A_k} = 1 - \sum_{l=1}^{k-1} 1_{A_l}$.

In der Regressionsschätzung $= b_0 + b_1\ 1_{A_1} + ... b_{k-1}\ 1_{A_{k-1}}$ erhält man für alle Einheiten (für j = 1, ..., n_i und i = 1, ..., k - 1) mit Ausprägung i den Wert:

$$\hat{y}(i,j) = b_0 + b_1 \underbrace{1_{A_1}(ij)}_{0} + \ldots + b_i \underbrace{1_{A_i}(i,j)}_{1} + \ldots b_{k-1} \underbrace{1_{A_{k-1}}(i,j)}_{0} = b_0 + b_i$$

Für Ausprägung k: (k, j) = b_0 (für j = 1, ..., n_k)
Die Schätzwerte für alle Untersuchungseinheiten mit gleicher Ausprägung auf A sind also gleich.

Die Regressionsschätzung ist dadurch charakterisiert, dass der Fehler $\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\hat{y}_{ij} - y_{ij})^2$

minimiert wird; ferner gilt:

$$\sum\sum(\hat{y}_{ij} - y_{ij})^2 = \sum\sum(\hat{y}_{ij} - \bar{y}_{ij})^2 + \sum\sum(\bar{y}_i - y_{ij})^2$$

Begründung: Wegen $\hat{y}_{ij} = \hat{y}_i$ ergibt sich:

$$2\sum_i \sum_j \left(\hat{y}_{ij} - \bar{y}_i\right)\left(\bar{y}_i - y_{ij}\right) = 2\sum_i \left(\hat{y}_i - \bar{y}_i\right)\underbrace{\sum_j \left(\bar{y}_i - y_{ij}\right)}_{0} = 0$$

$\sum\sum\left(\hat{y}_{ij} - y_{ij}\right)^2$ wird minimal, falls $\hat{y}_{ij} = \bar{y}_i$.

Es gilt also der Zusammenhang:

$$\left.\begin{array}{r} b_0 = \bar{y}_k \\ i \neq k : b_0 + b_i = \bar{y}_i \end{array}\right\} \; b_i = \bar{y}_i - \bar{y}_k \quad \text{(für i = 1, ..., k - 1)}$$

Die Konstante b_0 entspricht der ausnahmslos mit Nullen codierten Merkmalsausprägung k. Die übrigen b_i-Koeffizienten repräsentieren jeweils den Unterschied zwischen der i-ten und der k-ten Gruppe.

Aus $\sum_{i=1}^{k} 1_{A_i}$ folgt, dass die Regressionsschätzung sich auf folgende zwei Arten schreiben lässt:

$$\hat{y} = b_0 + \sum_{i=1}^{k-1} b_i 1_{A_i} = b_0 \sum_{i=1}^{k} 1_{A_i} + \sum_{i=1}^{k-1} b_i 1_{A_i} = \sum_{i=1}^{k-1} \left(b_0 + b_i\right) 1_{A_i} + b_0 1_{A_k} = \sum_{i=1}^{k} \bar{y}_i 1_{A_i}$$

Streuungszerlegung:

$$SS_y = SS_{res} + SS_{reg} = \sum_i \sum_j \left(y_{ij} - \hat{y}_{ij}\right)^2 + \sum_i \sum_j \left(\hat{y}_{ij} - \bar{y}\right)^2 = \underbrace{\sum_i \sum_j \left(y_{ij} - \bar{y}_i\right)^2}_{SS_{within}\ SS_{error}} + \underbrace{\sum_i n_i \left(\bar{y}_i - \bar{y}\right)^2}_{SS_{between}}$$

„Durch die Regression erklärt“ = „durch die Kategoriebildung erklärt“

Also: $Multiple\ R^2 = \dfrac{SS_y - SS_{error}}{SS_y} = \dfrac{SS_A}{SS_y} = \eta_A^2$ (Eta – Quadrat)

4.5 Die zweifache Varianzanalyse als Spezialfall der multiplen Regression mit Interaktionstermen

Im Weiteren wird die zweifache Varianzanalyse als Spezialfall der multiplen Regression mit Interaktionstermen betrachtet.

y bezeichnet eine abhängige metrische Variable, z.B. die Affinität zur CDU/CSU.

In SPSS lassen sich dichotome Dummy-Variablen (stellvertretend z.B. für die Ausprägungen der nominalen Variablen „Religion“) durch IF-Anweisung einfach erzeugen, wobei benutzt wird, dass alle nicht anders festgelegten Werte einer neuen Variablen gleich 0 gesetzt werden. Hat die nominale Kategorie „Religion“ 4 Ausprägungen, so reicht als Definition:

```
IF      (RELIGION EQ 1)     REL1 = 1
IF      (RELIGION EQ 2)     REL2 = 1
IF      (RELIGION EQ 3)     REL3 = 1
```

Neben der vierstufigen Religionsvariable wird eine zweite nominale Variable (z.B. Geschlecht) berücksichtigt:

```
IF      (SEX EQ 1)      SEX1 = 1
```

Die Regressionsschätzung von y auf Religion ist:

$$\hat{y} = b_0 + b_1\ REL1 + b_2\ REL2 + b_3\ REL3$$

Die Schätzwerte als Mittelwerte der Variablen y für die Einheiten mit einer bestimmten Religion sind dann:

Protestantisch	$b_0 + b_1$	Jüdisch	$b_0 + b_3$
Katholisch	$b_0 + b_2$	Ohne	b_0

Die Regressionsschätzung von y auf Geschlecht ist:

$$\hat{y} = \tilde{b}_0 + \tilde{b}_1\ SEX1$$

Die Schätzwerte als Mittelwerte der Einheiten mit einer der beiden Ausprägungen sind dann:

Männlich	$\tilde{b}_0 + \tilde{b}_1$
Weiblich	$\tilde{b}_0$

Betrachtet man nun eine dritte Regression, nämlich die Regression von y auf beide Variablengruppen, jedoch ohne Interaktionen, so geben die Koeffizienten den um die andere Variable bereinigten Einfluss wieder:

$$\hat{y} = \bar{b}_0 + \bar{b}_1\ REL1 + \bar{b}_2\ REL2 + \bar{b}_3 + REL3 + \bar{b}_4\ SEX1$$

Die Schätzwerte als Mittelwerte der Einheiten mit einer bestimmten Kombination von Religion und Geschlecht sind dann:

	Männlich	Weiblich
Protestantisch	$\bar{b}_0 + \bar{b}_1 + \bar{b}_4$	$\bar{b}_0 + \bar{b}_1$
Katholisch	$\bar{b}_0 + \bar{b}_2 + \bar{b}_4$	$\bar{b}_0 + \bar{b}_2$
Jüdisch	$\bar{b}_0 + \bar{b}_3 + \bar{b}_4$	$\bar{b}_0 + \bar{b}_3$
Ohne	$\bar{b}_0 + \bar{b}_4$	$\bar{b}_0$

In dem SPSS-Output zur multiplen Klassifikationsanalyse wird nun ausgedruckt:

Religion	Deviations	Adjusted deviations
Katholisch	$\bar{y}_1 - \bar{y} = b_0 + b_1 - \bar{y}$	$\bar{b}_0 + \bar{b}_1 - \bar{y} = \bar{y}_{1,2} - \bar{y}$
Protestantisch	$\bar{y}_2 - \bar{y} = b_0 + b_2 - \bar{y}$	$\bar{b}_0 + \bar{b}_2 - \bar{y} = \bar{y}_{2,2} - \bar{y}$
Jüdisch	$\bar{y}_3 - \bar{y} = b_0 + b_3 - \bar{y}$	$\bar{b}_0 + \bar{b}_3 - \bar{y} = \bar{y}_{3,2} - \bar{y}$
Ohne	$\bar{y}_4 - \bar{y} = b_0 - \bar{y}$	$\bar{b}_0 - \bar{y} = \bar{y}_{4,2} - \bar{y}$
Geschlecht		
Männlich	$\bar{y}_1 - \bar{y} = \tilde{b}_0 + \tilde{b}_1 - \bar{y}$	$\bar{b}_0 + \bar{b}_4 - \bar{y} = \bar{y}_{4,1} - \bar{y}$
Weiblich	$\bar{y}_2 - \bar{y} = \tilde{b}_0 - \bar{y}$	$\bar{b}_0 - \bar{y} = \bar{y}_{4,2} - \bar{y}$

Auf diese Weise lassen sich "bereinigte" Effekte im additiven, Interaktionsterme nicht berücksichtigenden Modell formulieren, ähnlich wie in der multiplen Regression ein bereinigter Prädiktor einen unabhängigen (zusätzlichen) Erklärungsbeitrag leistet.

In der zweifachen Varianzanalyse werden zusätzlich die Interaktionsterme berücksichtigt:
$\hat{y} = b_0 - b_1 REL1 + b_2 REL2 + b_3 REL3 + b_4 SEX1 + b_5 REL1 \cdot SEX1 + b_6 REL2 \cdot SEX1 + b_7 REL3 \cdot SEX1$
Es handelt sich um ein „*saturiertes*“ Modell, da alle Interaktionsterme berücksichtigt sind.

Für die Regressionsschätzung $\hat{y}_{iju}$ gilt wieder, dass sie für jede fest vorgegebene Kombination (i, j) unabhängig von u ist.

Deshalb gilt wieder:

$$\sum\sum\sum\left(\hat{y}_{iju} - y_{iju}\right)^2 = \sum\sum\sum\left(\hat{y}_{iju} - \bar{y}_{ij}\right)^2 + \sum\sum\sum\left(\bar{y}_{ij} - y_{iju}\right)^2 \text{ , denn:}$$

$$\sum\sum\sum\left(\hat{y}_{iju} - \bar{y}_{ij}\right)\left(\bar{y}_{ij} - y_{iju}\right) = \sum_i\sum_j\left(\hat{y}_{ij} - \bar{y}_{ij}\right)\underbrace{\sum_u\left(\bar{y}_{ij} - y_{iju}\right)}_{0}$$

Die Regressionsschätzung für jede fest vorgegebene Kombination (i, j) ist also gleich dem Durchschnitt der y-Werte über alle Einheiten mit Ausprägung (i, j), d.h. gleich dem Zellenmittelwert.

Schätzwerte (Mittelwerte der Einheiten mit einer bestimmten Kombination von Religion und Geschlecht):

	Geschlecht	
Religion	Männlich	Weiblich
Protestantisch	$b_0 + b_1 + b_4 + b_5$	$b_0 + b_1$
Katholisch	$b_0 + b_2 + b_4 + b_6$	$b_0 + b_2$
Jüdisch	$b_0 + b_3 + b_4 + b_7$	$b_0 + b_3$
Ohne	$b_0 + b_4$	b_0

$$R^2_{A,B,AB} = \frac{SS_y - SS_{error}}{SS_y}$$

$$SS_{error} = \sum_i \sum_j \sum_u \left(y_{iju} - \bar{y}_{ij}\right)^2, \text{ wobei} : \bar{y}_{ij} = \hat{y}_{ij}$$

4.5.1 Beispiel für die zweifache Varianzanalyse mit ungleichen Zellenhäufigkeiten: Untersuchung der Lebenszufriedenheit

Abhängige Variable: Lebenszufriedenheit

Primärbeziehung	Beruf	Mittelwert	N
klein	klein	,1875	160
	groß	,8571	140
	Gesamt	,5000	300
groß	klein	,4828	290
	groß	,7222	360
	Gesamt	,6154	650
Gesamt	klein	,3778	450
	groß	,7600	500
	Gesamt	,5789	950

Datenquelle: Mayntz et al. 1978

Die tabellenanalytische Interpretation des Beispiels befindet sich auf S. 56 f.

Ansatz der Varianzanalyse: Test der additiven und Interaktionseffekte

Beispiel für eine zweifache Varianzanalyse

Klassischer Ansatz (Experimentelle Methode)

ANOVA: Lebenszufriedenheit nach Primärbeziehung, Beruf

			Experimentelle Methode				
			Quadrat-summe	df	Mittel der Quadrate	F	Sig.
Lebens-zufrie-denheit	Haupteffekte	(Kombiniert)	35,993	2	17,996	91,454	,000
		Primärbeziehung	1,391	1	1,391	7,071	,008
		Beruf	33,260	1	33,260	169,020	,000
	2-Weg-Wechsel-wirkungen	Primärbeziehung * Beruf	9,432	1	9,432	47,934	,000
	Modell		45,425	3	15,142	76,947	,000
	Residuen		186,154	946	,197		
	Insgesamt		231,579	949	,244		

1) Zunächst wird getestet, ob die Haupteffekte insgesamt (additiven Effekte insgesamt) signifikant sind. $SS_{A,B} = 35{,}993$; gemäß dem F-Wert und der Significance leisten die additiven Effekte einen signifikanten Erklärungsbeitrag.

2) Die Interaktionseffekte, d.h. die Effekte der Kombinationen von A und B über die additiven Effekte hinaus, ergeben eine Variation von $SS_{AB} = 9{,}432$, die nach dem F-Wert und der Significance signifikant ist.

3) Das gesamte Modell weist eine erklärte Variation von $SS_{A,S,AB} = 45{,}425$ auf, welche nach dem F-Test und der Significance signifikant ist.

4) Die gesamte Variation beträgt $SS_y = 231{,}579$, davon entfällt auf die nicht erklärte Variation ein Betrag von $SS_{Residuen} = 186{,}154$.

5) Da die Erklärungsbeiträge von A und B sich überschneiden können, bzw. auch Suppressor etc. möglich sind, gilt nicht einfach die Additivität, denn i.a.: $SS_{A,B} \neq SS_A + SS_B$

Für dieses Problem gibt es nun unterschiedliche Lösungsvorschläge.

In der klassischen (experimentellen) Methode werden die Faktoren jeweils gegeneinander bereinigt, ihr Erklärungsbeitrag wird definiert als.

$$SS_{A,\,adjusted\ for\ B} = SS_{A,B} - SS_B = 35{,}993 - 34{,}601 = 1{,}391$$

$$SS_{B,\,adjusted\ for\ A} = SS_{A,B} - SS_A = 35{,}993 - 2{,}733 = 33{,}260$$

In der hierarchischen Methode wird der Erklärungsbeitrag des Faktors A vollständig berücksichtigt und der Faktor B dann um A bereinigt:

$$SS_A = 2{,}733$$

$$SS_{B,\,adjusted\ for\ A} = SS_{A,B} - SS_A 35{,}993 - 2{,}733 = 33{,}260$$

Im Regressions-Ansatz (Unique Methode) werden die Faktoren *und* die Interaktionen gegeneinander bereinigt. Dies entspricht der Regression von y auf A, B und (AB).

			Hierarchische Methode				
			Quadrat-summe	df	Mittel der Quadrate	F	Sig.
Lebens-zufrie-denheit	Haupt-effekte	(Kombiniert)	35,993	2	17,996	91,454	,000
		Primärbezie-hung	2,733	1	2,733	13,888	,000
		Beruf	33,260	1	33,260	169,020	,000
	2-Weg-Wechsel-wirkungen	Primärbezie-hung * Beruf	9,432	1	9,432	47,934	,000
	Modell		45,425	3	15,142	76,947	,000
	Residuen		186,154	946	,197		
	Insgesamt		231,579	949	,244		

			Eindeutige Methode				
			Quadrat-summe	df	Mittel der Quadrate	F	Sig.
Lebens-zufrie-denheit	Haupt-effekte	(Kombiniert)	44,906	2	22,453	114,102	,000
		Primärbezieh ung	1,310	1	1,310	6,659	,010
		Beruf	42,126	1	42,126	214,079	,000
	2-Weg-Wechsel-wirkungen	Primärbezieh ung * Beruf	9,432	1	9,432	47,934	,000
	Modell		45,425	3	15,142	76,947	,000
	Residuen		186,154	946	,197		
	Insgesamt		231,579	949	,244		

Gesamtvariation

$$SS_y = 950 \cdot \frac{550}{950} \cdot \frac{400}{950} = 231{,}579$$

Nicht erklärte Variation

$$SS_{\text{Re siduen}} = SS_{within} = \sum_i \sum_j \sum_{u=1}^{n_{ij}} \left(y_{iju} - \bar{y}_{ij.}\right)^2$$

$$= \sum_i \sum_j n_{ij} MS_{(ij)}$$

$MS_{(ij)}$ = Streuung von y in der Kombination (ij)

In dem Beispiel:

$$SS_{\text{Re siduen}} = 160 \cdot \frac{30}{160} \cdot \frac{30}{160} + 290 \cdot \frac{140}{290} \cdot \frac{150}{290} + 140 \cdot \frac{120}{140} \cdot \frac{20}{140} + 360 \cdot \frac{260}{360} \cdot \frac{100}{360} = 186{,}154$$

Erklärte Variation

Die erklärte Variation ist dann die Differenz:
$SS_{Modell} = SS_y - SS_{\text{Re siduen}} = 45{,}425$

Die erklärte Komponente ließe sich auch direkt berechnen:
$SS_{Modell} = \sum\sum\sum \left(\bar{y}_{ij.} - \bar{y}\right)^2$

Additive Komponente der Variation

Die Regression von y auf x,z ergab einen Erklärungsanteil von $\frac{35{,}993}{231{,}579} = 15{,}5\%$

$SS_{A,B} = 35{,}993$

Interaktionskomponente der Variation

Die Interaktionskomponente der Variation ergibt sich als Differenz der erklärten Variation und der additiven Komponente der Variation.

$$SS_{A,B} = SS_{A,B,AB} - SS_{A,B} = 45{,}425 - 35{,}993 = 9{,}432$$

Erklärte Varianz und Effekte

Einfache Varianzanalyse
In der einfachen Varianzanalyse von y durch das Merkmal A lässt sich die Varianz anschaulich darstellen durch das Zusammenwirken der Kovarianz von y und 1_{A_i} mit dem direkten Effekt $\bar{y}_{A_i} - \bar{y}$ von 1_{A_i} auf y.

Erklärte Variation: $\sum_{i=1}^{k} \frac{n_i}{n} \left(\bar{y}_{A_i} - \bar{y}\right)^2$

Regressionsschätzung: $\hat{y} = \sum_{i=1}^{k} \bar{y}_i 1_{A_i} = \bar{y} + \sum_{i=1}^{k} \left(\bar{y}_i - y\right) 1_{A_i}$

In der multiplen Regression ist $y - \hat{y}$ orthogonal zu den Prädiktoren x_i, also auch zu $\hat{y}$.

Deshalb gilt: $s_{\hat{y}}^2 = s_{\hat{y},\hat{y}} = s_{y,\hat{y}}$

$s_{y,\hat{y}} = \sum_{i=1}^{k} s_{y,x_i} \beta_i$

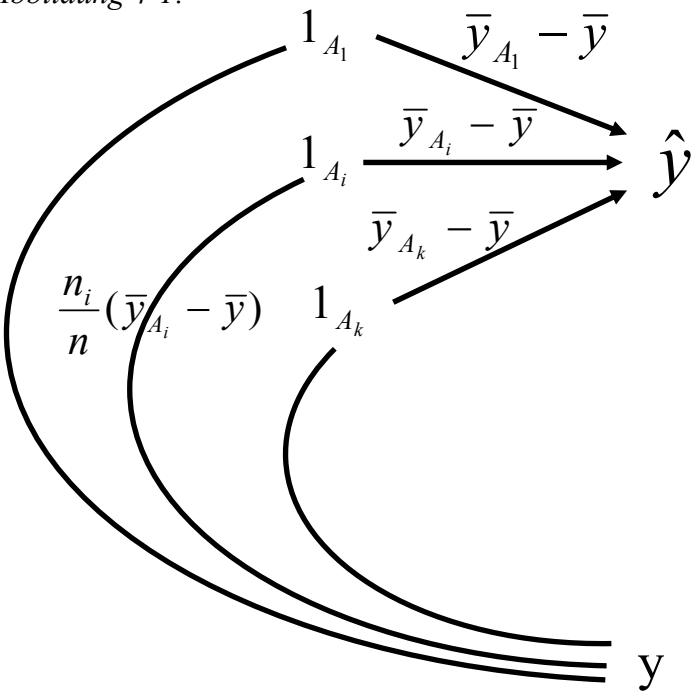
In der Varianzanalyse:

$$s_{y,1_{A_i}} = \frac{n_i}{n}\left(\bar{y}_{A_i} - \bar{y}\right)$$

Also: Multiple $R^2 = \sum_{i=1}^{k} \underbrace{\frac{n_i}{n}\left(\bar{y}_{A_i} - \bar{y}\right)}_{s_{y,1_{A_i}}} \cdot \underbrace{\left(\bar{y}_{A_i} - \bar{y}\right)}_{\text{Effekt von } 1_{A_i}}$

Meine „pfadanalytische“ Veranschaulichung der erklärten Varianz:

Abbildung 4-1:



Die erklärte Varianz ist gleich der Kovarianz von y auf $\hat{y}$. Die Kovarianz von y und $\hat{y}$ ergibt sich auch daraus, dass y mit den Prädiktoren 1_{A_i} kovariiert und dir Prädiktoren 1_{A_i} einen Effekt $\bar{y}_{A_i} - \bar{y}$ haben.

<u>Zweifache Varianzanalyse</u>

Wenn y durch die Merkmale A und B erklärt werden soll, so gibt es zunächst die Erklärungsbeiträge der Haupteffekte von A bzw. B, wobei sich die Erklärungsbeiträge überschneiden können.

Erklärte Variation durch A: $\sum_{i=1}^{k} \frac{n_i}{n}\left(\bar{y}_{A_i} - \bar{y}\right)^2$

Erklärte Variation durch B: $\sum_{j=1}^{l} \frac{n_j}{n}\left(\bar{y}_{B_j} - \bar{y}\right)^2$

In der Varianzanalyse werden $\bar{y}_{A_i} - \bar{y}$ als (Haupt-) Effekte von A und $\bar{y}_{B_j} - \bar{y}$ als (Haupt-) Effekte von B bezeichnet. Die Modell-prognose aufgrund des „additiven Modells“ würde dann für (i, j) lauten: $\bar{y} + \left(\bar{y}_{A_i} - \bar{y}\right) + \left(\bar{y}_{B_j} - \bar{y}\right)$

Achtung: Dies ist nicht identisch mit der Regression von y auf die 1_{A_i} und 1_{B_j}, da die direkten Effekte in der multiplen Regression berechnet werden, indem jeweils alle übrigen Prädiktoren in ihrem Einfluss herauspartialisiert werden. Das „additive Modell“ der Varianzanalyse ist insofern zu einfach, als zunächst ein unabhängiges Wirken von A und B angenommen wird und erst im nächsten Schritt die Abweichung als Interaktion betrachtet wird. Die Regression dagegen rechnet die Beziehung zwischen den Prädiktoren sofort heraus.

Aus der Forderung, dass die Abweichungen der Beobachtungen von den aufgrund des Modells zu erwartenden Beobachtungen (d.h. der Fehler) minimal sein soll, ergibt sich, dass der Effekt der Interaktion $1_{A_i}1_{B_j}$ lautet:

Effekt $1_{A_i}1_{B_j} = \bar{y}_{A_iB_j} - \left(\bar{y} + \left(\bar{y}_{A_i} - \bar{y}\right) + \left(\bar{y}_{B_j} - \bar{y}\right)\right)$

D.h. der Effekt der Interaktion $1_{A_i}1_{B_j}$ ist das Ausmaß, in dem der Durchschnitt der Kombination von dem bei additivem Wirken der Haupteffekte zu erwartenden Wert abweicht.

Die erklärte Varianz soll nun wieder mit Hilfe von Kovarianzen und Effekten dargestellt werden.

Die Kovarianz zwischen y und $1_{A_i}1_{B_j}$ lässt sich berechnen als:

$$s_{y,1_{A_i}1_{B_j}} = \frac{n_{A_iB_j}}{n}\left(\bar{y}_{A_iB_j} - \bar{y}\right)$$

Wegen dieser Ergebnisse können die Effekte von $1_{A_i}1_{B_j}$ und die Kovarianz von y und $1_{A_i}1_{B_j}$ unterschiedliches Vorzeichen haben.

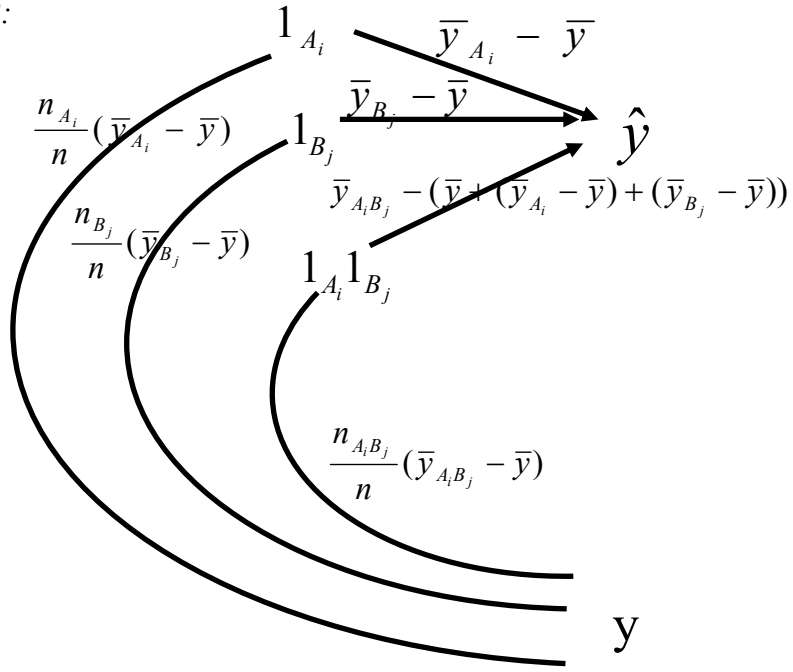
Die erklärte Varianz lautet:

$$s^2_{\hat{y}} = s_{y,\hat{y}} = \sum_{i=1}^{k}\frac{n_{A_i}}{n}\left(\bar{y}_{A_i} - \bar{y}\right)^2 + \sum_{j=1}^{l}\frac{n_{B_j}}{n}\left(\bar{y}_{B_j} - \bar{y}\right)^2$$

$$+ \sum_{i=1}^{k}\sum_{j=1}^{l}\frac{n_{A_iB_j}}{n}\left(\bar{y}_{A_iB_j} - \bar{y}\right)\left(\bar{y}_{A_iB_j} - \left(\bar{y} + \left(\bar{y}_{A_i} - \bar{y}\right) + \left(\bar{y}_{B_j} - \bar{y}\right)\right)\right)$$

Meine „pfadanalytische“ Veranschaulichung der erklärten Varianz für die zweifache Varianzanalyse:

Abbildung 4-2:



Die erklärte Varianz ergibt sich daraus, dass y mit den Prädiktoren 1_{A_i} und 1_{B_j} sowie $1_{A_i}1_{B_j}$ kovariiert, die jeweils den angegebenen Effekt haben.

<u>Berechnung der erklärten Varianz für das Beispiel Lebenszufriedenheit (y) in Abhängigkeit von Beziehungszufriedenheit (x) und Berufszufriedenheit (z)</u>

Varianz der Lebenszufriedenheit:

$$s_y^2 = \frac{n_{y_1}}{n}\left(1 - \frac{n_{y_1}}{n}\right) = 0{,}579 \cdot (1 - 0{,}579) = 0{,}244$$

Durch Faktor x erklärte Variation:

$$0{,}316 \cdot (-0{,}079)^2 + 0{,}684 \cdot (0{,}037)^2 = 0{,}003$$

Dies sind 1,18 % erklärte Varianz.

Durch Faktor z erklärte Variation:

$$0{,}474 \cdot (-0{,}201)^2 + 0{,}526 \cdot (0{,}181)^2 = 0{,}036$$

Dies sind 14,93 % erklärte Varianz.

Durch Interaktion $x_1 z_1$ erklärte Variation:

$$0{,}168 \cdot \underbrace{(0{,}188 - 0{,}579)}_{\text{-0,391}} \cdot (-0{,}111) = 0{,}007$$

Durch Interaktion $x_1 z_2$ erklärte Variation:

$$0{,}147 \cdot \underbrace{(0{,}857 - 0{,}579)}_{0{,}278} \cdot (0{,}176) = 0{,}007$$

Durch Interaktion $z_1 x_2$ erklärte Variation:

$$0{,}305 \cdot \underbrace{(0{,}483 - 0{,}579)}_{\text{-0,096}} \cdot (0{,}069) = -0{,}002$$

Durch Interaktion $x_z z_2$ erklärte Variation:

$$0{,}379 \cdot \underbrace{(0{,}722 - 0{,}579)}_{0{,}143} \cdot (-0{,}075) = -0{,}004$$

Durch Interaktionen insgesamt erklärte Variation: 0,009
Dies sind 3,49 % der erklärten Varianz.

Insgesamt erklärt:

Durch x:	0,003
Durch z:	0,036
Durch Interaktionen:	0,007
	0,007
	-0,002
	-0,004
	$\sum = 0{,}048$

Anteil erklärter Varianz = 0,048/0,244 = 19,60 % erklärte Varianz

Durch x werden 1,18 % erklärt und durch z werden 14,93 % der Varianz erklärt. Rechnerisch ergäbe die Summe 16,11 %, aber dies ist wegen der Überschneidung überzeichnet. Gemäß der Regression erklären x und z (ohne Interaktionen) nur 15,5 % der Varianz. Die Überzeichnung durch $SS_A + SS_B$ wird in der vorliegenden Berechnung gerade durch das Zusammenwirken von Kovariationen und Effekten ausgeglichen.

Die nicht erklärte Varianz ist die Varianz innerhalb der Kombinationen:

$$0{,}164 \cdot 0{,}188 \cdot (1 - 0{,}188)$$

$$+ \quad 0{,}305 \cdot 0{,}483 \cdot (1 - 0{,}483)$$

$$+ \quad 0{,}147 \cdot 0{,}857 \cdot (1 - 0{,}857)$$

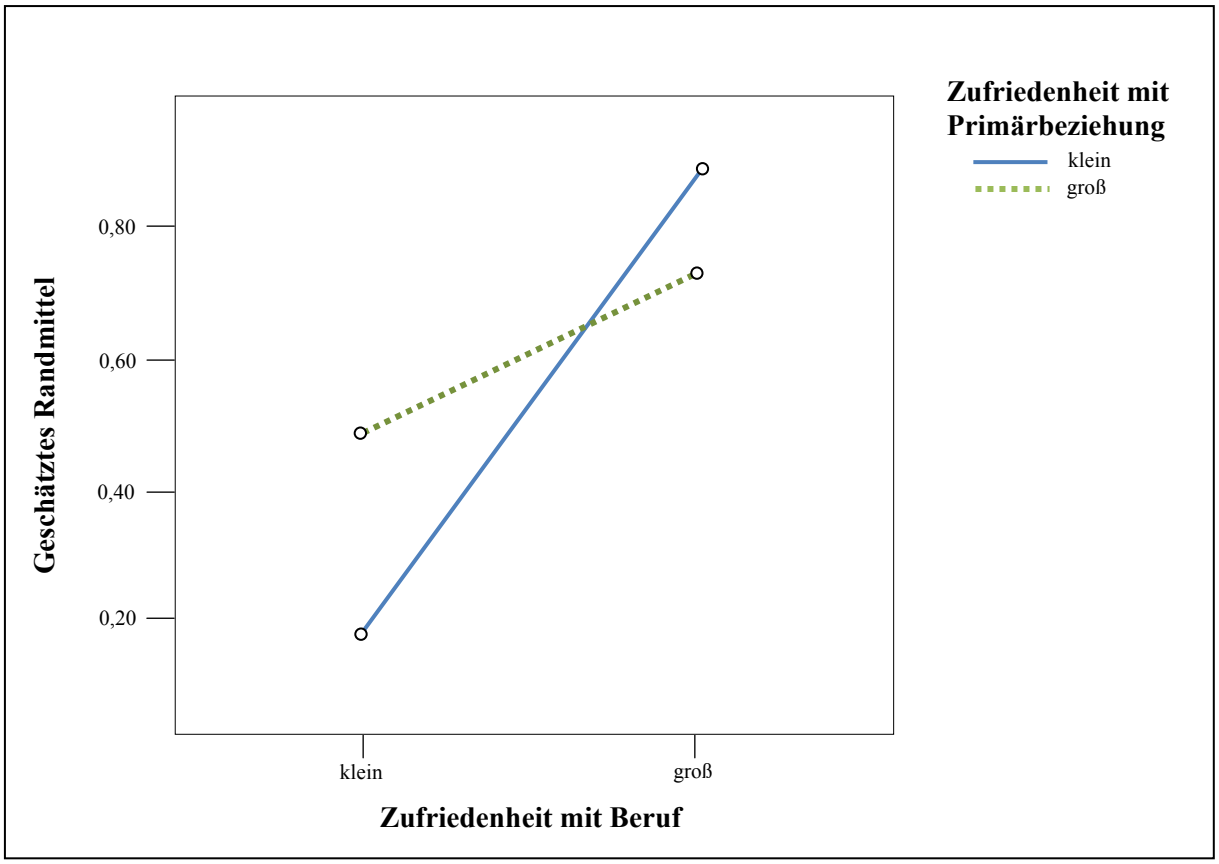
$$+ \quad 0{,}379 \cdot 0{,}722 \cdot (1 - 0{,}722)$$

$$= \quad 0{,}196$$

Die erklärte Varianz (0,048) und die nicht erklärte Varianz (0,196) ergeben zusammen die Gesamtvarianz (0,244).

Die erklärte Varianz für „Nicht-Zufriedenheit“ würde rechnerisch zu den gleichen Ergebnissen führen. (Es handelt sich ja auch um die spiegelbildliche Fragestellung.)

Abbildung 4-3: Geschätztes Randmittel von Lebenszufriedenheit



Bei großer Zufriedenheit mit den Primärbeziehungen unterscheiden sich Personen mit großer Berufszufriedenheit von solchen mit geringer Berufszufriedenheit um 72,22 – 48,28 = 23,94 %. Verglichen damit unterscheiden sich bei geringer Zufriedenheit mit den Primärbeziehungen Personen mit großer Berufszufriedenheit mit 85,71 – 18,75 = 66,96 % deutlich überproportional von solchen mit geringer Berufszufriedenheit. Die ist eine Charakterisierung der Interaktionseffekte, die sich graphisch veranschaulichen lässt. In dem Beispiel bedeutet dies, dass die sich Berufszufriedenheit besonders wichtig ist für die allgemeine Lebenszufriedenheit.

Stellenwert der Multiple Classification Analysis (MCA)

Der Kern der Varianzanalyse besteht in den Tests, ob die additiven Effekte (im Sinne der Regression von y auf die Variablen A und B) und die Interaktionseffekte (Kombinationseffekte über die additiven Effekte hinaus) einen signifikanten Erklärungsbeitrag leisten.

Zur Deskription der Effekte ist in SPSS die MCA (Multiple Classification Analysis) verfügbar. In dem Beispiel erklärt die MCA 15,5% der Varianz, d.h. es handelt sich um eine Deskription mit Hilfe der additiven Effekte im Sinne der multiplen Regression:

Effekt von A: $\beta_{A.B}(=0{,}078)$

Effekt von B: $\beta_{B.A}(=0{,}380)$

Falls es starke Interaktionseffekte gibt, macht eine Beschränkung auf die additiven Effekte wie in der MCA nicht viel Sinn. Das Design der Regressionsanalyse ist an dieser Stelle flexibler. Der Anwender erhält auch den Beta- Koeffizient der Interaktion, falls er eine Regression von y auf x, z, xz durchführt:

$$\beta_{y,x} = 0{,}278$$
$$\beta_{y,z} = 0{,}677$$
$$\beta_{y,xz} = -0{,}423$$

Die allgemeine Lebenszufriedenheit steigt also mit der Beziehungszufriedenheit (x) und noch stärker mit der Berufszufriedenheit (z), während die Interaktion einen negativen Effekt hat.

4.6 Unterschiedliche Codierungen in der Varianzanalyse

4.6.1 Codierung durch Dichotomien in der einfachen Varianzanalyse

Wird eine nominale Variable A mit k Ausprägungen durch Dichotomien vercodet, wobei jeweils der Wert α für das Zutreffen und der Wert β für das Nicht-Zutreffen einer Ausprägung steht, so erhält man (für i = 1, ..., k - 1):

$$\bar{y}_i = b_0 + \alpha b_i + \beta \sum_{\substack{j=i \\ j \neq i}}^{k-1} b_j$$

$$\bar{y}_k = b_0 + \beta \sum_{j=1}^{k-1} b_j$$

Also $\left(\text{für } i = 1,...,k-1\right): \bar{y}_i - \bar{y}_k = (\alpha - \beta) b_i$

Für die 1-0-Vercodung erhält man also:

$\left(\bar{y}_i = b_0 + b_i\right)$ $\quad b_i = \bar{y}_i - \bar{y}_k$ $\quad$ (i = 1, ..., k - 1)

$\left(\bar{y}_k = b_0\right)$ $\quad b_0 = \bar{y}_k$

Für die (+ 1)-(- 1)-Vercodung erhält man:

$$\left(\bar{y}_i = b_0 b_i - \sum_{\substack{j=1 \\ j \neq i}}^{k-1} b_j\right) \qquad b_i = \frac{\left(\bar{y}_i - \bar{y}_k\right)}{2} \qquad \text{(i = 1, ..., k - 1)}$$

$$\left(\bar{y}_k = b_0 - \sum_{j=1}^{k-1} b_j\right) \qquad b_0 = \bar{y}_k + \sum_{j=1}^{k-1} \frac{\left(\bar{y}_j - \bar{y}_k\right)}{2}$$

Ist die nominale Variable A eine Dichotomie, so erhält man als Regressionskoeffizienten in der (+ 1)-(- 1)-Vercodung gerade die "Effekte", d.h. die Abweichung vom nicht gewogenen Gesamtmittelwert:

$\left(\bar{y}_1 = b_0 + b_1\right)$ $\qquad b_0 = \frac{\bar{y}_1 + \bar{y}_2}{2} =: \bar{\bar{y}}$

$\left(\bar{y}_2 = b_0 - b_1\right)$ $\qquad b_1 = \bar{y}_1 - \bar{\bar{y}}$

Der Bezugspunkt der Effekte ist das *nicht gewogene* arithmetische Mittel der Gruppenmittelwerte.

4.6.2 Effekt-Codierung in der einfachen Varianzanalyse

Die Information über eine nominale Variable A mit k Ausprägungen lässt sich auch durch k - 1 Trichotomien derart vercoden, dass die Regressionskoeffizienten gerade die "Effekte", d.h. die Abweichungen der verschiedenen Ausprägungen vom Gesamtmittelwert, sind. Während bei der 1-0-Vercodung mit Dichotomien jeder Gruppenmittelwert $\bar{y}_i$ (für i = 1, ..., k - 1) mit dem Mittelwert einer bestimmten Bezugsgruppe (ohne Beschränkung der Allgemeinheit Gruppe k) verglichen wurde, wird jetzt jeder Gruppenmittelwert $\bar{y}_1$ (i = 1, ..., k) auf das nicht gewogene arithmetische Mittel $\bar{\bar{y}} = \frac{1}{k}\sum_{i=1}^{k} \bar{y}_i$ bezogen.

Ausprägung des Faktors A	Trichotomien			
	T_1	T_2	...	T_{k-1}
1	1	0		
	0	1		
⋮				
			⋱	
				1 0
				0 1
k	-1	-1	...	-1 -1

Anders formuliert: $T_i = 1_{A_i} - 1_{A_k}$ (für i = 1, ..., k - 1).

$\hat{y} = b_0 + b_1 T_1 + ... b_{k-1} T_{k-1}$

Da die Regressionsschätzung für die i-te Ausprägungsgruppe gerade der Mittelwert $\bar{y}_i$ dieser Gruppe ist (d.h. $\hat{y} = \sum_{i=1}^{k} \bar{y}_i 1_{A_i}$), so erhält man: $\bar{y}_i = b_0 + b_i$ (i = 1, ..., k - 1)

$$\bar{y}_k = b_0 - \sum_{j=1}^{k-1} b_j$$

Daraus folgt: $\sum_{j=1}^{k} \bar{y}_j = k \cdot b_0 + \sum_{j=1}^{k-1} b_j - \sum_{j=1}^{k-1} b_j = k \cdot b_0$

Also: $b_0 = \sum_{j=1}^{k} \frac{\bar{y}_j}{k} =: \bar{\bar{y}}$

Der Effekt der i-ten Gruppe (für i = 1, ..., k - 1) ist dann: $b_i = \bar{y}_i - \bar{\bar{y}}$

Mit den Parametern der Grundgesamtheit gilt:
$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$ oder: $\hat{y}_{ij} = y_{ij} - \varepsilon_{ij} = \mu + \alpha_i$

Mit den geschätzten Effekten gilt entsprechend:
$\hat{y}_{ij} = \bar{\bar{y}} + (\bar{y}_i - \bar{\bar{y}})$

Ist die Variable A eine Dichotomie, so reduziert sich die Effekt-Codierung auf die (+1)-(-1)-Codierung:

A	T_1
1	1
2	-1

$\bar{y}_1 = b_0 + b_1$ $\quad b_0 = \bar{\bar{y}}$

$\bar{y}_2 = b_0 - b_1$ $\quad b_1 = \bar{y}_1 - \bar{\bar{y}}$

(Der Unterschied der allgemeinen Formel für die Mittelwerte mit Hilfe der b_i aufgrund der (+1)-(-1)- bzw. Effekt-Codierung besteht in dem Summand $-\sum_{\substack{j=1 \\ j \neq i}}^{k-1} b_j$ (für i = 1, ..., k - 1), der für k = 2 verschwindet.)

Relativierung: Die bessere „Effekt-Codierung“ besteht darin, die unterschiedlichen Besetzungszahlen als Gewichte zu verwenden und dadurch den tatsächlichen allgemeinen Durchschnitt als Bezugspunkt zu wählen.

4.6.3 Effekt-Codierung in der zweifachen Varianzanalyse

Die Effekt-Codierung für die zweifache Varianzanalyse soll wegen der größeren Übersichtlichkeit nur für den Fall dargestellt werden, dass die beiden nominalen unabhängigen Variablen Dichotomien sind.

T sei die Vercodung für Merkmal A, U die Vercodung für Merkmal B.
Der Regressionsansatz lautet dann:
$\hat{y} = a_0 + b\,T + c\,U + d\,T\,U$

Dann erhält man:
$\bar{y}_{A,B} = a_0 + b + c + d$
$\bar{y}_{\bar{A},B} = a_0 - b + c - d$
$\bar{y}_{A,\bar{B}} = a_0 + b - c - d$
$\bar{y}_{\bar{A},\bar{B}} = a_0 - b - c + d$

$$\bar{\bar{y}} := \frac{(\bar{y}_{A,B} + \bar{y}_{\bar{A},B} + \bar{y}_{A,\bar{B}} + \bar{y}_{\bar{A},\bar{B}})}{4}$$

Dann gilt: $a_0 = \bar{\bar{y}}$

$$\bar{\bar{y}}_A := \frac{\bar{y}_{A,B} + \bar{y}_{A,\bar{B}}}{2}$$

$$\bar{\bar{y}}_B := \frac{\bar{y}_{A,B} + \bar{y}_{\bar{A},B}}{2}$$

Damit erhält man:
$\bar{\bar{y}}_A = a_0 + b$

Also: $b = \bar{\bar{y}}_A - \bar{\bar{y}}$

Entsprechend: $c = \bar{\bar{y}}_B - \bar{\bar{y}}$

Ferner: $\bar{y}_{A,B} - \bar{\bar{y}}_A - \bar{\bar{y}}_B + \bar{\bar{y}} = a_0 + b + c + d - (a_0 + b) - (a_0 + c) + a_0 = d$

Diese Codierung liefert also wieder als Regressionskoeffizienten genau die Effekte.

Wenn die Zellenhäufigkeiten gleich sind, sind die nicht gewogenen arithmetischen Mittelwerte gleich den gewogenen arithmetischen Mittelwerten.

Relativierung: Statt dieser verbreiteten „Effekt-Codierung“ ist es wiederum besser, die unterschiedlichen Besetzungszahlen als Gewichte zu verwenden und dadurch den tatsächlichen Durchschnitt als Bezugspunkt zu wählen.

4.7 Die Design-Matrix

Der varianzanalytische Ansatz lautet beispielsweise im Falle der zweifachen Varianzanalyse (für i = 1, ..., k; j = 1, ..., *l*; u = 1, ..., n_{ij}):
$y_{iju} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{iju}$
Er wird durch Regression von y auf Dummy-Variablen für die Merkmalsausprägungen gelöst. Arbeitet man mit der Effekt-Codierung und sind T_i die Codierung der i-ten Kategorie von A und U_j die Codierung der j-ten Kategorie von B, so erhält man z.B. für den Fall k = 3, *l* = 4 den Regressionsansatz:

$$\begin{aligned}\hat{y} = a_0 &+ b_1 T_1 + b_2 T_2 + c_1 U_1 + c_2 U_2 + c_3 U_3 \\ &+ d_{11} T_1 U_1 + d_{12} T_1 U_2 + d_{13} T_1 U_3 \\ &+ d_{21} T_2 U_1 + d_{22} T_2 U_2 + d_{23} T_2 U_3\end{aligned}$$

Schreibt man diese Gleichung für alle Beobachtungswerte untereinander, so erhält man in Matrixschreibweise:

$$\underbrace{\hat{y}}_{(n,\,1)} = \underbrace{X}_{(n,\,k\,l)} \cdot \underbrace{\begin{pmatrix} a_0 \\ b_1 \\ b_2 \\ c_1 \\ c_2 \\ c_3 \\ d_{11} \\ d_{12} \\ d_{13} \\ d_{21} \\ d_{22} \\ d_{23} \end{pmatrix}}_{(k\,l\,,\,1)}$$

$\bar{y} = a_0$

Haupteffekte:
- Effekte von T: b_1, b_2
- Effekte von U: c_1, c_2, c_3

Interaktionseffekte: d_{11}, d_{12}, d_{13}, d_{21}, d_{22}, d_{23}

Hierbei ist X die Design-Matrix, die unter Berücksichtigung der Effekt-Codierung folgende Form hat:

X=

1	T_1	T_2	U_1	U_2	U_3	$T_1 \cdot U_1$	$T_1 \cdot U_2$	$T_1 \cdot U_3$	$T_2 \cdot U_1$	$T_2 \cdot U_2$	$T_2 \cdot U_3$		Entsprechende Beobachtungswerte y
1	1	0	1	0	0	1	0	0	0	0	0	} n_{11}-mal	y_{11u}, u = 1, ..., n_{11}
1	1	0	0	1	0	0	1	0	0	0	0	} n_{12}-mal	y_{12u}, u = 1, ..., n_{12}
1	1	0	0	0	1	0	0	1	0	0	0	} n_{13}-mal	y_{13u}, u = 1, ..., n_{13}
1	1	0	-1	-1	-1	-1	-1	-1	0	0	0	} n_{14}-mal	y_{14u}, u = 1, ..., n_{14}
1	0	1	1	0	0	0	0	0	1	0	0	} n_{21}-mal	y_{21u}, u = 1, ..., n_{21}
1	0	1	0	1	0	0	0	0	0	1	0	} n_{22}-mal	y_{22u}, u = 1, ..., n_{22}
1	0	1	0	0	1	0	0	0	0	0	1	} n_{23}-mal	y_{23u}, u = 1, ..., n_{23}
1	0	1	-1	-1	-1	0	0	0	-1	-1	-1	} n_{24}-mal	y_{24u}, u = 1, ..., n_{24}
1	-1	-1	1	0	0	-1	0	0	-1	0	0	} n_{31}-mal	y_{31u}, u = 1, ..., n_{31}
1	-1	-1	0	1	0	0	-1	0	0	-1	0	} n_{32}-mal	y_{32u}, u = 1, ..., n_{32}
1	-1	-1	0	0	1	0	0	-1	0	0	-1	} n_{33}-mal	y_{33u}, u = 1, ..., n_{33}
1	-1	-1	-1	-1	-1	1	1	1	1	1	1	} n_{34}-mal	y_{34u}, u = 1, ..., n_{34}

Jede Spalte steht für einen Regressionsterm. In der varianzanalytischen Anwendung der Regression für das saturierte Modell, d.h. bei Berücksichtigung aller Interaktionsterme hat man 1 + (k - 1) + (*l* - 1) + (k - 1) (*l* - 1) = kl Regressionsterme, in dem Beispiel also $3 \cdot 4 = 12$.

Bei „proportionalen" Zellenhäufigkeiten (im Sinne der statistischen Unabhängigkeit in einer Kontingenztafel) sind die T_i-Spalten (i = 1, 2) unkorreliert mit den U_j-Spalten (j = 1, 2, 3). Bei gleichen Zellenhäufigkeiten sind die T_i U_j-Spalten (i = 1, 2; j = 1, 2, 3) unkorreliert mit den T_i-Spalten (i = 1, 2) und den U_j-Spalten (j = 1, 2, 3), falls man mit der Effekt-Codierung arbeitet.

Berechnung der Effekt-Schätzungen aufgrund der Regressionskoeffizienten:
Wenn man mit der Effekt-Codierung arbeitet, kann man die Bedingungen

$$\sum_{i=1}^{k} \alpha_i = 0 \ , \ \sum_{j=1}^{l} \beta_j = 0 \ , \ \sum_{i=1}^{k} (\alpha\beta)_{ij} = 0 = \sum_{j=1}^{l} (\alpha\beta)_{ij}$$

(für j = 1, ..., *l* bzw. i = 1, ..., k) benutzen, um *alle* Effekt-Schätzungen zu erhalten.

Zusammenhang zwischen den Regressionskoeffizienten und der Schätzung der Effekte:

Regressionskoeffizienten	Schätzung der Effekte
a_0	μ
b_1	α_1
b_2	α_2
	$\alpha_3 = -\alpha_1 - \alpha_2$
c_1	β_1
c_2	β_2
c_3	β_3
	$\beta_4 = -\beta_1 - \beta_2 - \beta_3$
d_{11}	$(\alpha\beta)_{11}$
d_{12}	$(\alpha\beta)_{12}$
d_{13}	$(\alpha\beta)_{13}$
	$(\alpha\beta)_{14} = -(\alpha\beta)_{11} - (\alpha\beta)_{12} - (\alpha\beta)_{13}$
d_{21}	$(\alpha\beta)_{21}$
d_{22}	$(\alpha\beta)_{22}$
d_{23}	$(\alpha\beta)_{23}$
	$(\alpha\beta)_{24} = -(\alpha\beta)_{21} - (\alpha\beta)_{22} - (\alpha\beta)_{23}$

Betrachtet man die Ausprägungskombination als Einheit der Analyse, so sind alle Besetzungszahlen identisch, nämlich gleich 1. In diesem Fall handelt es sich um ein „orthogonales" Design: Die Effekte von A_i und B_j sind unabhängig, und die Interaktionseffekte A_i B_j sind unabhängig von den Haupteffekten. – Die Anzahl der Einheiten ist gleich dem Produkt der Anzahl der Ausprägungen, d.h. $k \cdot l$. Betrachtet man das *saturierte Modell*, d.h. berücksichtigt man alle Interaktionsterme, so ist also die Anzahl der Fälle gleich der Anzahl der gesuchten Parameter.[11]

In dem Regressionsansatz $\hat{y}$ = Xb ist X dann eine quadratische (k*l*, k*l*)-Matrix. Die Regressionslösung liefert: $b = (X' X)^{-1} X' y$
In diesem Fall vereinfacht sich dies: $b = X^{-1} (X')^{-1} X' y = X^{-1} y$, d.h.: $y = Xb = \hat{y}$.

[11] In dem allgemeinen Regressionsansatz geht man implizit davon aus, dass es erheblich mehr Fälle als Variablen gibt. In der varianzanalytischen Anwendung ist die Situation durch die Einführung zahlreicher Dummy Variablen anders.

Die Daten werden dann durch den Regressionsansatz vollständig (d.h. ohne Fehler) erfasst, im saturierten Modell liefert dieser Ansatz nur eine Umformung der Daten.

Sobald man nicht saturierte Modelle betrachtet, d.h. z.B. nicht alle Interaktionsterme gleichzeitig berücksichtigt, ist die Anzahl der Regressionskoeffizienten kleiner als die Anzahl der Fälle. Es handelt sich dann nicht um eine Umformung, sondern um ein Schätzproblem, das eindeutig lösbar ist, weil es mehr Fälle als zu schätzende Koeffizienten gibt.

4.8 Kovarianzanalyse

Die Kovarianzanalyse unterscheidet sich von der Varianzanalyse dadurch, dass zusätzlich zu den nominalen unabhängigen Variablen (Faktoren) noch metrische unabhängige Variable (Kovariate) berücksichtigt werden.
Dabei wird varianzanalytisches Vorgehen mit regressionsanalytischer Verfahrensweise verkoppelt: Die Kovariate üben die Funktion von Kontrollvariablen aus, indem sie zur Schätzung der abhängigen Variablen herangezogen werden. Dabei wird der Einfluss der Kovariate aus der durch die Faktoren bei der abhängigen Variablen verursachten Varianz gleichsam herauspartialisiert.

y bezeichnet wieder eine abhängige metrische Variable, z.B. die Affinität zur CDU/CSU.

Arbeitet man wieder mit der Regression auf dichotomen Variablen, so führt die Einführung z.B. einer zusätzlichen metrischen Variablen „Einkommen“ (E) zu folgender Regressionsschätzung, falls noch keine Interaktionen berücksichtigt werden:

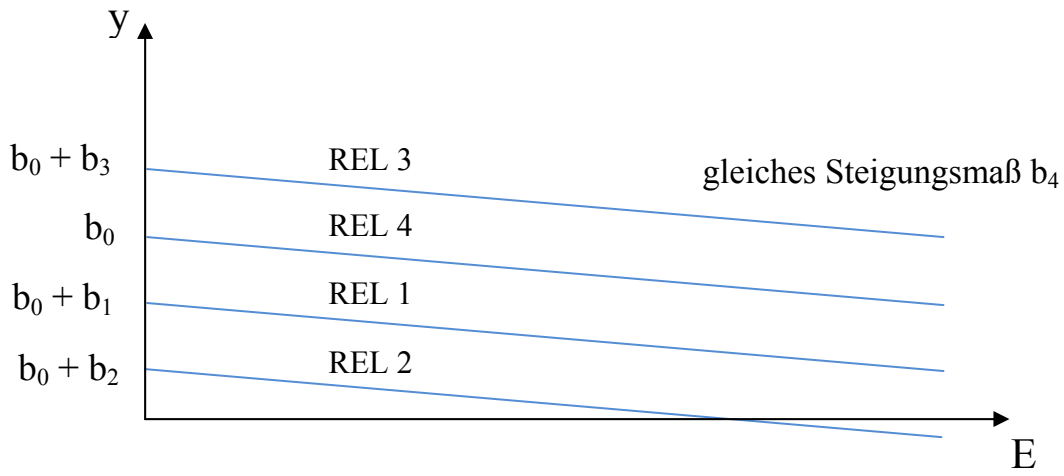
$$\hat{y} = b_0 + b_1\ \mathrm{REL1} + b_2\ \mathrm{REL2} + b_3\ \mathrm{REL3} + b_4\ \mathrm{E}$$

Innerhalb der einzelnen Kategorien wird ein gemeinsamer Steigungs-Koeffizient eingesetzt. Die Schätzgleichungen unterscheiden sich nur durch ihre gruppenspezifische Höhenlage (= Abweichung von der Referenzkategorie Rel 4):

$\hat{y}$ (Rel 1) = $b_0 + b_1 + b_4$ E $\hat{y}$ (Rel 3) = $b_0 + b_3 + b_4$ E

$\hat{y}$ (Rel 2) = $b_0 + b_2 + b_4$ E $\hat{y}$ (Rel 4) = $b_0 + b_4$ E

Abbildung 4-4: Graphische Darstellung des Zusammenhangs (mit Interaktionen)

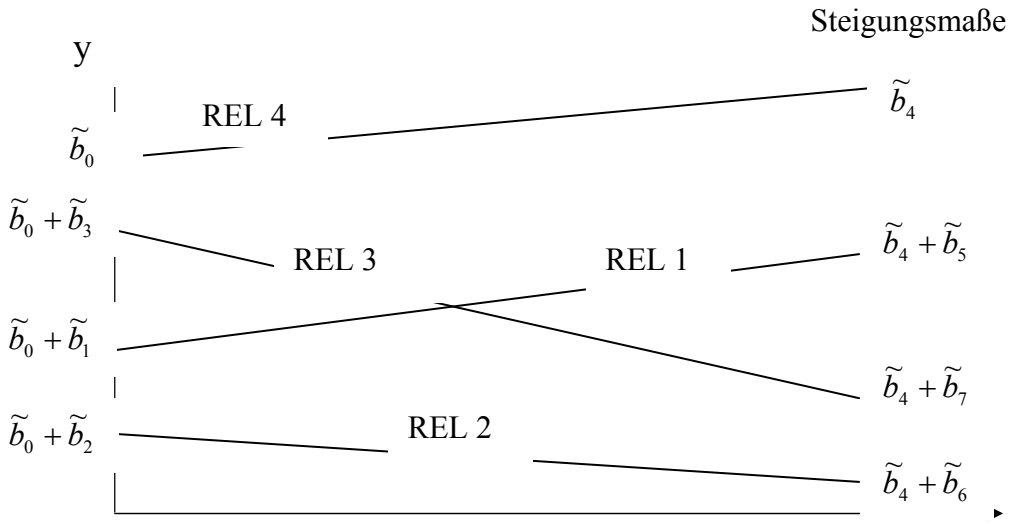


Geht man davon aus, dass die Wirkung der Kontrollvariablen „Einkommen“ je nach Gruppenzugehörigkeit unterschiedlich ist, sind zusätzlich Interaktionsterme zu berücksichtigen.

Das saturierte (alle möglichen Interaktionsterme enthaltende) Regressionsmodell lautet:
$\hat{y} = \tilde{b}_0 + \tilde{b}_1 \text{REL1} + \tilde{b}_2 \text{REL2} + \tilde{b}_3 \text{REL3} + \tilde{b}_4 \text{E} + \tilde{b}_5 (\text{REL1} \cdot \text{E}) + \tilde{b}_6 (\text{REL2} \cdot \text{E}) + \tilde{b}_7 (\text{REL3} \cdot \text{E})$

Die Interaktionseffekte zwischen den Faktoren und der metrischen Kontrollvariablen auf die abhängige Variable y können graphisch durch unterschiedliche Steigungen der Regressionsgeraden veranschaulicht werden.

Abbildung 4-5: Graphische Darstellung des Zusammenhangs (mit Interaktionen)



$$\hat{y}(REL1) = (\tilde{b}_0 + \tilde{b}_1) + (\tilde{b}_4 + \tilde{b}_5)E$$
$$\hat{y}(REL2) = (\tilde{b}_0 + \tilde{b}_2) + (\tilde{b}_4 + \tilde{b}_6)E$$
$$\hat{y}(REL3) = (\tilde{b}_0 + \tilde{b}_3) + (\tilde{b}_4 + \tilde{b}_7)E$$
$$\hat{y}(REL4) = \tilde{b}_0 + \tilde{b}_4 E$$

Im Allgemeinen wird in der Literatur das einfache Modell ohne Interaktionen behandelt. Es wird hierbei von einer homogenen Regression für die einzelnen Gruppen ausgegangen.

4.8.1 Kovarianzzerlegung

Die folgende Darstellung beschreibt den Zusammenhang zweier metrischer Variablen unter Kontrolle einer nominalen Variablen.

4.8.1.1 Kovarianzzerlegung nach einer nominalen unabhängigen Variablen

Der Zusammenhang zwischen zwei metrischen Variablen x und y kann derart untersucht werden, dass der Einfluss einer nominalen dritten Variablen berücksichtigt wird. Die dritte Variable möge k Ausprägungen haben, die mit n_1, ..., n_k $\left(\sum_{i=1}^{k} n_i = n\right)$ Elementen (Untersuchungseinheiten) besetzt sind.

Die Kovarianzzerlegung in erklärte und nicht erklärte Varianz kann folgendermaßen hergeleitet werden:

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(x_{ij}-\bar{x}\right)\left(y_{ij}-\bar{y}\right)$$

$$= \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\left(x_{ij}-\bar{x}_i\right)+\left(\bar{x}_i-\bar{x}\right)\right)\left(\left(y_{ij}+\bar{y}\right)+\left(\bar{y}_i-\bar{y}\right)\right)$$

$$= \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(x_{ij}-\bar{x}_i\right)\left(y_{ij}-\bar{y}\right)+\frac{1}{n}\sum_{i=1}^{k} n_i\left(\bar{x}_i-\bar{x}\right)\left(\bar{y}_i-\bar{y}\right)$$

$$= \frac{1}{n}\sum_{i=1}^{k} n_i\, s_i(x,y)+\frac{1}{n}\sum_{i=1}^{k} n_i\left(\bar{x}_i-\bar{x}\right)\left(\bar{y}_i-\bar{y}\right)$$

"within"	"between"
(nicht erklärt)	(erklärt)

Hierbei ist $s_i(x,y) := \frac{1}{n_i}\sum_{j=1}^{n_i}\left(x_{ij}-\bar{x}_i\right)\left(y_{ij}-\bar{y}_i\right)$ die Kovarianz von x und y in der i-ten Gruppe (i = 1, ..., k). Der zweite Summand, der die Unterschiede *zwischen* den Gruppen erfasst, heißt *externe* Kovarianz.
Die gesamte Kovarianz lässt sich also zerlegen in das gewogene arithmetische Mittel der Kovarianzen innerhalb der Gruppen und in die externe Kovarianz.

Die gemischten Glieder in der Zerlegung sind 0:

$$\sum_i\sum_j\left(x_{ij}-\bar{x}_i\right)\left(\bar{y}_i-\bar{y}\right) = \sum_i\left(\bar{y}_i-\bar{y}\right)\underbrace{\sum_j\left(x_{ij}-\bar{x}_i\right)}_{0} = 0$$

Für das zweite gemischte Glied gilt Entsprechendes.

Speziell für x = y erhält man die Zerlegung der Varianz:

$$s_y^2 = \sum_{i=1}^{k}\frac{n_i}{n}s_i^2(y) \;+\; \frac{1}{n}\sum_{i=1}^{k} n_i\left(\bar{y}_i-\bar{y}\right)^2$$

Streuung innerhalb der Gruppen	Streuung zwischen den Gruppen

Wie in der Varianzanalyse könnte man auch in der Kovarianzanalyse verallgemeinern auf den Fall zweier Faktoren A und B, wobei wieder die Interaktion AB zu berücksichtigen wäre. Es wären drei Fälle zu unterscheiden:

a) die Zellhäufigkeiten sind gleich,
b) die Zellhäufigkeiten sind proportional oder
c) die Zellhäufigkeiten sind beliebig.

In jedem Fall gilt die Zerlegung in die Streuung innerhalb und zwischen den Zellen:

$$\frac{1}{n}\sum_{i}^{k}\sum_{j}^{l}\sum_{u=1}^{n_{ij}}\left(x_{iju}-\bar{x}\right)\left(y_{iju}-\bar{y}\right)=\frac{1}{n}\sum\sum\sum\left(x_{iju}-\bar{x}_{ij}\right)\left(y_{iju}-\bar{y}_{ij}\right)+\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{l}n_{ij}\left(\bar{x}_{ij}-\bar{x}\right)\left(\bar{y}_{ij}-\bar{y}\right)$$

"within" (nicht erklärt) "between" (erklärt)

Die gemischten Glieder in der Zerlegung sind 0:

$$\sum_{i}\sum_{j}\sum_{u}\left(x_{iju}-\bar{x}_{ij}\right)\left(\bar{y}_{ij}-\bar{y}\right)=\sum_{i}\sum_{j}\left(\bar{y}_{ij}-\bar{y}\right)\underbrace{\sum_{u}\left(x_{iju}-\bar{x}_{ij}\right)}_{0}=0$$

Für das zweite gemischte Glied gilt Entsprechendes.

Speziell für x = y ist dies das bereits dargestellte Ergebnis der Varianzanalyse:
$SS_y = SS_{error} + SS_{A,\,B,\,AB}$

4.8.1.2 Anwendung der Kovarianzzerlegung: Aggregatdaten und Mehrebenenanalyse

Ökologische, d.h. auf räumliche Einheiten bezogene Daten, findet man häufig in der Amtlichen Statistik. Unter einem "ökologischen Fehlschluss" oder allgemeiner: einem Fehlschluss von Aggregatdaten auf Individualdaten versteht man den falschen Schluss von einem gefundenen Zusammenhang auf Gruppenebene auf einen entsprechenden Zusammenhang auf Individualebene. So sind nach Durkheim Ehescheidungen Indikatoren für ein ungünstiges Familienklima, welches Anomie (gemessen über den Indikator „Selbstmord“) verursacht. Es folgt nicht notwendig, dass Geschiedene selbstmordgefährdeter sind. Nach der Terminologie der Kovarianzzerlegung (s.o.) erfolgt der "ökologische Fehlschluss" von der externen Kovarianz auf die Gesamt-Kovarianz. Dieser Schluss ist aber nur zulässig, wenn die Kovarianzen in den Gruppen Null sind. Die Gesamt-Kovarianz beruht dann nur auf der Gruppierung in ökologische Einheiten. Im allgemeinen gilt aber:

$$s_{xy} > s_{\bar{x}_i,\bar{y}_i} \text{ und } s_i(x,y) \neq 0$$

Der umgekehrte Fehlschluss heißt "individualistischer Fehlschluss": Nach Michels "ehernem Gesetz der Oligarchie" kann man nicht von den anti-autoritären Einstellungen von Sozialisten auf nicht-bürokratische Formen der zugehörigen Organisationen (sozialistische Parteien) schließen. Hierbei wird von der Gesamt-Kovarianz auf die externe Kovarianz geschlossen. Dieser Schluss ist wiederum nur zulässig, wenn s_i (x, y) = 0.

Sind die Ausgangsdaten Aggregatdaten (z.B. für Gebietseinheiten, d.h. ökologische Daten), so kann man nur den Zusammenhang zwischen Mittelwerten verschiedener Variablen, die für die einzelnen Gebietseinheiten vorliegen, berechnen, wenn man über die Individualdaten nicht verfügt. Die Beziehung zwischen diesen verschiedenen Zusammenhangsmaßen lässt sich aber wie folgt angeben:

Der übliche Korrelationskoeffizient für Individualdaten ist:

$$r=\frac{\sum_{i}\sum_{j}\left(x_{ij}-\bar{x}\right)\left(y_{ij}-\bar{y}\right)}{\sqrt{\sum_{i}\sum_{j}\left(x_{ij}-\bar{x}\right)^2}\sqrt{\sum_{i}\sum_{j}\left(y_{ij}-\bar{y}\right)^2}}$$

Den Zwischengruppenkorrelationskoeffizienten r_z (ökologische Korrelation) erhält man, indem man die Mittelwerte mit der zugehörigen Besetzungszahl der Gruppe (Gebiet) gewichtet (weil man nur über die Mittelwerte für die einzelnen Gebiete verfügt):

$$r_z = \frac{\sum n_i (\bar{x}_i - \bar{x})(\bar{y}_i - \bar{y})}{\sqrt{\sum n_i (x_i - \bar{x})^2} \sqrt{\sum n_i (\bar{x}_i - \bar{x})^2}}$$

Um die Beziehung zwischen r und r_z zu formulieren, braucht man als 3. Korrelationskoeffizienten die "Intragruppenkorrelation" r_I, die wie folgt definiert werden muss:

$$r_I = \frac{\sum_i \sum_j (x_{ij} - \bar{x}_i)(y_{ij} - \bar{y}_i)}{\sqrt{\sum_i \sum_j (x_{ij} - \bar{x}_i)^2} \sqrt{\sum_i \sum_j (y_{ij} - \bar{y}_i)^2}}$$

Als Gewichtungskoeffizienten braucht man noch (mit der Bezeichnung A für die nominale Gruppen- bzw. Gebietsvariable) den Anteil der durch die Kategorienbildung nach A erklärten Varianz von x: $\eta_{xA}^2 = \frac{\sum_i n_i (\bar{x}_i - \bar{x})^2}{\sum_i \sum_j (x_{ij} - \bar{x})^2}$

Entsprechend ist der durch A erklärte Varianzanteil von y: $\eta_{yA}^2 = \frac{\sum_i n_i (\bar{y}_i - \bar{y})^2}{\sum_i \sum_j (y_{ij} - \bar{y})^2}$

Mit Hilfe von r_I und η_{xA}, η_{yA} lässt sich die Beziehung zwischen r und r_z wie folgt bestimmen:

$$r = \sqrt{1 - \eta_{xA}^2} \sqrt{1 - \eta_{yA}^2} r_I + \eta_{xA} \eta_{yA} r_z$$

Oder: $r_z = \frac{r - \sqrt{1 - \eta_{xA}^2} \sqrt{1 - \eta_{yA}^2}\, r_I}{\eta_{xA} \eta_{yA}}$

Der „ökologische“ und der „individualistische“ Fehlschluss beruhen auf der Gleichsetzung:
$r_z = r$

Beweis der Kovarianzzerlegung :

$$r = \frac{S_{xy}}{S_x \cdot S_y} = \frac{\sum_i \sum_j (x_{ij} - \bar{x}_i)(y_{ij} - \bar{y}_i)}{\sqrt{\sum_i \sum_j (x_{ij} - \bar{x})^2} \sqrt{\sum_i \sum_j (y_{ij} - \bar{y})^2}} + \frac{\sum n_i (\bar{x}_i - \bar{x})(\bar{y}_i - \bar{y})}{\sqrt{\sum_i \sum_j (x_{ij} - \bar{x})^2} \sqrt{\sum_i \sum_j (y_{ij} - \bar{y})^2}}$$

Ferner:

$$\eta_{xA}\eta_{yA}r_z = \frac{\sqrt{\sum_i n_i(\bar{x}_i-\bar{x})^2}}{\sqrt{\sum_i\sum_j (x_{ij}-\bar{x})^2}} \frac{\sqrt{\sum_i (\bar{y}_i-\bar{y})^2}}{\sqrt{\sum_i\sum_j (y_{ij}-\bar{y})^2}} \frac{\sum n_i(\bar{x}_i-\bar{x})(\bar{y}_i-\bar{y})}{\sqrt{\sum_i n_i(\bar{x}_i-\bar{x})^2}\sqrt{\sum_i n_i(\bar{y}_i-\bar{y})^2}}$$

$$= \frac{\sum n_i(\bar{x}_i-\bar{x})(\bar{y}_i-\bar{y})}{\sqrt{\sum_i\sum_j (x_{ij}-\bar{x})^2}\sqrt{\sum_i\sum_j (y_{ij}-\bar{y})^2}}$$

Wegen der Streuungszerlegung $\sum_i\sum_j (x_{ij}-\bar{x})^2 = \sum_i\sum_j (x_{ij}-\bar{x}_i)^2 + \sum_i n_i(\bar{x}_i-\bar{x})^2$ gilt:

$$\frac{\sum_i\sum_j (x_{ij}-\bar{x}_i)^2}{\sum_i\sum_j (x_{ij}-\bar{x})^2} = 1-\eta_{xA}^2$$

Entsprechend: $\frac{\sum_i\sum_j (y_{ij}-\bar{y}_i)^2}{\sum_i\sum_j (y_{ij}-\bar{y})^2} = 1-\eta_{yA}^2$

Daraus folgt:

$$\sqrt{1-\eta_{xA}^2}\sqrt{1-\eta_{xA}^2}\, r_I = \frac{\sqrt{\sum_i\sum_j (x_{ij}-\bar{x}_i)^2}}{\sqrt{\sum_i\sum_j (x_{ij}-\bar{x})^2}} \frac{\sqrt{\sum_i\sum_j (y_{ij}-\bar{y}_i)^2}}{\sqrt{\sum_i\sum_j (y_{ij}-\bar{y})^2}} \frac{\sum_i\sum_j (x_{ij}-\bar{x}_i)(y_{ij}-\bar{y}_i)}{\sqrt{\sum_i\sum_j (x_{ij}-\bar{x}_i)^2}\sqrt{\sum_i\sum_j (y_{ij}-\bar{y}_i)^2}}$$

$$= \frac{\sum_i\sum_j (x_{ij}-\bar{x}_i)(y_{ij}-\bar{y}_i)}{\sqrt{\sum_i\sum_j (x_{ij}-\bar{x})^2}\sqrt{\sum_i\sum_j (y_{ij}-\bar{y})^2}}$$

L

Erklärt die Gebietseinteilung x und y vollständig ($\eta_{xA} = 1$, $\eta_{yA} = 1$), so geht bei der Berechnung der ökologischen Korrelation keine Information über x und y verloren, so dass die ökologische Korrelation gleich der individuellen Korrelation ist.
Erklärt die Gebietseinteilung gar keinen Anteil der Variation von x und y ($\eta_{xA} = 0$, $\eta_{yA} = 0$), so ist die individuelle Korrelation gleich der Korrelation *innerhalb* der Gruppen, da die Gruppeneinteilung keinen Erklärungsbeitrag leistet ($r = r_I$).

Falls $r = r_z$, so folgt (mit einer Konstante $c \geq 1$):

$$r_I = \frac{1-\eta_{xA}\eta_{yA}}{\sqrt{1-\eta_{xA}^2}\sqrt{1-\eta_{yA}^2}} \quad r =: c \cdot r$$

Denn: $a^2 + b^2 - 2ab = (a - b)^2 \geq 0$, also: $a^2 + b^2 \geq 2ab$
$(1 - ab)^2 = 1 + a^2 b^2 - 2ab \geq 1 + a^2 b^2 - a^2 - b^2 = (1 - a^2)(1 - b^2)$

Ferner: $1 - \eta_{xA} \eta_{yA} \geq 0$, da $0 \leq \eta_{xA} \leq 1$, $0 \leq \eta_{yA} \leq 1$, also $0 \leq \eta_{xA} \eta_{yA} \leq 1$.

$|r_I| = c \cdot |r| \geq |r|$
Die ökologische Korrelation kann nur dann mit der individuellen Korrelation übereinstimmen, wenn die Intragruppenkorrelation dem absoluten Betrag nach größer oder gleich der individuellen Korrelation ist.

Empirisch sind ökologische Korrelationen *häufig* größer als die entsprechenden individuellen Korrelationen. An welchen Bedingungen kann dies liegen?
$(r < r_z) \Leftrightarrow (r_I < c \cdot r)$

Denn: $r > r_z = \dfrac{\left(r - \sqrt{1-a^2}\sqrt{1-b^2 r_I}\right)}{ab} \Leftrightarrow r(1-ab) > \sqrt{1-a^2}\sqrt{1-b^2 r_I}$

Ist nun die Intragruppenkorrelation kleiner als die individuelle Korrelation ($r_I < r$), so folgt:
$r_I < c \cdot r$ (falls $r > 0$), d.h. dass in diesem Fall die ökologische Korrelation größer ist als die individuelle Korrelation.

Empirisch wächst die ökologische Korrelation *häufig* mit der Größe der Gebietseinheiten.

Als Alternative zur „Aggregatpsychologie“ (Coleman), die sich auf Durchschnitte und Verteilungen von Individualmerkmalen beschränkt, wird die Mehrebenenanalyse verstanden. Als Ebenen könnten etwa dienen: 1) Personen 2) Gruppen 3) Organisationen 4) Gesamtgesellschaft. Für das Konzept der relativen Deprivation (vgl. Stouffer et al. 1949) ist es z.B. erforderlich, zwischen einem Individualmerkmal zu unterscheiden (Zufriedenheit mit Behörden) und der Gruppenvariablen (Wahrscheinlichkeit einer Beförderung). Die Benachteiligung wird in Relation zur Bezugsgruppe empfunden.

4.8.1.3 Die Kovarianz- und Korrelationszerlegung nach einer nominalen unabhängigen Variablen als Spezialfall einer allgemeinen Kovarianz- und Korrelationszerlegung nach metrischen Variablen

Betrachten wir zunächst die Beziehung zwischen zwei dichotomisierten Variablen i und j, die nach einer dritten dichotomen Variable k bzw. $\bar{k}$ („Verneinung von k“) zerlegt wird.

Nach der Lazarsfeld'schen Grundgleichung für die Differenz des Kreuzprodukts gilt (wobei sich die Anteile p immer auf die Gesamtzahl n und nicht auf die Fallzahl n_k der k-ten Untergruppe beziehen):

$$|ij| = \frac{|ij:k|}{p_k} + \frac{|ij:\bar{k}|}{p_{\bar{k}}} + \frac{|ik\|jk|}{p_k \cdot p_{\bar{k}}}$$

$|ij| = s_{ij}$, $|ik| = s_{ik}$, $|jk| = s_{jk}$, $s_k^2 = p_k \cdot p$

Aber: $|ij : k|$ ist nicht die Kovarianz in Gruppe k.

Denn: $|ij:k| = \begin{vmatrix} p_{ijk} & p_{ik} \\ p_{jk} & p_k \end{vmatrix} = \frac{n_{ijk}}{n}\frac{n_k}{n} - \frac{n_{ik}}{n}\frac{n_{jk}}{n} = \left(\frac{n_{ijk}}{n_k} - \frac{n_{ik}}{n_k}\frac{n_{jk}}{n_k}\right)\cdot p_k^2$

Also: $\frac{|ij:k|}{p_k} = p_k \cdot s_{k(i,j)}$, wobei s_k (i, j) die Kovarianz in der k-ten Gruppe bezeichnet.[12]

Die Zerlegung für die Kovarianz lautet also:

$$s_{ij} = p_k s_k(i,j) + p_{\bar{k}} s_{\bar{k}}(i,j) + \frac{s_{ik} s_{jk}}{s_k^2}$$

Die allgemeine Kovarianzzerlegung lautet (s.o.):

$$s_{xy} = \sum_{k=1}^{m} \frac{n_k}{n} s_k(x,y) + \frac{1}{n}\sum_{k=1}^{m} n_k\left(\bar{k}_k - \bar{x}\right)\left(\bar{y}_k - \bar{y}\right)$$

Lazarsfeld hat nun seine Grundgleichung noch verallgemeinert auf den Fall, in dem i und j zwar noch Dichotomien sein müssen, die dritte Variable aber k Ausprägungen haben kann.
Aus |ij : k| = p_{ijk} p_k - p_{ik} p_{jk} folgt:

$$p_{ijk} = \frac{|ij:k| + p_{ik}\, p_{jk}}{p_k}$$

Da $p_{ij} = \sum_{k=1}^{m} p_{ijk}$, so erhält man nach Lazarsfeld:

$$|ij| = \begin{vmatrix} p_{ij} & p_j \\ p_i & 1 \end{vmatrix} = \begin{vmatrix} \sum_{k=1}^{m} \frac{|ij:k|}{p_k} + \sum_{k=1}^{m} \frac{p_{ik}\, p_{jk}}{p_k} & p_j \\ p_i & 1 \end{vmatrix} = \sum_{k=1}^{m} \frac{|ij:k|}{p_k} + \left(\sum_{k=1}^{m} \frac{p_{ik}\, p_{jk}}{p_k} - p_i\, p_j\right)$$

Da $\frac{|ij:k|}{p_k} = p_k \cdot s_k(i,j)$,so folgt:

Die externe Kovarianz $\frac{1}{n}\sum_{k=1}^{m} n_k\left(\bar{x}_k - \bar{x}\right)\left(\bar{y}_k - \bar{y}\right) = \sum_{k=1}^{m} \frac{n_k}{n}\bar{x}_k\,\bar{y}_k - \bar{x}\,\bar{y}$ ist in diesem Fall gleich $\sum_{k=1}^{m} \frac{p_{ik}\, p_{jk}}{p_k} - p_i\, p_j$.

Lazarsfelds Zerlegung lässt sich verallgemeinern auf den Fall, dass nach zwei nominalen Merkmalen aufgegliedert wird (wie in der zweifachen Varianzanalyse und entsprechenden Kovarianzanalyse).

[12] Der Unterschied besteht darin, dass Lazarsfelds Konzept die Teilgruppe nur als Teil der Gesamtheit behandelt (Bezugsgröße: n), nicht aber als selbständige Einheit (Bezugsgröße: n_k).

Es gilt dann: $|ij : kl| = p_{ijkl}\, p_{kl} - p_{ikl}\, p_{jkl}$

Also: $p_{ijkl} = \dfrac{|ij : kl| + p_{ikl}\ p_{jkl}}{p_{kl}}$

Ferner: $p_{ij} = \sum_{k,l} p_{ijkl}$

Daraus folgt:

$$|ij| = \begin{vmatrix} p_{ij} & p_i \\ p_j & 1 \end{vmatrix} = \begin{vmatrix} \sum_{k,l} \dfrac{|ij : kl|}{p_{kl}} + \sum_{k,l} \dfrac{p_{ikl}\, p_{jkl}}{p_{kl}} & p_i \\ p_j & 1 \end{vmatrix} = \left(\sum_{k,l} \frac{|ij : kl|}{p_{kl}} \right) + \left(\sum_{k,l} \frac{p_{ikl}\, p_{jkl}}{p_{kl}} - p_i\ p_j \right)$$

Es gilt wieder: $p_{ijkl} = \dfrac{n_{ijkl}}{n}\dfrac{n_{kl}}{n} - \dfrac{n_{ikl}}{n}\dfrac{n_{jkl}}{n} = \left(\dfrac{n_{ijkl}}{n_{kl}} - \dfrac{n_{ikl}\ n_{jkl}}{n_{kl}\ n_{kl}} \right) \dfrac{n_{kl}^2}{n^2}$

Also: $\dfrac{p_{ijkl}}{p_{kl}} = p_{kl} \cdot s_{kl}(ij)$

Diese Zerlegung ist also die Kovarianzzerlegung nach zwei nominalen Variablen für den Spezialfall der Kovarianz zweier Dichotomien (vgl. die allgemeinere Form in Punkt 4.8.1.1).

Nach meiner Meinung spricht vieles für die allgemeine Kovarianzzerlegung zweier metrischer Variablen x und y nach einer nominalen Variablen und einer metrischen Variablen T.

1. Die Zerlegung nach der nominalen Variablen mit m Ausprägungen:

$$s_{xy} = \sum_{k=1}^{m} \frac{n_k}{n} s_k(x,y) + \sum_{k=1}^{m} \frac{n_k}{n} (\bar{x}_k - \bar{x})(\bar{y}_k - \bar{y})$$

Sind x und y Dichotomien, so ist die externe Kovarianz

$\sum_{k=1}^{m} \dfrac{n_k}{n} (\bar{x}_k - \bar{x})(\bar{y}_k - \bar{y}) = \sum_{k=1}^{m} \dfrac{n_k}{n} \bar{x}_k\ \bar{y}_k - \bar{x}\ \bar{y} = \sum_{k=1}^{m} \dfrac{p_{ik}\ p_{jk}}{p_k} - p_i\ p_j$, woraus Lazarsfelds Version folgt, wenn man zusätzlich berücksichtigt: $p_k s_k(x,y) = \dfrac{|ij : k|}{p_k}$

2. Die Zerlegung nach einer metrischen Variablen T:

$$s_{xy} = s_{xy.T} + \frac{s_{xT} s_{yT}}{s_T^2}$$

Sind x, y und T Dichotomien, so ist $s_{xy.T} = p_k\, s_k\,(x, y) + p_k\, s_k\,(x, y)$. Lazarsfelds Zerlegungsformel folgt hieraus wieder[13], weil: $p_k s_k(x,y) = \frac{|ij:k|}{p_k}$

Letzteres lässt sich noch weiter verallgemeinern:
Sind x, y und $T_1, ..., T_m$ metrische Variablen, so seien $\hat{x}$ und $\hat{y}$ die Regressionsschätzungen von x und y auf $T_1, ..., T_m$. Dann gilt: $\langle x - \hat{x}, y - \hat{y} \rangle = \langle x, y - \hat{y} \rangle = \langle x, y \rangle - \langle x, \hat{y} \rangle$, denn $y - \hat{y}$ ist orthogonal zu $T_1, ..., T_m$, also auch zu $\hat{x}$.

Also: $s_{xy} = s_{xy.T_1,...,T_m} + s_{x,\hat{y}(T_1,...,T_m)}$
(Wegen der Symmetrie der Formel folgt: $s_{x,\hat{y}} = s_{\hat{x},y}$)

Für die Regressionsschätzung gilt: $\langle x, T_i \rangle = \langle \hat{x}, T_i \rangle$
Also gilt auch: $\langle x, \hat{y} \rangle = \langle \hat{x}, \hat{y} \rangle$

Eine symmetrische Formulierung der allgemeinen Kovarianzzerlegung lautet also:
$s_{xy} = s_{xy.T_1,...,T_m} + s_{\hat{x}(T_1,...,T_m),\hat{y}(T_1,...,T_m)}$

⌜

Speziell für m = 1: $\hat{y}(T) = \frac{s_{yT}}{s_T^2}(T - \overline{T}) + \text{constans}$

$$s_{x,\hat{y}(T)} = \frac{s_{yT}}{s_T^2} \cdot \frac{< x - \overline{x}, T - \overline{T} >}{n} = \frac{s_{yT} s_{xT}}{s_T^2}$$

Dies zeigt, dass es sich um eine Verallgemeinerung von (2., S. 204) handelt.

⌞

Speziell für x = y gilt also: $s_x^2 = s_{x-\hat{x}(T_1,...,T_m)}^2 + s_{\hat{x}(T_1,...,T_m)}^2$

⌜

Speziell für m = 1: $\hat{x}(T) = \frac{s_{xT}}{s_T^2}(T - \overline{T}) + \text{constans}$

$$s_{\hat{x}(T)} = \frac{s_{xT}^2}{s_T^4} \cdot \frac{< T - \overline{T}, T - \overline{T} >}{n} = \frac{s_{xT}^2}{s_T^2}$$

⌞

Die allgemeine Kovarianzzerlegung nach einer nominalen Variablen mit m Ausprägungen folgt aus dieser allgemeinen Kovarianzzerlegung, wenn $T_1, ..., T_{m-1}$ Dichotomien sind, die das Zutreffen (= 1) oder das Nicht-Zutreffen (= 0) der Ausprägungen anzeigen.

---

[13] Dies ergibt natürlich auch sofort den Zusammenhang der Begriffe „partiell“ und „bedingt“: $s_{xy.T} = \frac{|xy:k|}{p_k} + \frac{|xy:\bar{k}|}{p_{\bar{k}}}$

Beweis:
Aus der Betrachtung der Varianzanalyse als Spezialfall der Regression folgte:

$$\hat{y}(T_1,...,T_{m-1}) = \sum_{k=1}^{m} \bar{y}_k 1_{A_k}$$

$$s_{x,1_{A_k}} = \sum_{i,j} \frac{(x_{ij} - \bar{x})(1_{A_k}(i,j) - p_k)}{n} = \sum_{i,j} \frac{(x_{ij} - \bar{x}) 1_{A_k}(i,j)}{n} = \sum_{j} \frac{(x_{kj} - \bar{x})}{n}$$

$$= \frac{(n_k \bar{x}_k - n_k \bar{x})}{n} = p_k (\bar{x}_k - \bar{x})$$

Also: $s_{x,\hat{y}} = \sum_{k=1}^{m} \bar{y}_k s_{x,1_{A_k}} = \sum_{k=1}^{m} \bar{y}_k p_k (\bar{x}_k - \bar{x}) = \sum_{k=1}^{m} p_k (\bar{y}_k - \bar{y})(\bar{x}_k - \bar{x})$

Mittelwertberechnung:

$$\sum_{i,j} \frac{(y_{ij} - \bar{y}_k) 1_{A_k(i,j)}}{n} = \sum_{j} \frac{(y_{kj} - \bar{y}_k)}{n} = 0$$

Deshalb: $s_{x,(y-\bar{y}_k)1_{A_k}} = \sum_{i,j} \frac{(x_{ij} - \bar{x})(y_{ij} - \bar{y}_k) 1_{A_k(i,j)}}{n}$

$$= \sum_{j} \frac{(x_{kj} - \bar{x})(y_{kj} - \bar{y}_k)}{n}$$

$$= \sum_{j} \frac{(x_{kj} - \bar{x}_k)(y_{kj} - \bar{y}_k)}{n}$$

$$= p_k \, s_k(x,y)$$

Insgesamt[14]: $s_{x-\hat{x},y-\hat{y}} = s_{x,y-\hat{y}} = s_{x,\sum_{k=1}^{m}(y-\bar{y}_k)1_{A_k}}$

$$= \sum_{k=1}^{m} s_{x,(y-\bar{y}_k)1_{A_k}}$$

$$= \sum_{k=1}^{m} p_k s_k(x,y)$$

Die entsprechende allgemeine Zerlegung des Korrelationskoeffizienten lautet:

$$r_{xy} = \frac{s_{xy}}{s_x \cdot s_y} = \frac{s_{xy.T_1,...,T_m}}{s_x \cdot s_y} + \frac{s_{\hat{x}(T_1,...,T_m),\hat{y}(T_1,...,T_m)}}{s_x \cdot s_y}$$

[14] Für eine dichotome Kontrollvariable T, $\bar{T}$ (d.h. m = 2) folgt: $s_{xy.T} = p_k s_k(x,y) + p_k s_k(x,y)$

Der erste Summand lautet:

$$r_{xy.T_1,...,T_m} \cdot \frac{s_{x-\hat{x}}(T_1,...,T_m) \cdot s_{y-\hat{y}}(T_1,...,T_m)}{s_x \cdot s_y} = r_{xy.T_1,...,T_m} \cdot \sqrt{1-R^2_{x.T_1,...,T_m}} \sqrt{1-R^2_{y.T_1,...,T_m}}$$

(Für m = 1 folgt das bekannte Ergebnis: $r_{xy.T} \sqrt{1-r^2_{xT}} \sqrt{1-r^2_{yT}}$)

Zum zweiten Summand:

$$\frac{\frac{<\hat{x},x>}{n}}{s_x} = \frac{\frac{<\hat{x},\hat{x}>}{n}}{s_x} = R^2_{x.T_1,...,T_m}$$

Daraus folgt: $\frac{\frac{<\hat{x},\hat{y}>}{n}}{s_x \cdot s_y} = \frac{\frac{<\hat{x},\hat{y}>}{n}}{s_{\hat{x}} s_{\hat{y}}} \cdot \frac{s_{\hat{x}}}{s_x} \cdot \frac{s_{\hat{y}}}{s_y} = r_{\hat{x},\hat{y}} \cdot R_{x.T_1,...,T_m} R_{y.T_1,...,T_m}$

(Für m = 1 folgt das bekannte Resultat: $r_{\hat{x}(T),\hat{y}(T)} r_{x.T} r_{y.T} = r_{xT} r_{yT}$)

Die allgemeine Zerlegung des Korrelationskoeffizienten für metrische Variablen x, y, T_1, ..., T_m lautet also:

$$r_{xy} = r_{xy.T_1,...,T_m} \sqrt{1-R^2_{x.T_1,...,T_m}} \sqrt{1-R^2_{y.T_1,...,T_m}} + r_{\hat{x}(T_1,...,T_m),\hat{y}(T_1,...,T_m)} \cdot R_{x.T_1,...,T_m} \cdot R_{y.T_1,...,T_m}$$

Oder: $r_{xy.T_1,...,T_m} = \frac{r_{xy} - r_{\hat{x}(T_1,...,T_m),\hat{y}(T_1,...,T_m)} \cdot R_{x.T_1,...,T_m} \cdot R_{y.T_1,...,T_m}}{\sqrt{1-R^2_{x.T_1,...,T_m}} \cdot \sqrt{1-R^2_{y.T_1,...,T_m}}}$

Oder: $r_{\hat{x}(T_1,...,T_m),\hat{y}(T_1,...,T_m)} = \frac{r_{xy} - r_{xy.T_1,...,T_m} \sqrt{1-R^2_{x.T_1,...,T_m}} \sqrt{1-R^2_{y.T_1,...,T_m}}}{R_{x.T_1,...,T_m} \cdot R_{y.T_1,...,T_m}}$

(Speziell für m = 1:

$$r_{xy} = r_{xy.T} \sqrt{1-r^2_{xT}} \sqrt{1-r^2_{yT}} + r_{xT} \cdot r_{yT}$$

Oder: $r_{xy.T} = \frac{r_{xy} - r_{xT}\, r_{yT}}{\sqrt{1-r^2_{xT}} \sqrt{1-r^2_{yT}}}$)

Die allgemeine Korrelationszerlegung nach einer nominalen Variablen A mit m Ausprägungen ist ein Spezialfall hiervon für die Dichotomien, die m - 1 Ausprägungen repräsentieren:

$$r_{xy} = \sqrt{1-\eta^2_{xA}} \sqrt{\eta^2_{yA}}\; r_I + \eta_{xA}\, \eta_{yA}\, r_z$$

Beweis:

Bei der Darstellung der Varianzanalyse als Spezialfall der multiplen Regression war gezeigt worden: $\eta^2_{xA} = R^2_{x.T_1,...,T_m}$ (Für y gilt Entsprechendes.)

Ferner wurde gezeigt: $s_{x-\hat{x},y-\hat{y}} = \sum_{k=1}^{m} p_k \; s_k(x,y)$

Speziell für x = y: $s^2_{x-\hat{x}} = \sum_{k=1}^{m} p_k s_k(x,x) = \sum_{k=1}^{m}\sum_{j=1}^{n_k} \frac{(x_{kj} - \bar{x}_k)}{n}$

Entsprechendes gilt für $s^2_{y-\hat{y}}$.

Insgesamt: $\frac{s_{x-\hat{x},y-\hat{y}}}{s_{x-\hat{x}} \cdot s_{y-\hat{y}}} = r_I$

Mittelwertberechnung:

$$\sum_i \sum_j \frac{\hat{y}(i,j)}{n} = \sum_{k=1}^{m} \bar{y}_k \sum_i \sum_j \frac{1_{A_k(i,j)}}{n} = \sum_{k=1}^{m} \frac{n_k \bar{y}_k}{n} = \bar{y}$$

Für x entsprechend.

Also: $s_{\hat{x},\hat{y}} = \frac{< \sum_{k=1}^{m} (\bar{x}_k - \bar{x}) 1_{A_k}, \sum_{l=1}^{m} (\bar{y}_l - \bar{y}) 1_{A_l} >}{n}$

$$= \sum_{k=1}^{m} (\bar{x}_k - \bar{x})(\bar{y}_k - \bar{y}) \sum_i \sum_j \frac{1_{A_k}(i,j)}{n}$$

$$= \sum_{k=1}^{m} p_k (\bar{x}_k - \bar{x})(\bar{y}_k - \bar{y})$$

Speziell für x = y: $s_{\hat{x},\hat{x}} = \sum_{k=1}^{m} p_k (\bar{x}_k - \bar{x})^2$

Insgesamt: $r_{\hat{x},\hat{y}} = \frac{s_{\hat{x},\hat{y}}}{s_{\hat{x}} s_{\hat{y}}} = r_z$

Rechnet man den Einfluss der Kontrollvariablen sowohl aus y als auch aus x heraus und führt mit der bereinigten Variablen $y - \hat{y}$ (T_1, ..., T_m) eine Regression auf die bereinigte Variable $x - \hat{x}$ (T_1, ..., T_m) durch, so lässt sich der Zusammenhang des Beta-Koeffizienten $\beta_{yx.T_1,...,T_m}$ dieser Regression mit dem Beta-Koeffizienten β_{yx} der üblichen einfachen Regression von y auf x angeben.

Aus der allgemeinen Kovarianzzerlegung nach Kontrollvariablen folgt:

$$\beta_{yx} = \frac{s_{yx}}{s_x^2} = \frac{s_{xx.T_1,\ldots,T_m}}{s_x^2}\frac{s_{yx.T_1,\ldots,T_m}}{s_{xx.T_1,\ldots,T_m}} + \frac{s^2_{\hat{x}(T_1,\ldots,T_m)}}{s_x^2}\frac{s_{\hat{y}(T_1,\ldots,T_m),\hat{x}(T_1,\ldots,T_m)}}{s^2_{\hat{x}(T_1,\ldots,T_m)}}$$

$$= \left(1 - R^2_{x.T_1,\ldots,T_m}\right)\beta_{yx.T_1,\ldots,T_m} + R^2_{x.T_1,\ldots,T_m}\beta_{\hat{y}(T_1,\ldots,T_m),\hat{x}(T_1,\ldots,T_m)}$$

Hierbei ist $\beta_{\hat{y},\hat{x}}$ der Beta-Koeffizient der einfachen Regression der Variablen $\hat{y}$ auf die Variable $\hat{x}$.
Für den Spezialfall, dass es eine kategoriale Kontrollvariable A mit m Ausprägungen gibt, erhält man unter Verwendung von m - 1 unabhängigen Indikatorfunktionen:

$$\beta_{yx} = \left(1 - \eta^2_{xA}\right)\beta_{yx.A} + \eta^2_{xA}\,\beta_{\hat{y}(A),\hat{x}(A)}\text{, wobei: } \beta_{yx.A} = \frac{\sum_{i=1}^{m} n_i\, s_{i(y,x)}}{\sum_{i=1}^{m} n_i\, s_{i(x,x)}} \text{ und}$$

$$\beta_{\hat{y}(A),\hat{x}(A)} = \frac{\sum_{i=1}^{m} n_i\left(\bar{y}_i - \bar{y}\right)\left(\bar{x}_i - \bar{x}\right)}{\sum_{i=1}^{m} n_i\left(\bar{x}_i - \bar{x}\right)}$$

Für eine Gebietsvariable A lässt sich dies wie folgt anwenden: Berechnet man eine Regression von y auf x aufgrund von Aggregatdaten, so ist $\beta_{\hat{y}(A),\hat{x}(A)}$ der zugehörige Beta-Koeffizient. Die abgeleitete Gleichung drückt also den Zusammenhang des Beta-Koeffizienten der Regression von y auf x bei der Verwendung von Aggregatdaten mit dem Beta-Koeffizienten der üblichen einfachen Regression von y auf x aus.
Erklären die Gebietsunterschiede die Variation von x völlig (d.h. η^2_{xA}= 1), so stimmen der Regressionskoeffizient aufgrund von Aggregatdaten und der übliche Regressionskoeffizient überein. Erklären die Gebietsunterschiede gar keinen Anteil der Variation von x (d.h. η^2_{xA} = 0), so ist der übliche Beta-Koeffizient identisch mit dem Beta-Koeffizienten innerhalb der Gruppen $\left(\beta_{yx.A}\right)$.

4.9 Kontrastgruppenanalyse (tree analysis)

Dieses Verfahren zur Untersuchung von Abhängigkeiten wurde von Sonquist und Morgan entwickelt. (Einen Vergleich der "tree analysis" mit der "multiple classification analysis" zur Analyse von statistischen Interaktionen findet man in Sonquist (1975).)
Die "tree analysis" besteht in der schrittweisen Anwendung der einfachen Varianzanalyse jeweils auf dichotomisierte Kategorien. Wenn man eine abhängige metrische Variable y durch eine Dichotomie statistisch erklären will, so kann man von der folgenden Streuungszerlegung ausgehen:

$$\begin{array}{ccccc} \mathrm{SS_y} & = & \mathrm{SS_{Fehler}} & + & \mathrm{SS_{erklärt}} \\ \sum_{i=1}^{2}\sum_{j=1}^{n}\left(y_{ij} - \bar{y}\right)^2 & = & \sum_{i=1}^{2}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}_i\right)^2 & + & \sum_{i=1}^{2} n_i\left(\bar{y}_i - \bar{y}\right)^2 \end{array}$$

In der Varianzanalyse war für diesen Fall k = 2 gezeigt worden, dass:

$$SS_{erklärt} = \frac{1}{\frac{1}{n_1} + \frac{1}{n_2}} \left(\bar{y}_1 - \bar{y}_2\right)^2$$

Hat man also verschiedene Dichotomien zur Auswahl, so erhält man die beste Erklärung durch diejenige Dichotomie, nach der die abhängige Variable in die heterogensten beiden Untergruppen zerfällt. Um die Erklärung zu maximieren, muss die Differenz $\left|\bar{y}_1 - \bar{y}_2\right|$ maximiert werden. Dies erfolgt durch Auswahl der dazu am besten geeigneten Dichotomie.[15] Für Anteilswerte bedeutet dies, $\left|p_1 - p_2\right|$ zu maximieren.[16] Analog dazu wird die Varianz innerhalb der Gruppen minimiert, was bezüglich der abhängigen Variable zu homogeneren Untergruppen führt.

Im nächsten Schritt wird versucht, entsprechend $\sum_{j=1}^{n_1}\left(y_{1j} - \bar{y}_1\right)^2$ und $\sum_{j=1}^{n_2}\left(y_{2j} - \bar{y}_2\right)$, d.h. die noch nicht erklärte Variation, optimal zu erklären durch die jeweils geeigneten Dichotomien usw.

Will man also eine abhängige metrische Variable y allgemeiner durch nominale Variablen erklären, so kann man jeweils dichotomisierte Ausprägungen bilden.[17] Es wird dann wieder jeweils diejenige Dichotomie von Ausprägungen ausgewählt, mit der die zugehörigen Variablen y am stärksten unterschieden werden können (deshalb der Name: Kontrastgruppenanalyse). Wendet man die Auswahl von Dichotomien, die bezüglich der abhängigen Variablen am stärksten diskriminieren, sukzessive an, so entsteht durch diese Verzweigungen eine „Baum"-Darstellung („tree analysis", wobei der Baum „auf dem Kopf" steht).

Man braucht „Stop-Regeln" zur Beendung der Verzweigungen, wie z.B.:
- eine Untergrenze für die Anzahl der Einheiten einer Zelle,
- eine Obergrenze für den Erklärungszuwachs durch eine weitere Dichotomie - beispielsweise wenn die Gruppe selbst schon relativ homogen ist (die Untergruppen sollten i.d.R. mehr als 1-2 % der Gesamtvarianz erklären),
- eine Obergrenze für die Gesamtzahl der Zellen, um nicht unübersichtlich viele Typen zu bekommen.

Das folgende Beispiel stammt von Liepelt. Hier wird das dichotome Merkmal SPD-Wahl (ja/nein) durch entsprechende Untergruppen erklärt. Die Bezugsgröße ist der Anteil der SPD-Wähler in der Gesamtheit (p = 45 %).

---

15 $SS_{erklärt} = \sum n_i \, \bar{y}_i^2 - n \, \bar{y}^2$. Formal äquivalent ist es also, $n_1 \, \bar{y}_1^2 + n_2 \, \bar{y}_2^2$ zu maximieren, da $\bar{y}$ bei dieser Auswahl eine Konstante ist.

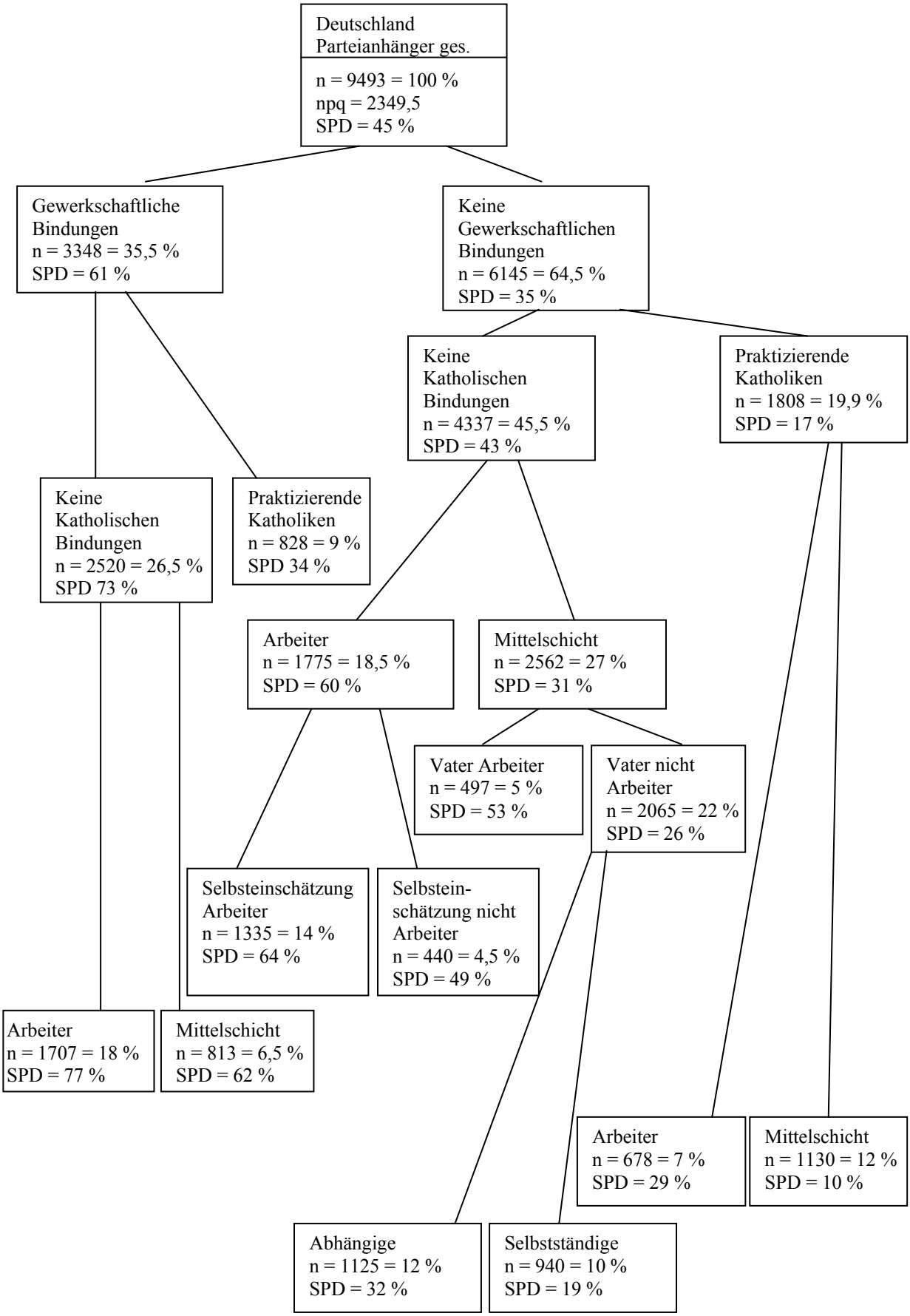
16 Die Varianzzerlegung für Anteilswerte lautet entsprechend: $np(1-p) = \sum_{i=1}^{2} n_i\left[p_i(1-p_i)\right] + \sum_{i=1}^{2} n_i\left(p_i - p\right)$

Hierbei ist p der Anteilswert der Untersuchungseinheiten, die eine positive Ausprägung auf der abhängigen Variable besitzen. Entsprechend ist p_i der Anteilswert in der jeweiligen Kontrastgruppe.

17 Bei einer nominalen Variablen mit mehr als zwei Ausprägungen braucht man nicht alle Möglichkeiten zu berücksichtigen, sondern kann sich nach Ericson (in: Sonquist und Morgan 1964) auf folgendes beschränken: Die Kategorien A_1, ..., A_k seien so durchnummeriert, dass $\bar{y}_{A_1} \le \bar{y}_{A_2} \le \ldots \le \bar{y}_{A_k}$. Dann braucht man nur die Alternativen zu berücksichtigen:

A_1 versus Rest (= A_2, ..., A_k)
A_1, A_2 versus Rest (= A_3, ..., A_k)
⋮
A_1, ..., A_{k-1} versus Rest (= A_k)

Abbildung 4-6: SPD-Wählertypologie Westdeutschlands nach Liepelt (1968)



Gesamtanteil der erklärten Varianz:[18]
6,4 % + 4,7 % + 5,0 % + 0,6 % + 4,1 % + 0,6 % + 0,7 % + 1,2 % + 0,4 % = 23,7 %
SS = npq (weil s_y^2 = p (1 - p) für Anteilswerte).

Das Beispiel von Liepelt (1968) führt zu folgender Typologie der SPD-Wähler (+ = ja, - = nein):

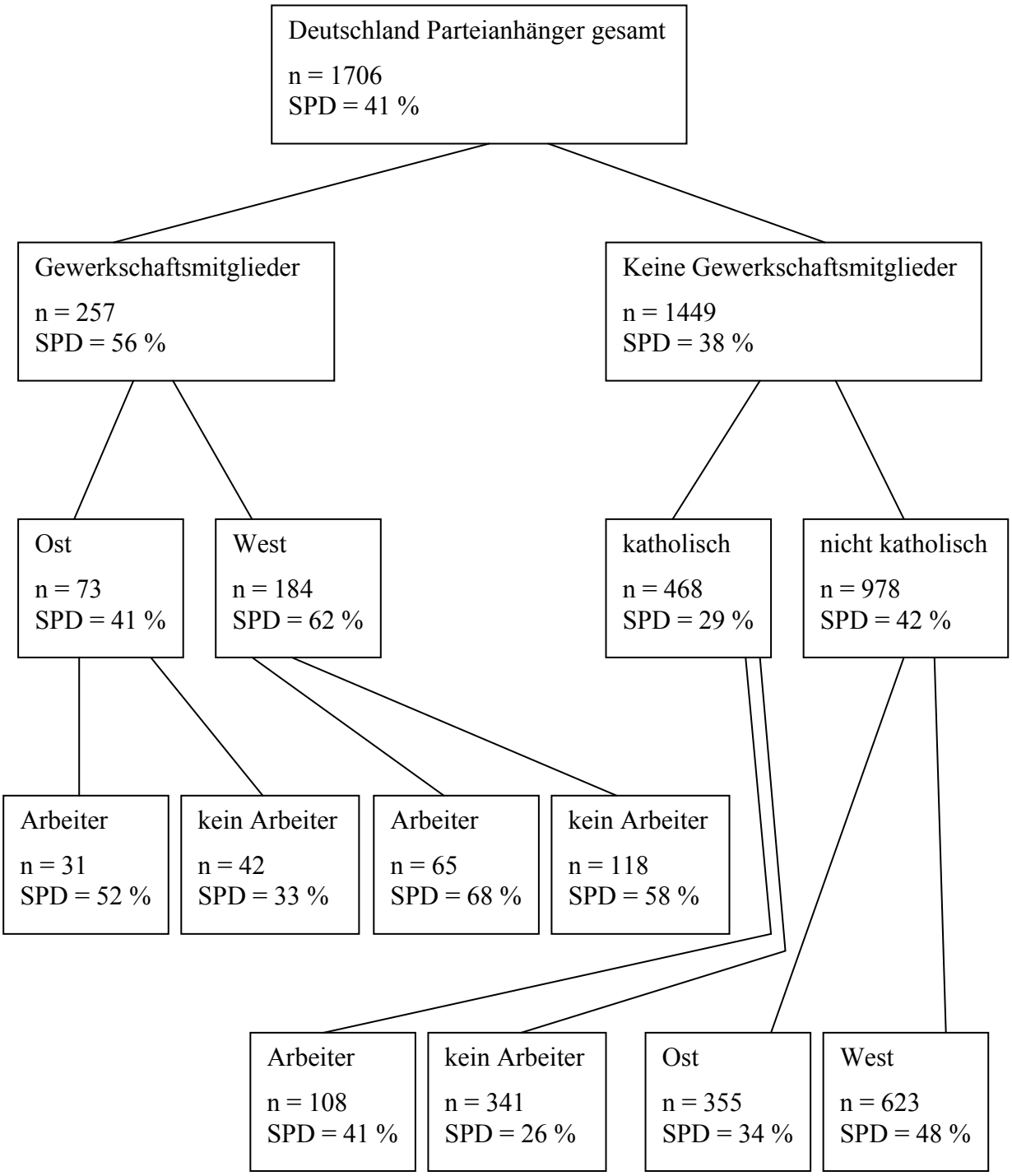
SPD-Anteil	Gewerkschaftliche Bindung	Katholische Bindung	Arbeiter	Selbsteinschätzung Arbeiter	Vater Arbeiter	abhängig beschäftigt	Anzahl der berücksichtigten unabhäng. Variablen
77 %	+	-	+				3
64 %	-	-	+	+			4
62 %	+	-	-				3
53 %	-	-	-		+		4
49 %	-	-	+	-			4
34 %	+	+					2
32 %	-	-	-		-	+	5
29 %	-	+	+				3
19 %	-	-	-		-	-	5
10 %	-	+	-				3

Diese Typologie ist insofern nicht befriedigend, als jeweils unterschiedlich viele unabhängige Variablen zur Charakterisierung verwendet werden. Würde man alle möglichen Typen berücksichtigen (wobei jeweils alle 6 unabhängige Variablen verwendet werden), so hätte man 2^6 = 64 Typen, was ebenfalls völlig unübersichtlich wäre.

Wenn man die aktuelle SPD-Wählertypologie für Gesamtdeutschland aufgrund des Allbus 2002 vergleicht, so ist die Gewerkschaftsmitgliedschaft noch immer der trennschärfste Indikator (vgl. Abbildung 4-7). Allerdings haben die Gewerkschaftsmitglieder im Westen eine deutlich stärkere SPD-Präferenz als im Osten. Unter den Nicht-Mitgliedern ist die katholische Konfession die stärkste Trennlinie. Schließlich haben in den resultierenden Teilgruppen Arbeiter jeweils eine stärkere SPD-Präferenz als die übrigen Befragten.

[18] Dies ist gleich der Varianzreduktion: $\frac{\frac{SS_{erklärt}}{n}}{\frac{SS_y}{n}} = \frac{\frac{\left(SS_y - \left(SS_y - SS_{erklärt}\right)\right)}{n}}{\frac{SS_y}{n}}$

Abbildung 4-7: SPD-Wählertypologie Deutschland 2002



Datenquelle: ALLBUS 2002; pairwise deletion of missing values.

Sonquist und Morgan sehen die Hauptleistung ihres Verfahrens in der Behandlung von Interaktionseffekten, woher auch ihre Bezeichnung als „automatic interaction detection technique“ (AID) rührt. Dass dieses Verfahren Vorteile gegenüber der herkömmlichen multiplen Regression auf Dichotomien hat, scheint mir zweifelhaft. Mehr als eine anschauliche Darstellung liefert das Verfahren nicht.

Bei der Regression auf Dichotomien können Interaktionen wie üblich berücksichtigt werden:
$y = a + b_1 x_1 + b_2 x_2 + b_3 x_1 \cdot x_2$
(Im Falle von 2 Prädiktoren, etc.)

Zum Vergleich der „tree analysis“ mit der schrittweisen Regression im Falle von Dichotomien als unabhängigen Variablen lässt sich folgendes festhalten:

Da die Varianzanalyse nur ein Spezialfall der multiplen Regression und die „tree analysis“ in jedem Einzelschritt ein Spezialfall der Varianzanalyse ist, würde in einer schrittweisen Regression ebenfalls in dem angeführten Beispiel der SPD-Wahl zunächst die Variable „Gewerkschaftliche Bindungen - ja/nein“ ausgewählt, da diese Regression am meisten erklärt.
SPD = f (Gewerkschaft)

In Kombination mit dieser Variablen möge die Variable „Katholische Bindungen - ja/nein“ am meisten der verbleibenden Varianz erklären.
SPD = f (Gewerkschaft, katholisch)

In Kombination mit diesen beiden Variablen möge die Variable „Arbeiter - ja/nein“ am meisten der verbleibenden Varianz erklären.
SPD = f (Gewerkschaft, katholisch, Arbeiter)

Insgesamt: Im Gegensatz zur schrittweisen Regression wendet die „tree analysis“ auf jede Ausprägung der Dichotomie die Maximierungsregel der Varianzklärung separat und unabhängig an. So kann z.B. auftreten, dass bei „keine gewerkschaftlichen und keine katholischen Bindungen“ für Arbeiter als nächste Variable die Selbsteinschätzung berücksichtigt wird und für Nicht-Arbeiter im nächsten Schritt, ob der Vater Arbeiter ist oder nicht. In der multiplen Regression kann jeweils nur eine Variable die nächste für beide Ausprägungen sein.[19] In der schrittweisen Regression erhält man also im Unterschied zur „tree analysis“ einen symmetrischen „Baum“ und entsprechende Typen. Die schrittweise Regression liefert im Falle von Dichotomien als unabhängigen Variablen eine symmetrische „tree analysis“. Die „tree analysis“ liefert also über die schrittweise Regression hinaus nur einige asymmetrische Zusammenhänge.

Das Konzept der tree analysis lässt sich subsumieren unter eine spezifische Art der Regression, die man „bedingte“ Regression nennen könnte:

Hat eine Dichotomie X_1 die Ausprägungen a und b, eine Dichotomie X_2 die Ausprägungen c und d, so ist die Regression von y auf X_1 und X_2 abhängig von 4 Ausprägungskombinationen:
f (a, c), f (a, d), f (b, c), f (b, d)

Die Regression von y auf X_2 unter der Bedingung X_1 = a ist eine Funktion der Art f (a, X_2), ausführlicher also:
f (a, c), f (a, d)

[19] Die „tree analysis“ kann also insgesamt mehr Varianz erklären, da sie sich den Daten bei den Verzweigungen einer Ebene unabhängig voneinander anpassen lässt.

In der tree analysis werden solche bedingten Regressionen verwendet. Eine bedingte Regression ist quasi nur eine Seite der gesamten Regression.

In der tree analysis ist es möglich, dass unter der Bedingung $X_1 = b$ die Regression von y auf eine andere Dichotomie X_3 (X_2) betrachtet wird. Solche Asymmetrien in dem Baumdiagramm spiegeln bedingte Zusammenhänge wider.

Die Regression auf die Interaktion $X_1 \cdot X_2$ ist sowohl verschieden von dem additiven Regressionsansatz als von der bedingten Regression, man benötigt die Funktionswerte der multiplikativ verknüpften Kombinationen:
$f(a \cdot c), f(a \cdot d), f(b \cdot c), f(b \cdot d)$

Das Konzept der Interaktion in der Regression ist also verschieden von dem Konzept der bedingten Zusammenhänge in der tree analysis.

Eine adäquate Behandlung der Interaktionen dürfte die folgende sein: Zunächst werden nur solche Prädiktoren beibehalten, die einen signifikanten Beitrag leisten; von den zweifachen Interaktionen dieser Prädiktoren werden solche berücksichtigt, die einen signifikanten zusätzlichen Effekt haben; entsprechend für die dreifachen Interaktionen etc.

Ergebnis der schrittweisen Regression im Falle von Dichotomien als unabhängigen Variablen[20]:

Abbildung 4-8: Baumdarstellung

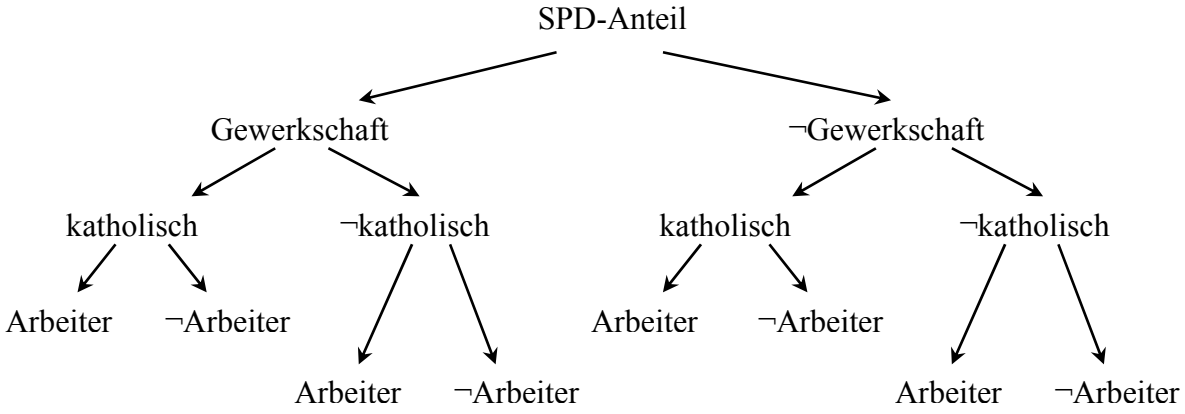


Tabelle 4-6: Typen aufgrund der drei Prädiktoren

$\bar{y}$	Gewerkschaft	katholisch	Arbeiter	n
	+	+	+	
	+	+	-	
	+	-	+	
	+	-	-	
	-	+	+	
	-	+	-	
	-	-	+	
	-	-	-	

[20] Hier sind nur die wichtigsten 3 unabhängigen Variablen berücksichtigt, die Darstellung lässt sich natürlich leicht auf k unabhängige Variablen verallgemeinern.

Diese unterschiedliche Darstellung einer Variablen y in Abhängigkeit von k unabhängigen Variablen x_1, ..., x_k durch die 2^k (im Beispiel: $2^3 = 8$) möglichen Variationen in Form eines Baumes oder einer Tabelle aller Typen gilt allgemein. Die letzte Zeile des Baumes ist identisch mit der Spalte der abhängigen Variablen in der Tabelle (vgl. auch Harder 1974: 177). Der Baum enthält also noch zusätzliche Informationen.

4.10 Anwendungsbeispiel zur Varianzanalyse: Vergleich der Erklärungskraft verschiedener Berufsstruktur- und Klassenmodelle für die Bundesrepublik Deutschland

Im Rahmen des von Erik Olin Wright (University of Wisconsin, Madison) initiierten Forschungsverbundes „Comparative project on class structure and class consciousness“ wurde an der Universität-Gesamthochschule-Duisburg ein von der Deutschen Forschungsgemeinschaft finanziertes Projekt zur Klassenstruktur und zum Klassenbewusstsein in der Bundesrepublik auf der Basis einer repräsentativen Befragung von 1815 deutschen Erwerbstätigen durchgeführt. Aus diesem Projektkontext soll im Folgenden mein Versuch dargestellt werden, die für die Bundesrepublik vorliegenden Berufsstruktur- und Klassenmodelle im Hinblick auf ihre Erklärungskraft zu vergleichen. Mein Hauptgesichtspunkt bei diesem Vergleich besteht darin, dass man sich im Unterschied zu bisherigen Berufsstruktur- und Klassenuntersuchungen nicht mit der Auszählung der theoretisch begründeten Modelle begnügen, sondern vorher eindeutige Qualitätskriterien formulieren sollte. Als Kriterien werden im Folgenden vorgeschlagen, dass die Schicht- oder Klassenlagen der Modelle homogene soziale Lagen sein müssen (als bester Indikator erwies sich das Einkommen), die gesellschaftliche Folgen erwarten lassen (als Indikator für wahrscheinliches zukünftiges Handeln wird ein Bewusstseins-Index verwendet). Mit diesen Kriterien wird ein Vergleich der Erklärungsleistung der verschiedenen Berufsstruktur- und Klassenmodelle für die Bundesrepublik durchgeführt.

Berücksichtigt wurden die umfangreichen Arbeiten des Instituts für Marxistische Studien und Forschungen (IMSF) und des Projekts Klassenanalyse (PKA) als marxistische Klassenmodelle der siebziger Jahre für die Bundesrepublik[21]. Dem wurde das in Weber'scher Tradition stehende Klassenmodell Walter Müllers gegenübergestellt. Als weitere Modelle wurden Wrights altes und neues Klassenmodell berücksichtigt. Ferner wird ein Berufsstrukturmodell auf Basis der bundesdeutschen Sozialstatistik zum Vergleich herangezogen, das sich nach meiner Auffassung empirisch als eine bessere Bündelung sozialer Lagen in der Bundesrepublik erweist als alle vorher genannten Klassenmodelle.

4.10.1 Probleme des Modellvergleichs und Kriterien zur Beurteilung der Erklärungskraft

Die verschiedenen in den Vergleich einbezogenen Ansätze haben unterschiedliche Grundannahmen und Ansprüche, was jedoch einen empirischen Vergleich nicht unmöglich macht. Zentral ist nach meiner Auffassung die Überprüfung der Homogenität der jeweiligen Klassenlagen der verschiedenen Modelle, denn bei homogenen Gruppen lässt sich am ehesten Klassenhandeln durch kollektive Akteure erwarten. Dabei bezieht sich die Homogenität zunächst auf die materielle Lage, welche die Basis gemeinsamer Interessen bilden kann. Da das Ziel von Klassenanalysen die Erklärung von Statik und Dynamik der Sozialstruktur ist, spielt ferner das Bewusstsein der potentiellen Akteure eine zentrale Rolle, weshalb das Bewusstsein neben der materiellen Lage die zweite Hauptdimension ist, bezüglich der die Homogenität von Gruppen untersucht wird.

[21] Da diese beiden Modelle sich empirisch gar nicht bewährten, werden sie in der folgenden Darstellung nicht berücksichtigt.

Die Homogenität wird durch den „absoluten“ bzw. „relativen“ Erklärungsanteil Eta^2 bzw. Eta^2/k gemessen, womit die nominale Information der Verschiedenheit von Gruppierungen erfasst wird, aber keine ordinale Information. Aussagen zu letzterem werden aber im Folgenden in einem graphischen Bezugsrahmen überprüft.

4.10.1.1 Indikatoren für die Hierarchie der materiellen Lage

Als Kriterien für die Hierarchie der materiellen Lage werden in dem Vergleich verwendet: Schicht-Selbsteinstufung und Oben-Unten-Skala, Berufsprestige-Skala sowie das Netto-Arbeitseinkommen. Mit dem Argument, dass Schichten als Feingliederungen von Klassen betrachtet werden können, wollte ich die Schicht-Selbsteinstufung als zentralen Indikator verwenden. Es stellte sich aber heraus, dass das Einkommen sich viel besser durch die verschiedenen Berufsstruktur- und Klassenmodelle diskriminieren lässt als die Schicht-Selbsteinstufung. Das Berufs-Prestige erweist sich nach dem Erklärungsanteil (Eta^2) sogar als noch besserer Indikator, aber das Berufs-Prestige ist nicht die zentrale Zielvariable einer Klassenanalyse. Zwar basieren viele Klassenmodelle in der Operationalisierung auf Berufsvariablen, aber als Zielgröße einer Klassenanalyse wurde das Einkommen als überzeugendster Indikator zur Erfassung der Hierarchie der materiellen Lage ausgewählt, einerseits weil die Einkommenshierarchie sehr erfolgreich für die Handlungschancen ist, andererseits weil Einkommen ein eher „objektiver“ Maßstab ist, auch wenn er „subjektiv“ erfragt wird.

4.10.1.2 Indikatoren für den ideologischen Standort (Bewusstsein)

Als Haupt-Kriterium zur Erfassung des ideologischen Standorts würde sich zunächst die Links-Rechts-Dimension anbieten, die sich bei Untersuchungen der politischen Landschaft als wichtigste Dimension erweist. Es zeigt sich aber, dass dieser Indikator tatsächlich genau zur Parteienlandschaft passt, jedoch nicht so gut zu den Berufsstruktur- und Klassenmodellen. Wright hatte in seinen Analysen mit einem Bewusstseins-Index gearbeitet, und die deutsche Modifikation dieses Index[22] erweist sich als recht geeignet zur Beurteilung von Berufsstruktur- und Klassenmodellen. Da ein Index als zusammengesetzte Messung weniger überschaubar ist - er erfasst die Polarität Arbeit versus Kapital anhand von vier Items - werden ferner die Werte für das beste einzelne Item angegeben („Eigentümer haben Vorteile“). Allerdings ist die Diskriminanzkraft des Bewusstseins-Index deutlich höher als die der einzelnen Items. Deshalb wird der Bewusstseins-Index als zentraler Indikator zur Erfassung des ideologischen Standorts (Bewusstseins) verwendet.

[22] Analog zu Wright wurde der Bewusstseins-Index als einfacher additiver Durchschnittsindex über die folgenden Indikatoren gebildet: 1. „In Unternehmen haben Eigentümer Vorteile auf Kosten der Arbeitnehmer und Konsumenten.“ 2. „Im Falle eines Streiks sollte das Management gesetzlich daran gehindert werden, anstelle der Streikenden andere Arbeitnehmer einzustellen.“ 3. „Wenn die Arbeitnehmer in ihrem Betrieb die Chance hätten, ohne das Management zu arbeiten, dann könnten sie alle Angelegenheiten wirksam genauso gut erledigen.“ 4. „Arbeitnehmer in unserer Gesellschaft brauchen Gewerkschaften, um ihre Interessen durchzusetzen“.

4.10.2 Vergleich der Erklärungskraft der verschiedenen Berufsstruktur- und Klassenmodelle

In der folgenden Tabelle sind die Erklärungsanteile (Eta^2) der verschiedenen Berufsstruktur- und Klassenmodelle nach den berücksichtigten Kriterien zusammengestellt.

(1) Die Einkommensunterschiede lassen sich am besten erklären durch das Berufsstrukturmodell auf Basis der bundesdeutschen Sozialstatistik (NV 17). Das Einkommen wird durch dieses Modell zu 39,3 % erklärt, dies ist sehr viel für eine einzelne Variable.

Walter Müllers Ansatz, der damit am ehesten verwandt ist, schneidet mit 31,9 % recht gut ab. Wrights Modelle schneiden erst besser ab, wenn man die Modifikation (von Klasse 1/2 zu MODKL 1/2) berücksichtigt. Wrights altes Klassenmodell (MODKL 1) passt weniger gut zur Empirie. Sein neues Klassenmodell (MODKL 2) erreicht aber einen Erklärungsanteil von 32,6 %.

Das Berufsstrukturmodell auf Basis der bundesdeutschen Sozialstatistik (NV 17) erweist sich insgesamt eindeutig als beste Bündelung von sozialen Lagen in der Bundesrepublik nach dem Kriterium der Hierarchie der materiellen Lagen.

(2) Nach dem Kriterium des Bewusstseins-Index erweist sich Wrights neues Klassenmodell als das beste; mit der Modifikation (MODKL 2) wird ein Erklärungsanteil von 17,5 % erreicht. Allerdings ist die modifizierte Berufsvariable (NV 17) mit 17,1 % nicht viel schlechter. Auch die übrigen Berufsstruktur- und Klassenmodelle schneiden nach diesem Kriterium im Vergleich ähnlich ab.

(3) Es zeigt sich, dass man mit Berufsstruktur- und Klassenmodellen viel eher die Hierarchie der materiellen Lagen erklären kann (39,3 % als höchster Erklärungsanteil für das Einkommen durch die modifizierte Berufsvariable) als das Bewusstsein (17,1 % für die modifizierte Berufsvariable bzw. 17,5 % für MODKL 2). Der Anspruch der Klassenanalyse, die Verankerung von Interessen in der materiellen Lage herauszuarbeiten, um dadurch kollektive Akteure, Koalitionen und Allianzen des Klassenhandelns prognostizieren zu können, lässt sich weniger gut einlösen als eine Beschreibung von Bündelungen sozialer Lagen, die die Sozialstruktur einer Gesellschaft charakterisieren. Es sollte nicht überraschen, dass „das Bewusstsein“ sich nur zum Teil aus der Klasse oder der Stellung im Beruf ableiten lässt, „das Sein“ umfasst ja eine Vielzahl weiterer Faktoren wie Geschlechterrolle, Haushaltskontext etc. Dieser Anteil ist allerdings groß genug, um das ganze Programm, das Handeln zum Teil aus der materiellen Interessenlage abzuleiten, nicht als unfruchtbar bezeichnen zu müssen.

(4) Nach meinen beiden Hauptkriterien würde es sich anbieten, eine Gruppierung nur dann einem einheitlichen Typus von potentiellen Akteuren zuzuordnen, wenn die Gruppierung bezüglich Einkommen und Bewusstseins-Index relativ homogen ist.

Tabelle 4-7: Kriterienvariablen nach Klassenlagen für verschiedene Klassen- und Berufsstrukturmodelle

Kriterien für die Hierarchie der materiellen Lage (Eta^2)

	V 17: Berufliche Stellung k = 24	NV 17: Modifizierte Berufsvariable k = 23	RED-NV 17 k = 12	GROB-NV 17 k = 8	REST-NV 17 k = 5	M-KLAS-SE k = 17	KLAS-SE 1 k = 8	KLAS-SE 2 k = 12	MOD-KL 1 k = 11	MOD-KL 2 k = 15
V801: Prestige	50, 1 %	46,0 %	40,4 %	39,0 %	36,1 %	48,7 %	14,9 %	48,1 %	15,3 %	45.6 %
V480: Schicht	25,7 %	23,3 %	19,8 %	18,0 %	17,4 %	21,6 %	11,3 %	17,6 %	13,4 %	17,6 %
V488: Oben-Unten-Skala	19,0 %	18,3 %	15,6 %	14,6 %	14,2 %	17,5 %	9,1 %	13,5 %	12,4 %	15,1 %
EINB: Einkommen	34,9 %	39,3 %	35,1 %	31,0 %	28,0 %	31,9 %	18,7 %	23,7 %	27,7 %	32,6 %
(Einkommen relativ)	1,454	1,709	2,925	3,875	5,600	1,876	2,338	1,975	2,518	2,173

Legende:

NV 17: Als gesonderte Gruppierungen: Arbeitslos; in Ausbildung; Mithelfende.
REDNV 17: Zusammengefasst: Freiberufler; größere Selbständige; Angestellte und Beamte in vier Stufen zusammengefasst.
GROBNV 17: Beamte, Angestellte und Arbeiter in drei Stufen zusammengefasst.
RestNV 17: Alle Selbständigen und Mithelfenden zusammengefasst.
MKLASSE: Klassenmodell nach Walter Müller (1977).
KLASSE 1: Wrights erstes Klassenmodell.
KLASSE 2: Wrights zweites Klassenmodell.
MODKL 1: Wrights erstes Klassenmodell, wobei arbeitslos, in Ausbildung, Mithelfende gesonderte Gruppierungen bilden.
MODKL 2: Wrights zweites Klassenmodell, wobei arbeitslos, in Ausbildung, Mithelfende gesonderte Gruppierungen bilden.

Fortsetzung von Tabelle 4-7: Kriterien für Bewusstsein (Eta^2)

	V 17: Berufliche Stellung k = 25	NV 17: Modifizierte Berufsvariable k = 23	RED-NV 17 k = 12	GROB-NV 17 k = 8	REST-NV 17 k = 5	M-KLAS-SE k = 17	KLAS-SE 1 k = 8	KLAS-SE 2 k = 12	MOD-KL 1 k = 11	MOD-KL 2 k = 15
INDEX	17, 1 %	17,1 %	16,2 %	15,0 %	14,9 %	15,9 %	15,8 %	16,6 %	16,9 %	17,5 %
(INDEX relativ)	0,684	0,770	1,350	1,875	2,980	0,935	1,975	1,383	1,536	1,1671
V305: Eigentümer Vorteile	11,6 %	12,0 %	11,4 %	10,8 %	10,6 %	11,4 %	10,7 %	11,1 %	11,8 %	12,2 %
V489: Links-Rechts	5,5 %	6,7 %	5,0 %	4,5 %	4,3 %	5,5 %	4,0 %	4,2 %	5,7 %	6,0 %

4.10.3 Graphische Darstellung der verschiedenen Berufsstruktur- und Klassenmodelle

Vor dem Vergleich hatte ich keine Präferenz für eines der Berufsstruktur- oder Klassenmodelle. Da sich im empirischen Vergleich das Berufsstrukturmodell auf Basis der bundesdeutschen Sozialstatistik (NV 17) als am besten geeignet erwies, homogene soziale Lagen in der Bundesrepublik zu bündeln, habe ich die Darstellungsweise gewählt, dass hier vertiefend dieses nach den vorher gewählten Kriterien beste Modell diskutiert wird.

In den folgenden Graphiken habe ich die beiden orthogonalen Koordinatenachsen nicht wie in der Faktorenanalyse und multidimensionalen Skalierung im Nachhinein inhaltlich bestimmt, sondern vorgegeben, nämlich als horizontale Achse den Bewusstseins-Index und als vertikale Achse das Einkommen. Als Nachteil der gewählten Vorgehensweise könnte man ansehen, dass die beiden inhaltlich vorgegebenen Bezugsachsen orthogonal dargestellt werden, obwohl sie empirisch leicht korrelieren (r = 0,25). Mir scheinen jedoch die Vorzüge zu überwiegen: Auf diese Weise wird es möglich, die Bewährung der Modelle nach den beiden Kriterien, die sich in der bisherigen Analyse als am wichtigsten herausgestellt haben, anhand der einzelnen Klassenlagen in diesem Bezugssystem graphisch zu illustrieren. Diese Darstellungsart erlaubt es, die Konfigurationen oder Strukturen der verschiedenen Klassenmodelle prägnant zu veranschaulichen. (Als ein Bezugspunkt ist in den Graphiken jeweils der empirische Schwerpunkt bzw. Mittelwert der Verteilungen nach den beiden Kriterien angegeben, nämlich $(\bar{x}, \bar{y})$, wobei x der Bewusstseins-Index ist und y das Einkommen.)

4.10.4 Berufsstrukturmodell auf Basis der bundesdeutschen Sozialstatistik nach Einkommen und Bewusstseins-Index

Die Berufsstrukturvariablen V 17 („Stellung im Beruf“) bzw. modifiziert NV 17 sind keine rein technischen Berufsvariablen, insofern lässt sich Wrights Gegenüberstellung von Klassen als soziale Beziehungen der Produktion versus Berufe als technische Beziehungen der Produktion (Wright 1980) nicht auf das Berufsstrukturmodell auf Basis der bundesdeutschen Sozialstatistik (NV 17) übertragen. Gerade weil dieses Berufsstrukturmodell die Stellung im Beruf, die Anzahl der Beschäftigten bei Selbständigen, Qualifikation, Entscheidungs- und Anweisungsbefugnis etc. umfasst, ist es zur Bündelung von sozialen Lagen sehr geeignet. Hervorzuheben ist, dass diese modifizierte Berufsvariable bei deutlich geringerem Aufwand (nämlich im Wesentlichen auf Basis einer einzigen Frage) sogar erfolgreicher ist als die konkurrierenden Ansätze.

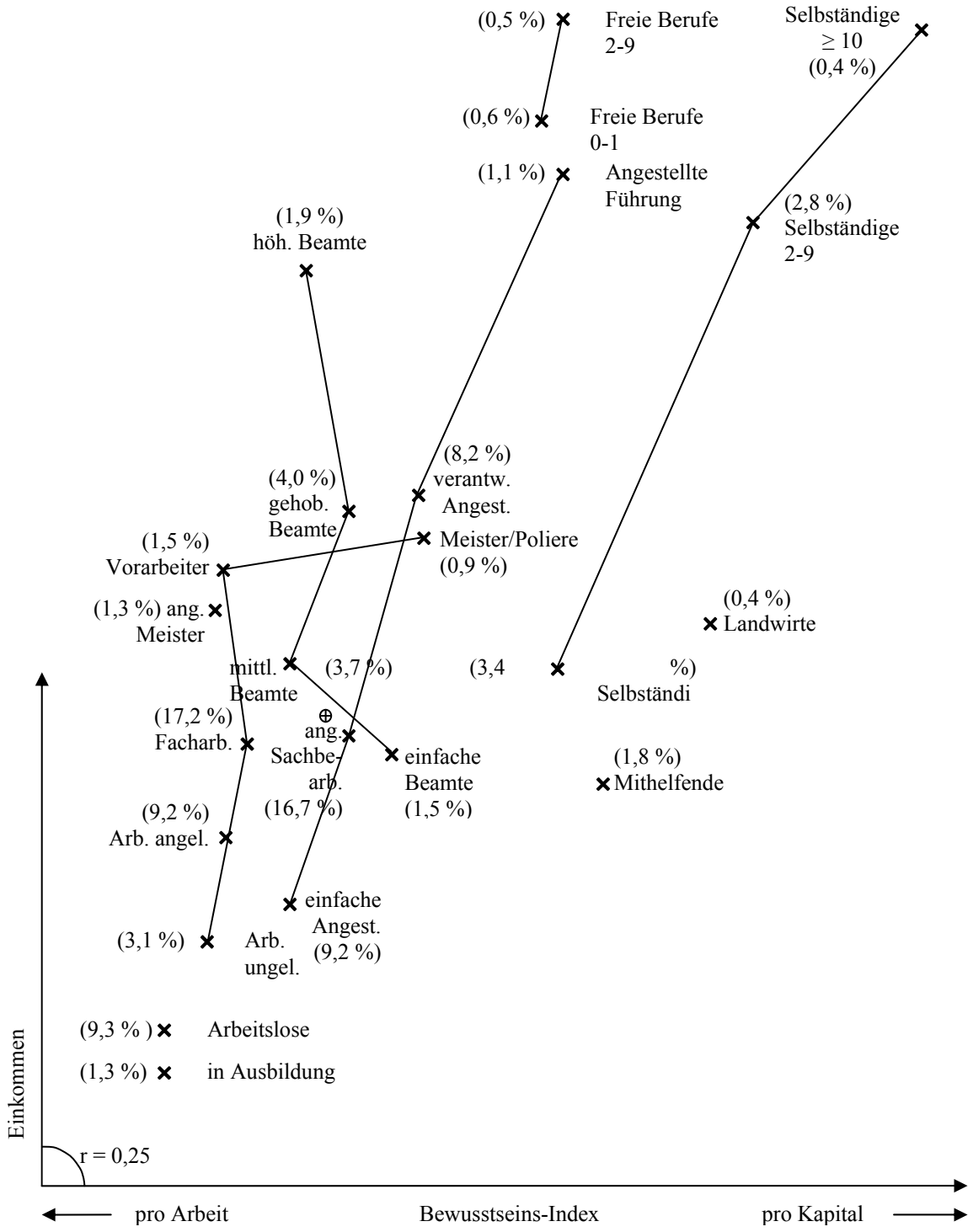
Die Landwirte werden unterschieden nach der Größe der landwirtschaftlich genutzten Fläche. Angesichts der geringen Fallzahl kann diese Unterteilung in der Analyse nicht weiter verfolgt werden. Die Landwirte rangieren in der Nähe der sonstigen Selbständigen mit 0-1 Mitarbeiter, ideologisch stehen sie weiter rechts (vgl. Abbildung 4-9).

Die Freiberufler und sonstigen Selbständigen werden unterschieden nach der Zahl der Mitarbeiter. Die sonstigen Selbständigen haben umso mehr Einkommen und stehen umso weiter rechts, je mehr Mitarbeiter sie haben. Die Verbindungslinien zwischen diesen Stellungen im Beruf weisen ungefähr einen linearen Trend auf. Für die Freiberufler gibt es einen ähnlichen Trend, nur dass sie ideologisch weniger rechts stehen als die sonstigen Selbständigen.

Die Mithelfenden rangieren in der Nähe der kleinen Selbständigen, sie erhalten nur etwas weniger Einkommen. (Per Definition sollten sie gar kein eigenes Einkommen haben, die meisten geben aber ein Einkommen an.)

Abbildung 4-9: Berufsstrukturmodell (NV 17) nach Einkommen und Bewusstseins-Index

Einkommen: Eta² = 39,3 % Bew.-Index: Eta² = 17,1 %
Mitte: ⊕
(Anteile der Berufsgruppen an den Befragten)



Auch bei den Lohnabhängigen handelt es sich nicht einfach um die bloße Stellung im Beruf, die erfasst wird. Da die verschiedenen Stellungen im Beruf unterschiedliche Laufbahnen implizieren, werden sinnvoller Weise auch die diese Laufbahnen jeweils strukturierenden Kriterien mit erhoben: Bei Beamten, Angestellten und Arbeitern ist eine jeweils spezifische Mischung von Schulabschluss, beruflichem Abschluss und beruflicher Erfahrung charakteristisch für die Stellung in der Betriebshierarchie nach Entscheidungs- und Anweisungsbefugnis.

Die Beamten insgesamt rangieren ideologisch etwa in der Mitte, und die Einkommensunterschiede entsprechen erwartungsgemäß der Laufbahnhierarchie. Die Beamtenlaufbahn ist weitgehend durch den Bildungsabschluss als Eingangsvoraussetzung strukturiert: Für den höheren Dienst benötigt man einen Hochschulabschluss, für den gehobenen Dienst einen mittleren Abschluss und für den einfachen und mittleren Dienst reichen Volks- bzw. Hauptschulabschluss. Im Gegensatz zu den Angestellten rangieren Beamte ideologisch auch dann in der Mitte, wenn sie selbst an der Spitze der Hierarchie stehen. Dies dürfte daran liegen, dass der öffentliche Dienst in der Polarität Arbeit versus Kapital eher einen neutralen Platz einnimmt: Die Arbeitnehmer im öffentlichen Dienst sehen sich nicht einem „Privatkapitalisten“ gegenüberstehen; deshalb ist die Polarität Arbeit versus Kapital für den öffentlichen Dienst weniger bewusstseinsrelevant.

Die Privatwirtschaft hat diese Hierarchie bei den Angestellten (die es natürlich auch bereits im öffentlichen Dienst gibt) kopiert. Es gibt einen fast perfekten linearen Trend von den einfachen Angestellten über die angestellten Sachbearbeiter sowie über die verantwortlichen Angestellten zu den leitenden Angestellten: Je höher Angestellte in der Hierarchie rangieren, desto höher ihr Einkommen und desto stärker ist ihr ideologischer Standort pro Kapital. Aus dem Rahmen fallen nur die angestellten Meister, die der Spitze der Arbeiter (eher den Vorarbeitern als den Meistern/Polieren) ähneln. Die einfachen Angestellten verrichten zum Teil ähnliche Tätigkeiten wie Arbeiter.

Die Arbeiter stehen insgesamt links von der Mitte: Die Betriebshierarchie fällt von den Meistern/Polieren, die als „leitende Arbeiter“ bereits den Arbeitgebern nahe stehen, über die Vorarbeiter zu den Facharbeitern sowie zu den angelernten und schließlich ungelernten Arbeitern, was sich im Einkommen widerspiegelt. Die letzteren beiden Gruppen dürften in der Stichprobe unterproportional berücksichtigt sein, da keine Ausländer befragt wurden. Die Strukturierung der Arbeiterschaft verläuft entsprechend der beruflichen Ausbildung als Eingangsvoraussetzung.

Verlängert man den linearen Trend über die ungelernten Arbeiter hinaus, so kommt man zu der Lage der Arbeitslosen und schließlich zu der Lage der Personen, die sich in Ausbildung befinden. Das Risiko der Arbeitslosigkeit ist für die ungelernten Arbeiter am höchsten, die soziale Lage der ungelernten Arbeiter ist unter den Erwerbstätigen der sozialen Lage der Arbeitslosen am ähnlichsten. Insofern strukturiert die berufliche Ausbildung auch noch die Betroffenheit durch Arbeitslosigkeit. Sich noch in der Ausbildung zu befinden, ist gemäß dem aktuellen Einkommen eine ähnlich benachteiligende Lage wie die der Arbeitslosen. Innerhalb der Gruppe der sich noch in Ausbildung Befindenden wird es allerdings erhebliche Unterschiede in der subjektiven Perspektive geben, je nachdem, welche beruflichen Aussichten mit der Ausbildung verbunden sind.

Eine reine Variable „Stellung im Beruf“ würde alle Angestellten zusammenfassen, angesichts der ausgeprägten Variation im Einkommen und Bewusstsein, die ich aufgezeigt habe, eine sehr ungünstige Vorgehensweise. Für die sonstigen Selbständigen, die Freiberufler, die Beamten und die Arbeiter gilt entsprechendes: Eine Vergröberung auf die reine Stellung im Beruf würde auf wichtige Informationen verzichten.

Nach der Graphik haben einfache Angestellte und ungelernte/angelernte Arbeiter eine vergleichbare materielle Lage. Trotzdem würde eine „klassenanalytische“ Zusammenfassung solcher Lagen ebenfalls zu einem Informationsverlust führen: Die Angestellten stehen ideologisch insgesamt weiter rechts als die Arbeiter, ferner sind die Arbeiter ideologisch homogener. Die Beamten rangieren ideologisch insgesamt in der Mitte, während die Angestellten mit einer höheren Stellung in der Hierarchie auch ideologisch weiter rechts einzuordnen sind.

Es soll nun anhand des Variationskoeffizienten (= Standardabweichung/Mittelwert) geprüft werden, wie homogen die angegebenen sozialen Lagen nach den Kriterien Einkommen und Bewusstseins-Index sind. Am heterogensten bezüglich des Einkommens sind die Mithelfenden, weitere Selbständigen-Gruppen sowie die leitenden Angestellten. Dies dürfte daran liegen, dass die Spitze der Einkommenshierarchie breit gefächert ist. Am homogensten bezüglich des Einkommens sind mittlere Beamten-, Arbeiter- und Angestelltengruppen. Bezüglich des Bewusstseins-Index sind die Landwirte am homogensten, sonst sind aber die Selbständigen-Gruppierungen nach diesem Kriterium gerade am heterogensten, was sich angesichts der großen Bandbreite ihrer materiellen Lage auch erwarten ließ. Dass die Landwirte trotz großer Streuung in der materiellen Lage ideologisch eher homogen sind, verweist auf Besonderheiten des Berufsstandes der Landwirte, für die die Erblichkeit und Verbundenheit mit Grund und Boden wohl einen besonderen Stellenwert haben. Ansonsten sind die Gruppierungen der Arbeiter ideologisch am homogensten, was für die Konstituierung als kollektiver Akteur besonders günstig ist. Allerdings handelt es sich dabei nur um etwa 30 % der in der Erhebung berücksichtigten Erwerbstätigen, in der Gesamt-Wahlpopulation wäre dieser Prozentsatz also noch geringer. Diese Gruppierungen lassen sich am ehesten durch die traditionelle Arbeiterbewegung, die Gewerkschaften und die Sozialdemokratie, mobilisieren.

Mögliche Reduktionen der Berufsvariablen

Da ich die vorher genannten Differenzierungen für informativ halte, sind weitere Reduktionen nicht zu befürworten, wenn es keine technischen Erfordernisse dafür gibt. Will man aber die Berufsvariablen nach Geschlecht oder anderen Variablen aufgliedern, sinkt die Fallzahl sehr schnell, so dass sich bei solchen Analysen Reduktionen der Berufsvariablen empfehlen könnten. Ferner lässt sich zeigen, dass diese reduzierten Berufsvariablen bessere Erklärungsanteile für die Kriterien aufweisen als die Klassenmodelle mit der gleichen (oder ungefähr gleichen) Anzahl der Kategorien, wie sich mit „relativen“ Erklärungsanteilen belegen lässt.

Reduzierte Berufsvariable (REDNV 17, vgl. Abbildung 4-10)

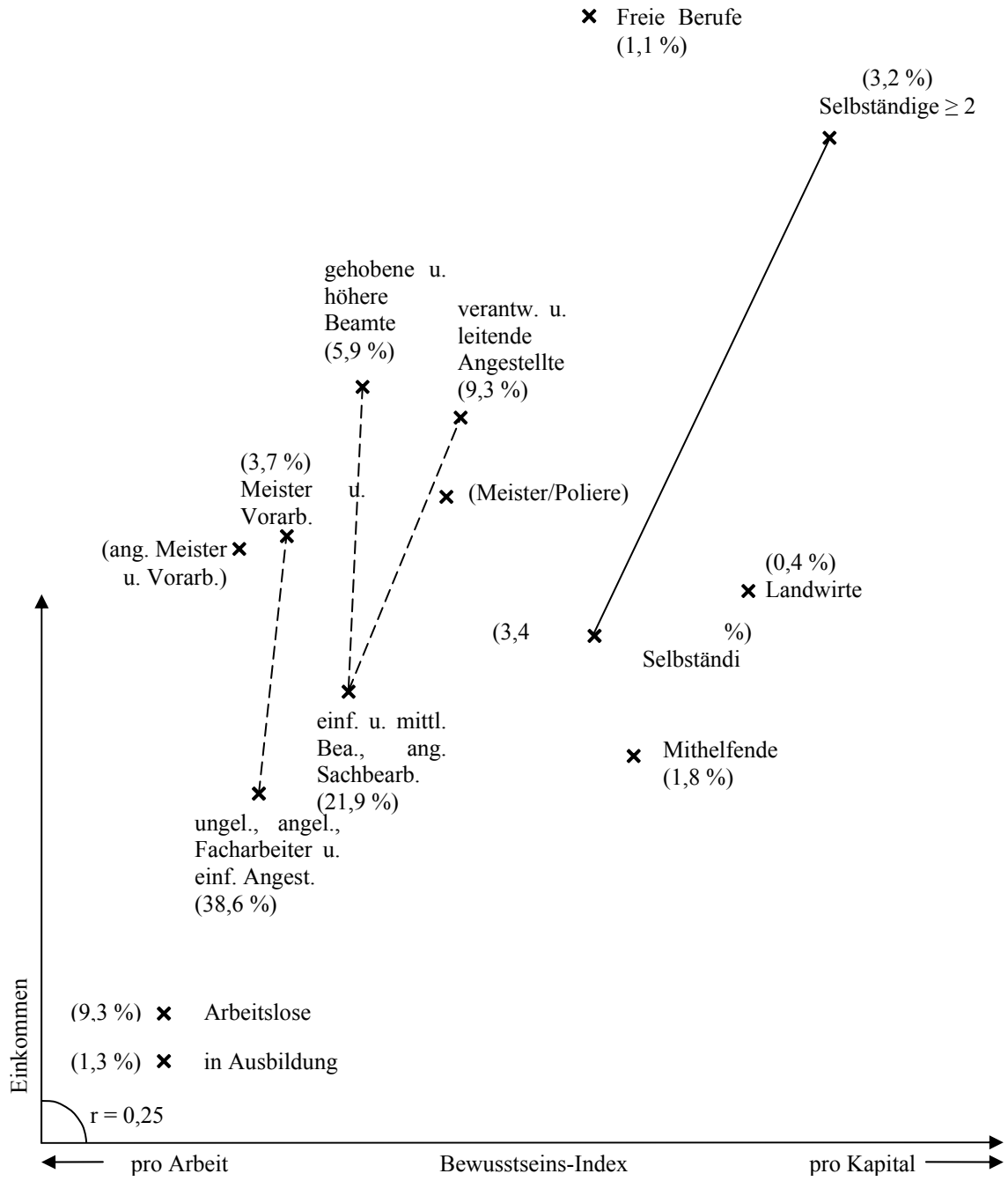
Da es in den Daten keine Freiberufler mit 10 oder mehr Beschäftigten gibt, bilden sie insgesamt eine recht homogene Gruppe. Auch die sonstigen Selbständigen mit 2 oder mehr Beschäftigten lassen sich als Gruppe betrachten.

Die beiden oberen Gruppen der Beamten sowie der Angestellten lassen sich jeweils zusammenfassen. Die beiden unteren Beamtengruppen lassen sich mit den angestellten Sachbearbeitern zusammenführen.

Die Meister und Vorarbeiter bilden eine relativ homogene Gruppe. Die Arbeiter bis hin zu den Facharbeitern und die einfachen Angestellten lassen sich zusammenfassen.

Abbildung 4-10: Reduzierte Berufsvariable (REDNV 17) nach Einkommen und Bewusstseins-Index

Einkommen: Eta² = 35,1 % Bew.-Index: Eta² = 16,2 %
Mitte: wie „einf. u. mittl. Bea., ang. Sachbearb.“
(Anteile der Berufsgruppen an den Befragten)



Nach diesen induktiv orientierten Zusammenfassungen sind die stärksten empirischen Zusammenhänge noch erhalten: Die Strukturierung der sonstigen Selbständigen nach Beschäftigtenzahl, die Hierarchien der Beamten und Angestellten von einer für beide Gruppen ähnlichen Basis aus, schließlich die Hierarchie der Arbeiter (wobei die einfachen Angestellten und die angestellten Meister eingeschlossen wurden).

Grobe Berufsvariable (GROBNV 17, vgl. Abbildung 4-11)

Eine weitere Vergröberung besteht darin, alle sonstigen Selbständigen zusammenzufassen.

Die oberen Beamten- und Angestelltengruppen könnte man alle zusammenführen.

Die Meister/Vorarbeiter könnte man mit den unteren Beamtengruppen und angestellten Sachbearbeitern zusammenfassen.

Ferner ließen sich Arbeitslose und Personen in Ausbildung zusammen betrachten.

In dieser induktiv orientierten Vergröberung verblieben als Strukturierung insbesondere die drei Hierarchie-Stufen der lohnabhängigen Erwerbstätigen.

Rest-Berufsvariable (RESTNV 17, vgl. Abbildung 4-11)

Schließlich könnte man noch alle Selbständigen und Mithelfenden zusammenfassen.

Literaturverzeichnis zum Anwendungsbeispiel

Holtmann, D., 1990: *Die Erklärungskraft verschiedener Berufsstruktur- und Klassenmodelle für die Bundesrepublik Deutschland. Ein Vergleich der Ansätze von IMSF, PKS, Walter Müller, Erik O. Wright und des Berufsstrukturmodells auf der Basis der bundesdeutschen Sozialstatistik*. In: Zeitschrift für Soziologie 19: 26-45 und 141-143.

Holtmann, D., Strasser, H., 1990: *Klassen in der Bundesrepublik heute: Zur Theorie und Empirie der Ausdifferenzierung von Handlungsressourcen*. In: Schweizer Zeitschrift für Soziologie 16: 79-106.

Müller, W., 1977: *Klassenlagen und soziale Lagen in der Bundesrepublik*. In: Handl, J., Mayer, K.U., Müller, W.: *Klassenlagen und Sozialstruktur*. Frankfurt: Campus.

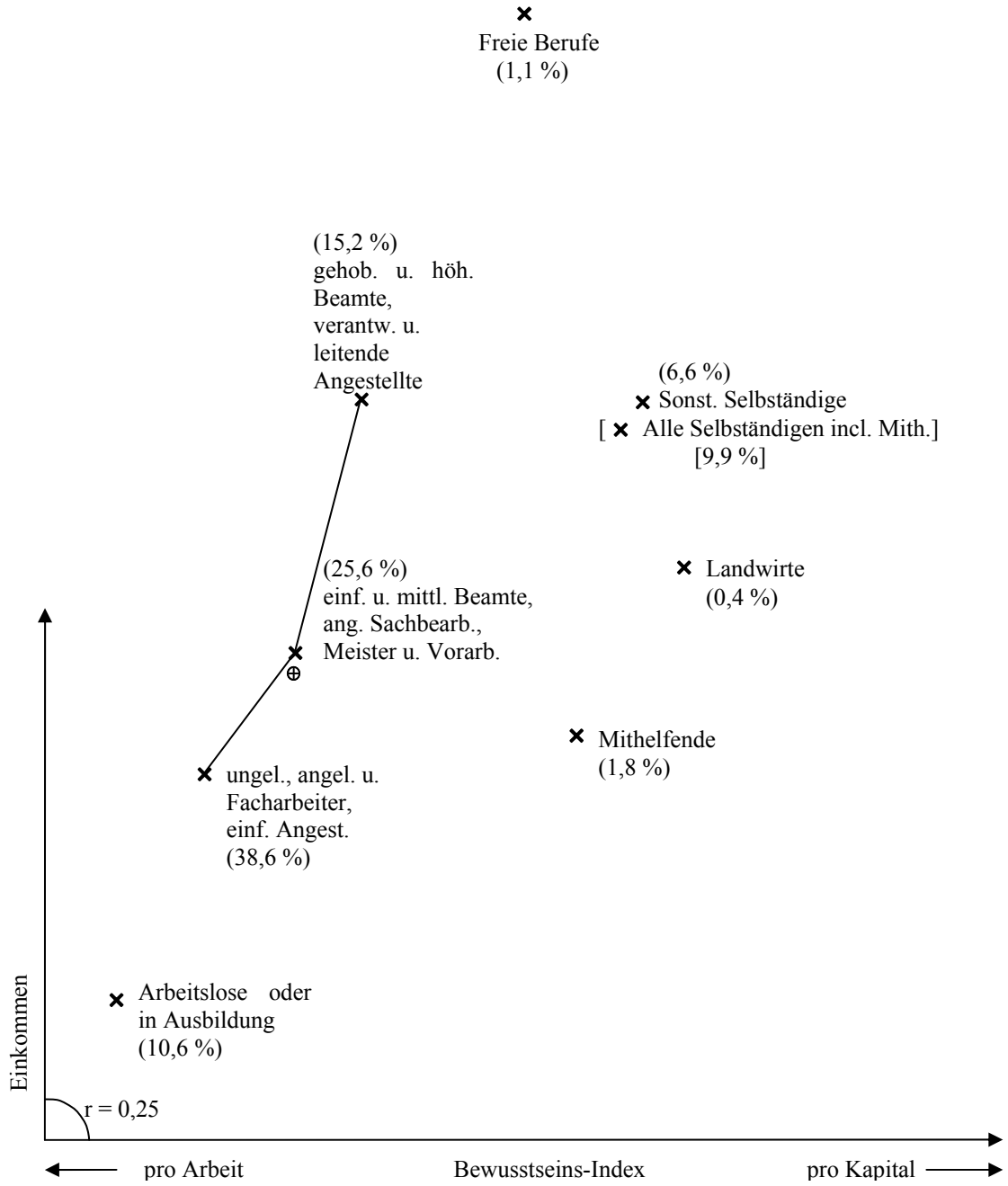
Roemer, J.E., 1982: *A general theory of exploitation and class*. Cambridge, Massachusetts: Harvard University Press.

Roemer, J.E., 1986: *Should Marxists be interested in exploitation?* In: Roemer, J.E. (Hg.), *Analytical Marxism*. Cambridge, Massachusetts: Cambridge University Press.

Wright, E.O., 1985: *Classes*. London: Verso.

Abbildung 4-11: Grobe Berufsvariable (GROBNV 17) nach Einkommen und Bewusstseins-Index (RESTNV 17)

Einkommen: Eta² = 31,0 % [28,0 %] Bew.-Index: Eta² = 15,0 % [14,9 %]
Mitte: ⊕
(Anteile der Berufsgruppen an den Befragten)



Literaturverzeichnis

Andrews, F.M., Morgan, J.N., Sonquist, J.A., Klem, L., 1973[2]: *Multiple Classification Analysis*. Ann Arbor: Institute for Social Research, The University of Michigan.

Bortz, J., 2005[6]: *Statistik für Human- und Sozialwissenschaftler*. Berlin: Springer.

Clauß, G., Finze, F.-R., Partzsch, L., 2002: *Statistik für Soziologen, Pädagogen, Psychologen und Mediziner*. Band 1. Frankfurt a.M.: Deutsch, Thun.

Cohen, J., Cohen, P., West S.P., Aiken L.S., 2003[3]: *Applied Multiple Regression. Correlation Analysis for the Behavioral Sciences*. Mahwah, New York: Lawrence Erlbaum.

Glaser, W.R., 1978: *Varianzanalyse*. Stuttgart: Fischer.

Harder, T., 1974: *Werkzeug der Sozialforschung*. München: Fink.

Hochstädter, D., Kaiser, U., 1988: *Varianz- und Kovarianzanalyse*. Frankfurt: Harri Deutsch.

Holm, K. (Hg.), 1975: *Die Befragung*. Band 6. München: Francke.

Hummell, H.J., 1972: *Probleme der Mehrebenenanalyse*. Stuttgart: Teubner.

Hummell, H.J., Ziegler, R., 1982: *Zur Verwendung linearer Modelle bei der Kausalanalyse nicht-experimenteller Daten*. In: Dieselben (Hg.): *Korrelation und Kausalität*. Band 1. Stuttgart: Enke.

Kerlinger, F.N., Pedhazur, E.J., 1973: *Multiple regression in behavioral research*. New York: Holt.

Lazarsfeld, P.F., 1955: *Interpretation of statistical relations as a research operation*. In: Lazarsfeld, P.F., Rosenberg, M. (Hg.): The language of social research. 7. printing. New York: The Free Press (1967), 115-125

Lazarsfeld, P.F., 1961: *The algebra of dichotomous systems*. In: Solomon, H. (Hg.), *Studies in item analysis and prediction*. Stanford: Stanford University Press.

Lazarsfeld, P.F., Henry, N.W., 1968: *Latent structure analysis*. Boston: Houghton Mifflin.

Liepelt, K., Mitscherlich, A., 1968: *Thesen zur Wählerfluktuation*. Frankfurt: EVA.

Mayntz, R., Holm, K., Hübner, P., 1978: *Einführung in die Methoden der empirischen Soziologie*. Opladen: Westdeutscher Verlag.

Morgan, J.N., Sonquist, J.A., 1963: *Problems in the analysis of survey data, and a proposal*. In: Journal of the American Statistical Association 58: 415-434.

Nie, N.H. et al., 1975[2]: *Statistical package for the social sciences (SPSS)*. New York: McGraw-Hill.

Overall, J.E., Klett, C.J., 1983: *Applied multivariate analysis*. Melbourne, Florida: Krieger.

Pappi, F.U., 1977: *Aggregatdatenanalyse*. In: Van Koolwijk, J., Wieken-Mayser, M. (Hg.), *Techniken der empirischen Sozialforschung*. Band 7. München, Wien: Oldenbourg.

Pfanzagl, J., 1967: *Allgemeine Methodenlehre in der Statistik.* Band II. Berlin: DeGryter.

Robinson, W.S., 1950: *Ecological correlations and the behavior of individuals*. In: American Sociological Review 15: 351-357.

Rosenberg, M., 1968: *The logic of survey analysis*. New York: Basic Books.

Scheffé, H.A., 1999: *The analysis of variance*. New York: Wiley InterSciences.

Schuessler, K.F., 1971: *Analyzing social data. A statistical orientation*. Boston: Houghton Mifflin.

Sonquist, J.A., Morgan, J.N., 1964: *The detection of interaction effects*. A report on a computer program for the selection of optimal combinations of explanatory variables. Ann Arbor: Survey Research Center, University of Michigan.

Sonquist, J.A., 1975[3]: *Multivariate Model Building*. The validation of a search strategy. Ann Arbor: Survey Research Center, University of Michigan.

Stouffer, S. et al., 1949: *The American Soldier*. Princetown: Princeton University Press.

Winer, B.J. et al., 1991[3]: *Statistical principles in experimental design*. New York: McGraw-Hill.

Wonnacott, Th.H., Wonnacott, R.J., 1990[4]: *Introductory Statistics for Business and Economics*. New York: Wiley.

Anhang: Multiple Regressionsanalyse mit Hilfe von Determinanten

Ein Gleichungssystem

$y_1 = a_{11} x_1 + \ldots + a_{1n} x_n + b_1$
$y_i = a_{i1} x_1 + \ldots + a_{in} x_n + b_i$
$y_m = a_{m1} x_1 + \ldots + a_{mn} x_n + b_m$

kann mit Hilfe der Matrixalgebra allgemein gelöst werden.

Eine **Matrix** ist ein rechteckiges Schema mit m Zeilen (horizontal) und n Spalten (vertikal).

$$A := \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & & \ddots & \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}$$

Kurzform: $A = (a_{ij})$, wobei a_{ij} das Element in i-ter Zeile und j-ter Spalte ist.

Matrixmultiplikation:

Ist A eine (m, n)-Matrix und B eine (n, k)-Matrix, so ist das Produkt der Matrizen AB = C definiert durch:

$$c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}$$

Das Produkt von Matrizen lässt sich nur berechnen, wenn die Anzahl der Spalten der ersten Matrix der Anzahl der Zeilen der zweiten Matrix entspricht. Das Produkt BA kann daher nicht berechnet werden, falls $m \neq k$.

Transponierte Matrix:

Die zu der (m, n)-Matrix A transponierte Matrix A' ist die (n, m)-Matrix mit der Eigenschaft:
$a'_{ij} = a_{ji}$
Also: (A')' = A, (AB)' = B' A'

Das Gleichungssystem lässt sich nun in Matrixform schreiben.

Mit $y := \begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix}$, $A := \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix}$, $x := \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$, $b := \begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix}$ lautet das Gleichungssystem in Matrixschreibweise: y = Ax + b

Die **Determinante** (Abkürzung: det (A) oder |A|) einer quadratischen Matrix $A := \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$ ist definiert als: $\det(A) = a_{11} a_{22} - a_{12} a_{21}$
Für eine 2 x 2 Matrix entspricht sie dem Kreuzprodukt.

Die Determinante für größere Matrizen kann rekursiv definiert werden durch folgenden Zusammenhang:

Sei A_{ij} die (n - 1, n - 1)-Matrix, die dadurch entsteht, dass die i-te Zeile und j-te Spalte von A weggelassen werden. Die Auflösung der Determinante nach der i-ten Zeile lautet dann:

$$\det(A) = \sum_{j=1}^{n} (-1)^{i+j} a_{ij} \det(A_{ij})$$

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \quad \det(A) = a_{11} \cdot \begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} - a_{12} \begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix} + a_{13} \begin{vmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{vmatrix}$$

Dadurch, dass die Determinante einer Matrix sich als Linearkombination der Determinanten von quadratischen (n - 1, n - 1)-Teilmatrizen berechnen lässt, ist die Determinante rekursiv definiert für jede quadratische Matrix.

$E = \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix}_{(n,n)}$, die **Einheitsmatrix**, ist das neutrale Element bzgl. der Matrixmultiplikation von quadratischen (n, n)-Matrizen: $E \cdot A = A \cdot E = A$

Wenn es zu einer quadratischen Matrix A eine **inverse Matrix** (auch einfach: Inverse) A^{-1} gibt, d.h., $A\,A^{-1} = A^{-1}\,A = E$, so lässt sich diese Inverse wie folgt berechnen:

Das allgemeine Element g_{ij} von A^{-1} lautet:

$$g_{ij} = \frac{(-1)^{j+i} \det(A_{ji})}{\det(A)}$$

(Die Indizes sind also transponiert).

Für die gesamte Matrix gilt:

$$A^{-1} = \frac{adj(A)}{\det(A)}$$

Die einzelnen Elemente der **adjungierten Matrix** adj(A) bestehen aus Unterdeterminanten. Die Unterdeterminanten erhält man aus der transponierten Matrix (A'), indem man dort Zeile und Spalte entsprechend dem Index der Unterdeterminante streicht. Das Vorzeichen lautet:
$(-1)^{i+j}$

Beispiel:
$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$$

$$adj(A) = \begin{pmatrix} +\det(A_{11}) & -\det(A_{12}) & +\det(A_{13}) \\ -\det(A_{21}) & +\det(A_{22}) & -\det(A_{23}) \\ +\det(A_{31}) & -\det(A_{32}) & +\det(A_{33}) \end{pmatrix}$$

Es gilt: $(AB)^{-1} = B^{-1}\,A^{-1}$

Falls die Matrix A invertierbar ist (d.h. A^{-1} existiert), so ist die Lösung des Gleichungssystems y = Ax einfach: $x = A^{-1} \cdot y$

Cramers Regel: Das Gleichungssystem y = Ax lässt sich auch auf andere Weise relativ elegant lösen.
A (i → y) sei die Matrix, die dadurch entsteht, dass die i-te Spalte von A ersetzt wird durch den Vektor y.

Dann lässt sich die Lösung auch wie folgt schreiben:

$$x_i = \frac{\det(A(\mathrm{i} \to \mathrm{y}))}{\det(A)}$$

Beispiel:

In der multiplen Regression für k = 2 ist zu lösen: R β= r,
wobei $R = \begin{pmatrix} 1 & r_{12} \\ r_{21} & 1 \end{pmatrix}, \quad r = \begin{pmatrix} r_{y1} \\ r_{y2} \end{pmatrix}$

Also: $\det(R) = 1 - r_{12}^2$

$$\det(R(1 \to r)) = \det\begin{pmatrix} r_{y1} & r_{12} \\ r_{y2} & 1 \end{pmatrix} = r_{y1} - r_{12} r_{y2}$$

$$\det(R(2 \to r)) = \det\begin{pmatrix} 1 & r_{y1} \\ r_{21} & r_{y2} \end{pmatrix} = r_{y2} - r_{12} r_{y1}$$

Deshalb erhält man für $\beta = R^{-1} r$:

$$\beta_1 = \frac{\det(R(1 \to r))}{\det(R)} = \frac{r_{y1} - r_{12} r_{y2}}{1 - r_{12}^2}$$

$$\beta_2 = \frac{\det(R(2 \to r))}{\det(R)} = \frac{r_{y2} - r_{12} r_{y1}}{1 - r_{12}^2}$$

Für k = 3 sei nur die Determinante als Beispiel angegeben, sowie β_1.

$$R = \begin{pmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{pmatrix}$$

Nach der ersten Zeile aufgelöst:
$\det(R) = 1 \cdot (1 - r_{23}^2) - r_{12}(r_{21} - r_{23} r_{31}) + r_{13}(r_{21} r_{32} - r_{31})$
$\det(R(1 \rightarrow r)) = r_{y1} \cdot (1 - r_{23}^2) - r_{y2}(r_{21} - r_{23} r_{31}) + r_{y3}(r_{21} r_{32} - r_{31})$
Also: $\beta_1 = \det(R(1 \rightarrow r)) / \det(R)$

Falls man nicht nur einige wenige unabhängige Variable hat, so hat es sich herausgestellt, dass die Berechnung der β_i nach Cramers Regel weniger günstig ist als das folgende Vorgehen:

$$\text{Sei } S := \begin{pmatrix} r_{yy} & r_{y1} & \cdots & r_{yk} \\ r_{1y} & r_{11} & \cdots & \\ \vdots & & \ddots & \\ r_{ky} & & \cdots & r_{kk} \end{pmatrix} = \begin{pmatrix} r_{yy} & r_{y1} & \cdots & r_{yk} \\ r_{1y} & & & \\ \vdots & & R & \\ r_{ky} & & & \end{pmatrix}$$

S_{ab} sei die (k, k)-Matrix, die entsteht, wenn die Zeile der Variablen mit dem Zeilenindex a und die Spalte der Variablen mit dem Spaltenindex b gestrichen wird. Z.B. ist also $S_{yy} = R$.

Es gilt: $\beta_i = (-1)^{i-1} \det(S_{yi}) / \det(S_{yy})$ (für i = 1, ..., k)

Die Determinante $\det S_{yi}$ soll durch Auflösung nach der 1. Spalte berechnet werden:

$$\det(S_{yi}) = \sum_{j=1}^{k} (-1)^{j+1} r_{jy} \underbrace{\det\left((S_{yi})_{jy}\right)}_{\det(R_{ji})}$$

$$\beta_i = \sum_{j=1}^{k} (-1)^{i+j} r_{jy} \det(R_{ji}) / \det(R)$$

Dies ist genau die i-te Komponente des Vektors $R^{-1} r$.

$$\left(\sum_{j=1}^{k} g_{ij} r_{yj} = \sum_{j=1}^{k} (-1)^{j+i} r_{yj} \det(R_{ji}) / \det(R) \right)$$

Z.B. lässt sich β_1 nach diesem Verfahren wie folgt berechnen:

$$\det(S_{y1}) = \det\begin{pmatrix} r_{yy} & r_{y1} & r_{y2} \\ r_{1y} & r_{11} & r_{12} \\ r_{2y} & r_{21} & r_{22} \end{pmatrix} = r_{1y} - r_{12} r_{2y}$$

$$\det(S_{yy}) = \det\begin{pmatrix} r_{yy} & r_{y1} & r_{y2} \\ r_{1y} & r_{11} & r_{12} \\ r_{2y} & r_{21} & r_{22} \end{pmatrix}_{yy} = \det(R) = 1 - r_{12}^2$$

$$\beta_1 = \frac{(-1)^2 (r_{1y} - r_{12} r_{2y})}{1 - r_{12}^2} = \frac{r_{1y} - r_{12} r_{2y}}{1 - r_{12}^2}$$

Auch Multiple R^2 lässt sich durch Determinanten ausdrücken:
det (S), aufgelöst nach der 1. Zeile, ergibt:

$$\det(S) = \sum_{i=1}^{k} (-1)^{i+2} r_{yi} \det(S_{yi}) + 1 \cdot \underbrace{\det(S_{yy})}_{\det(R)}$$

$$\frac{\det(R) - \det(S)}{\det(R)} = \sum_{i=1}^{k} (-1)^{i+1} r_{yi} \det(S_{yi}) / \det(R)$$

$$= \langle y, \underbrace{\sum_{i=1}^{k} (-1)^{i+1} x_i \det(S_{yi}) / \det(R)}_{\sum_{i=1}^{k} \beta_i x_i} \rangle / n$$

$$= \langle y, \hat{y} \rangle / n = \text{Multiple } R^2$$

Also: $\textit{Multiple } R^2 = 1 - \frac{\det(S)}{\det(R)}$

$$s^2_{y-\hat{y}} = 1 - \textit{Multiple } R^2 = \frac{\det(S)}{\det(R)}$$

Entsprechend: $s^2_{x_i - \hat{x}_i} = \frac{\det(R)}{\det(R_{ii})}$

$$s^2_{y-\hat{y}(i)} = \frac{\det(S_{ii})}{\det(R_{ii})}$$

Damit lassen sich auch die übrigen Koeffizienten in Determinantenform ausdrücken:

Part Correlation

$$r_{y,\, x_i - \hat{x}_i} = \beta_i \cdot s_{x_i - \hat{x}_i} = \frac{(-1)^{i+1} \det(S_{yi}) \sqrt{\det(R)}}{\det(R) \sqrt{\det(R_{ii})}} = \frac{(-1)^{i+1} \det(S_{yi})}{\sqrt{\det(R) \cdot \det(R_{ii})}}$$

Partial Correlation

$$r_{y - \hat{y}_{(i)}, x_i - \hat{x}_i} = \beta_i \frac{s_{x_i - \hat{x}_i}}{s_{y_i - \hat{y}(i)}} = \beta_i \cdot \sqrt{\frac{\det(R)}{\det(S_{ii})}} = \frac{(-1)^{i+1} \det(S_{yi})}{\sqrt{\det(R) \cdot \det(S_{ii})}}$$

Sachregister

Allgemeines lineares Modell 2, 7
Autokorrelation 111
bereinigter Zusammenhang 142
Clusteranalyse 2
Codierungen in der Varianzanalyse 192
conjoint influence 43
Delta 24
Design-Matrix 195
Determinante 232
Differenz der Kreuzprodukte 20, 25
Diskriminanzanalyse 2, 5, 146
Distorter-Phänomen 31, 35, 143, 152
Drittvariablenkontrolle 15,67
Durbin-Watson-Test auf Autokorrelation 111, 116
Effektkodierung 49, 193, 194
Effekte, additive 96
Effekte, multiplikative 96
Effekte in der multiplen Regression 95, 96, 102, 119, 153, 154, 157
Effekte in der Pfadanalyse 135, 138, 139, 153, 154, 157
Effekte in der Tabellenanalyse 18, 28, 41, 42, 66
Effekte in der Varianzanalyse 46, 47, 56, 58
Einheitsmatrix 83, 120, 233
endogene Variablen 127
erklärte Varianz in der multiplen Regression 95, 102, 118, 153, 154, 157
erklärte Varianz in der Pfadanalyse 137, 138, 139, 145, 153, 154, 157
erklärte Varianz in der Tabellenanalyse 102
erklärte Varianz in der Varianzanalyse 169, 187, 189
Erklärungskraft, gesamte – vgl. erklärte Varianz, Multiple R und Kap. 4.10 (S. 218)
Erklärungskraft eines Prädiktors, bereinigte 93, 107, 153, 154, 156, 157
Eta-Quadrat 164, 170, 180
exogene Variablen 127
Faktorenanalyse 1, 6, 8, 9, 10
Fundamentaltheorem der Pfadanalyse 127
graphische Darstellung der Varianzanalyse 223
Haupt- und Interaktionseffekte in der Varianzanalyse 47
Heteroskedastizität vs. Homoskedastizität 111, 114
inneres Produkt 68
Interaktion 54
Interaktion in der Regression 96
Interaktionseffekte in der Tabellenanalyse 45, 47
intervenierende Variable 29
inverse Matrix 84, 233
Kanonische Korrelation(sanalyse) 1, 6
Kausalbegriff, statistischer 13
Kausalbeziehung 13
kausale Geschlossenheit 128
kausale Interpretation von Zusammenhängen 65
kausaler Effektkoeffizient in der Pfadanalyse 130
kausale Ordnung 125
Kausalstrukturen, Typologie von 29
kleinsten Quadrate, Methode der 77
Kollinearität – vgl. Multikollinearität

Konfigurationsfrequenzanalyse 2
Konstanthalten (kontrollieren) einer Variablen 15, 67
Kontrastgruppenanalyse 211
Kontrollvariable (Drittvariablenkontrolle) 15, 67
Korrelationskoeffizient 80
Korrelationskoeffizient , multipler (R) 84
Korrelationskoeffizient, partiell 67
Korrelationskoeffizient, semi-partiell 87
Korrelationsmatrix 83
Kovarianz 40, 41, 67, 71, 95, 156
Kovarianzanalyse 198
Kovarianzmatrix 120
Kovarianzzerlegung 35, 38, 40, 63, 65, 199, 204
Kreuzprodukte 20
latent structure analysis 2
Logistische Regression 103
Logits 104
log-lineares Modell 21
Log-Odds 104
Matrix 82, 232
Mechanismen zur Erklärung einer Korrelation 134
Messniveau, erforderliches 4
Methode der kleinsten Quadrate 77
multidimensionale Skalierung 2
Multikollinearität 109, 112
Multiple Classification Analysis (MCA) 191
Multiple R (Multiple Korrelation) 84, 138
Multiple Regression 75
multivariate Modelle 1
Nichtlinearität 110, 113
Nicht rekursive versus rekursive Modelle 6, 125
Odds 104
Orthogonalität 90
Part Correlation 87, 92
partielle Korrelation 67, 68, 92, 158, 160
Phi 27
Pfadanalyse 5, 124
Pfaddiagramm 129, 135
Pfadkoeffizienten 126, 158
Pfadtheorem 134
Q, Yules 26
Regression, einfache (bivariate) 77
Regression, multiple 5, 8, 9, 10, 75
Regression, multivariate 120
Regression, schrittweise 88, 94
Regressionskoeffizienten 79, 80, 86, 91
Regressionskoeffizienten, F-Test auf Signifikanz 107
Rekursivität 125
Residualvariablen in der Pfadanalyse 127
Residuen, Normalverteilung der 111, 115
saturiertes Modell 58, 182, 197
scheinbare Nicht-Kausalität 144, 158
Schein-Effekt 143
scheinkausale Korrelation 16, 29

Scheinkorrelation 16, 29
schrittweise Regression 88, 94
Skalarprodukt 68
Spezifikation 43, 54
standardisierte vs. unstandardisierte Koeffizienten 133
Standardisierung 79
Streudiagramm 114
Streuungszerlegung 84, 89, 145, 165
Suppressor-Phänomen 30, 32, 142, 151
Tabellenanalyse 15
Toleranz (Regression) 113
transponierte Matrix 82, 232
tree analysis 211
typologische Effekte 45, 54
Überschneidung in der Erklärung 144
unstandardisierte vs. standardisierte Koeffizienten 133
unvollständiges Modell 126, 133
Varianz, erklärte 95, 102
Varianzanalyse 1, 5, 8, 9, 10, 46, 54, 163
Varianzanalyse, als Test in der multiplen Regression 109
Varianzanalyse, dreifache 177
Varianzanalyse, Effekte in der 184, 186, 187, 189, 193, 194
Varianzanalyse, eindeutige (unique) Methode 177, 185
Varianzanalyse, einfache 163, 164, 179, 192, 193
Varianzanalyse, erklärte Varianz und Effekte 186, 187, 189
Varianzanalyse, experimenteller Ansatz 176, 184
Varianzanalyse, graphische Darstellung 223
Varianzanalyse, Haupteffekte 46, 47, 55, 56, 184, 185, 188
Varianzanalyse, hierarchischer Ansatz 176, 185
Varianzanalyse, Interaktionseffekt 46, 47, 54, 56, 184, 185, 188
Varianzanalyse, klassischer Ansatz 176, 184
Varianzanalyse, Regressions-Ansatz 177, 185
Varianzanalyse, zweifache 171, 181, 183, 194
varianzanalytische Interpretation in der Tabellenanalyse 46, 54, 56
Varianzhomogenität (Homoskedastizität) vs. Heteroskedastizität 111, 114
Varianzzerlegung – vgl. Streuungszerlegung
vollständiges Modell 126
Vorzeichenregel nach Davis 39
Vorzeichenregel auf Basis der Kovarianz 40, 156
Weighted Least Squares (WLS) 120
Zerlegung der Vier-Felder-Tafel 17
Zerlegungsformel für Maßzahlen 24
z-Transformation 79

Zur adäquaten Analyse sozialwissenschaftlicher Phänomene ist die Anwendung multivariater Modelle hilfreich, die die Analyse von Zusammenhängen und Abhängigkeiten zwischen vielen Merkmalen ermöglichen.

Als grundlegende Modelle werden im folgenden Band behandelt:
Die Elaboration von Zusammenhängen lässt sich durch Teilgruppenvergleich (Tabellenanalyse) auf nominalem Messniveau und durch partielle Korrelation auf metrischem Messniveau durchführen. In der multiplen Regression wird die Variation eines interessierenden Phänomens auf die Variation einer Reihe von Erklärungsfaktoren zurückgeführt. Die wichtigsten Interpretationshilfen dabei sind der Anteil der erklärten Varianz und die Effekte. In der Pfadanalyse werden alle Mechanismen herausgearbeitet, durch deren Zusammenwirken die Höhe jedes statistischen Zusammenhangs bestimmt wird: Direkte und indirekte Kausaleffekte, scheinkausale Komponenten und Assoziationseffekte. In der Varianzanalyse wird die Variation eines interessierenden Phänomens auf Haupteffekte und Interaktionseffekte einer Reihe von Erklärungsfaktoren zurückgeführt.

ISBN 978-3-86956-084-7