

ZHSF – Herbstseminar 1999, Seminarabschluß.
Materialien zur Einführung in multivariate
Analyseverfahren

Ulrich Kohler
Universität Mannheim

30. Januar 2001

Inhaltsverzeichnis

1	Einleitung: Typen multivariater Verfahren	1
1.1	Die allgemeine Idee der Regression	1
1.2	Klassifikationsverfahren, Skalierungsverfahren	4
2	Multiple Regression	5
2.1	Drittvariablenkontrolle in Tabellen	5
2.2	Prozentsatzdifferenzen und lineare Regression	7
2.3	Mögliche Konstellationen bei der Drittvariablenkontrolle	9
2.4	Kontrolle in linearen Regressionsmodellen	10
2.5	Etwas Formaler: Lineare Regression mit drei und mehr Variablen	12
2.6	Das allgemeine lineare Modell, oder: Was tun bei kategorialen unabhängigen Variablen?	16
2.6.1	Codierung einer dichotomen unabhängigen Variablen	16
2.6.2	Codierung einer polytomen nominalskalierten unabhängigen Variablen	17
2.7	Literatur	22
3	Strukturgleichungsmodelle	25
3.1	Was sind Pfaddiagramme	25
3.2	Standardisierte Pfadkoeffizienten	26
3.3	Wrights Rules	27

3.4	Die Berechnung der Pfadkoeffizienten aus der Korrelationsmatrix	28
3.5	Die Berechnung der Pfadkoeffizienten aus der Kovarianzmatrix	29
3.6	Pfadkoeffizienten für Pfadmodelle mit latenten Variablen . . .	30
3.7	Meßfehler und ihre Konsequenzen	32
3.8	Probleme linearer Strukturgleichungsmodelle	34
3.9	Grundschema linearer Strukturgleichungsmodelle	35
3.9.1	Das Strukturmodell	36
3.9.2	Das Meßmodell auf der ξ -Seite	37
3.9.3	Das Meßmodell auf der η -Seite	38
3.10	Überblick über die Lisrel Notation	39
3.11	Faktorenanalyse	39
3.12	Literatur	40
3.13	Software für SEM's	41
4	Logistische Regression	43
4.1	Warum nicht einfach die lineare Regression?	43
4.2	Odds, Odds-Ratio oder: Wetten dass!	44
4.2.1	odds	45
4.2.2	odds-ratio	46
4.3	Regressionsmodell	47
4.4	Gesamtfit des Modells	50
4.5	Test des Einfluß einzelner Variablen	51
4.6	Zum Maximum-Likelihood-Verfahren	52
4.7	Literatur	55
5	Loglineare Modelle	57
5.1	Bestimmung des Modellfits	61
5.2	Verbesserungen des Modells	62
5.3	Ein multivariates Beispiel	63

<i>INHALTSVERZEICHNIS</i>	iii
5.4 Literatur	66
6 Ereignisdatenanalyse	69
6.1 Literatur	70
7 Skalierungs- und Klassifikationsverfahren	71
7.1 Nochmal: Faktorenanalyse (hier: Hauptkomponentenanalyse)	71
7.2 Hierarchische Clusteranalyse	74
7.3 Multidimensionale Skalierung	78
7.4 Literatur	79

Kapitel 1

Einleitung: Typen multivariater Verfahren

Als multivariate Verfahren werden in der Regel *Verfahren zur Analyse von mehr als zwei abhängigen Variablen* bezeichnet. Im weiteren Sinne werden mit dem Ausdruck jedoch auch Verfahren bezeichnet, bei denen — in welcher Form auch immer — mehr als zwei Variablen beteiligt sind. Im folgenden werden wir uns mit multivariaten Verfahren in diesem weiteren Sinne befassen.

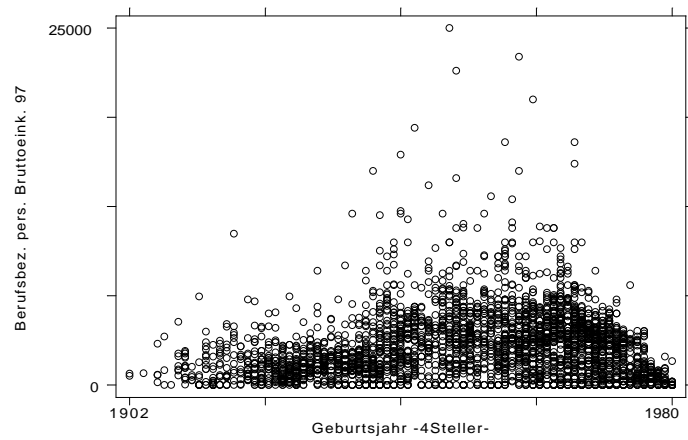
Der Schwerpunkt der Darstellung liegt auf den *Kausalanalyseverfahren*. Dies sind Verfahren, die danach Fragen, wie stark eine oder mehrere Variablen von anderen Variablen abhängen. Kausalanalyseverfahren beruhen auf der allgemeinen Idee der Regression.

1.1 Die allgemeine Idee der Regression

(Nachfolgende Darstellung wurde von Josef Brüderl für das Seminar „Regressionsverfahren“ im Sommersemester 1999 an der Uni-Mannheim angefertigt)

Wir betrachten zwei Variablen (Y, X) . Unsere Daten sind die realisierten Werte dieser Variablen. In einem Streudiagramm betrachten wir die zweidimensionale Verteilung von Y und X .

2KAPITEL 1. EINLEITUNG: TYPEN MULTIVARIATER VERFAHREN



In einer Regression betrachtet man die bedingte Verteilung von Y in Abhängigkeit von den Werten von X (Regression von Y auf X). Y wird als abhängige Variable bezeichnet und X als unabhängige. In der Regressionsanalyse beschäftigt man sich also mit der bedingten Verteilung

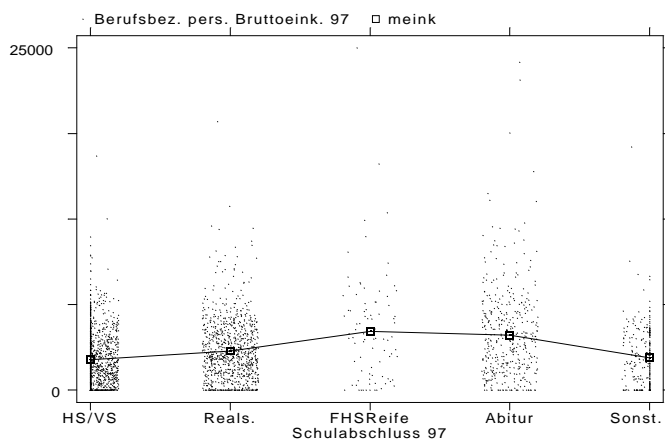
$$f(Y = y|X = x). \quad (1.1)$$

Wir ordnen damit jedem Merkmalswert von X eine Funktion zu, und zwar die bedingte Verteilung von Y . Dies ist praktisch nur schwer darstellbar. Deshalb charakterisiert man die bedingten Verteilungen durch (eine) Kennzahl(en):

- Y metrisch: bedingter Mittelwert
- Y metrisch, ordinal: bedingtes Quantil
- Y nominal: bedingte Häufigkeiten

Es hängt vom Meßniveau von Y ab, welche Kennzahl, man verwenden kann. Aber selbst für nominales Y läßt sich die bedingte Verteilung durch Kennzahlen beschreiben. Damit ist eine Regression für jedes Y -Meßniveau durchführbar.

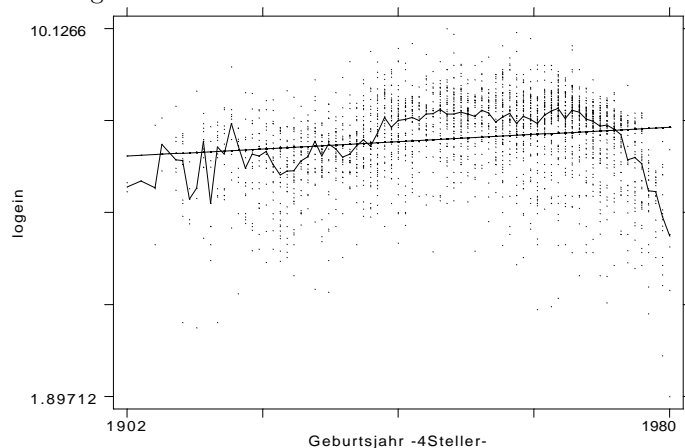
Regression mit diskretem X In diesem Fall errechnet man für jeden vorkommenden X -Wert die Kennzahl:



Regression mit stetigem X In diesem Fall ist die direkte Errechnung der Kennzahl nicht praktikabel, weil für manche X -Werte nur wenige Y -Werte vorliegen. Es kommen zwei Verfahren in Betracht. In *Lokalen Verfahren* werden für jeden (im Prinzip) möglichen X -Wert werden die Y -Werte in einer Umgebung von x benutzt um die Kennzahl zu berechnen. Die Kennzahlen verbindet man dann mit einer Linie (geglättete Regressionsfunktion). Je größer man die Umgebung wählt, desto glatter wird die Regressionsfunktion. Man bezeichnet diese Verfahren auch als nicht-parametrische Regression. Beispiele: Lokale Mean (Median) Regression, Lowess.

In *Regressionsmodellen* nimmt man an, dass die Kennzahlen einer Funktion folgen: $g(x; \theta)$. Man unterstellt also ein parametrisches Regressionsmodell. Gegeben die Daten und das gewählte Modell, schätzt man die Parameter θ so, dass die Regressionsfunktion am besten auf die Daten paßt. Man muß sich also zusätzlich noch für ein Schätzkriterium entscheiden. Üblich sind OLS und Maximum-Likelihood-Verfahren.

Folgende Graphik zeigt den Vergleich einer lokalen „Median“ Regression mit einer parametrischen Regression nach dem OLS-Verfahren.



4KAPITEL 1. EINLEITUNG: TYPEN MULTIVARIATER VERFAHREN

Im folgenden werden parametrische Regressionsverfahren besprochen. Zunächst wird die Regression mit Mittelwerten (*Multiple Regression*) und eine spezielle Anwendung dieser Technik, die *Strukturgleichungsmodelle* behandelt.

Ist die abhängige Variable nominal, so sind Mittelwert-Regressionen nicht sinnvoll. Man kann jedoch die relativen Häufigkeiten betrachten, und untersuchen, wie dieselben von den Werten der X -Variablen bedingt werden. Bei dichotomen abhängigen Variablen führt dies zur *logistischen Regression*.

Betrachtet man die relativen Häufigkeiten mehrerer Ausprägungen erhält man die *multinomiale logistische Regression*. Besprochen wird der Sonderfall der multinomialen logistischen Regression, bei der alle unabhängigen Variablen kategorial sind (*loglineare Modelle*).

Schließlich wird der Fall betrachtet, in der die Zeitdauer bis zum Eintritt eines Ereignisses die abhängige Variable ist (*Ereignisdatenanalyse*).

1.2 Klassifikationsverfahren, Skalierungsverfahren

Die zweite Gruppe von multivariaten Verfahren sind Klassifikations- und Skalierungsverfahren. Klassifikationsverfahren fragen danach, ob sich innerhalb der Daten Beobachtungen finden lassen, die sich einander hinsichtlich bestimmter Merkmale ähnlich sind. Bei Skalierungsverfahren fragt man danach, ob bestimmten Variablen eine gemeinsame Dimension zugrunde liegt. Die häufig angewandte Faktorenanalyse kann dabei auch als ein Sonderfall eines Strukturgleichungsmodells aufgefasst werden, und wird darum bereits in diesem Rahmen behandelt. .

Kapitel 2

Multiple Regression

2.1 Drittvariablenkontrolle in Tabellen

Folgende Tabelle zeigt den Zusammenhang zwischen dem Geschlecht und dem persönlichen Bruttoeinkommen erwerbstätiger Personen in Deutschland 1997 (Daten: 50 Prozent Teilstichprobe des SOEP 1997):

Mann j/n	Hohes pers. Bruttoeink		Total
	nein	ja	
nein	525 77.21	155 22.79	680 100.00
ja	425 46.81	483 53.19	908 100.00
Total	950 59.82	638 40.18	1588 100.00

Die Prozentsatzdifferenz dieser Tabelle beträgt

$$53.19 - 22.79 = 30.40,$$

d.h. Männer haben um 30 Prozentpunkte häufiger ein hohes Bruttoeinkommen als Frauen. Man kann argumentieren, dass Männer deswegen ein höheres Einkommen haben, weil Frauen — aus welchen Gründen auch immer — häufiger als Männer Teilzeit arbeiten. Insofern kann es interessant sein, den Zusammenhang zwischen Geschlecht und Einkommen für Voll- und Teilzeitbeschäftigte getrennt zu betrachten. Dazu wird der Zusammenhang für unterschiedliche Teilpopulationen betrachtet:

-> est=Vollzeit

Mann j/n	Hohes pers. Bruttoeink		Total
	nein	ja	
nein	333 69.09	149 30.91	482 100.00
ja	400 45.51	479 54.49	879 100.00
Total	733 53.86	628 46.14	1361 100.00

-> est=Teilzeit

Mann j/n	Hohes pers. Bruttoeink		Total
	nein	ja	
nein	192 96.97	6 3.03	198 100.00
ja	25 86.21	4 13.79	29 100.00
Total	217 95.59	10 4.41	227 100.00

Nun läßt sich die Prozentsatzdifferenz für beide Populationen berechnen:

$$\text{Vollzeit: } 54.49 - 30.91 = 23.58$$

$$\text{Teilzeit: } 13.79 - 3.03 = 10.76$$

Der Zusammenhang zwischen Geschlecht und Einkommen ist bei den Vollzeitbeschäftigten viel stärker als bei den Teilzeitbeschäftigten. Diesen unterschiedlichen Zusammenhang zwischen zwei Variablen (Geschlecht und Einkommen) für unterschiedliche Ausprägungen einer Dritten (Erwerbsstatus) bezeichnet man als „*Interaktionseffekt*“.

Abgesehen von diesem Interaktionseffekt stellen wir fest, dass der Zusammenhang zwischen Geschlecht und Einkommen in *beiden* Populationen kleiner ist, als bei der gemeinsamen Betrachtung der beiden Populationen.

Der durchschnittliche Effekt in beiden Gruppen beträgt:

$$\frac{23.58 \times 1361 + 10.76 \times 227}{1588} = 21.75.$$

2.2. PROZENTSATZDIFFERENZEN UND LINEARE REGRESSION 7

Damit messen wir den Einfluss des Geschlechts auf das Einkommen unter Kontrolle des Erwerbsstatus. Unter Kontrolle bedeutet: Betrachtung des Zusammenhangs unter „Konstanthaltung“ einer anderen Variablen. Hier wird der Zusammenhang zwischen dem Geschlecht und dem Einkommen unter Kontrolle des Erwerbsstatus betrachtet.

2.2 Prozentsatzdifferenzen und lineare Regression

Prozentsatzdifferenzen lassen sich auch durch lineare Regressionsmodelle bestimmen. So ergibt sich die Prozentsatzdifferenz ohne Kontrolle des Erwerbsstatus als b -Koeffizient der einfachen linearen Regression des Einkommens gegen das Geschlecht:

Source	SS	df	MS	Number of obs =	1588
Model	35.9321568	1	35.9321568	F(1, 1586) =	164.83
Residual	345.742906	1586	.217996788	Prob > F =	0.0000
				R-squared =	0.0941
				Adj R-squared =	0.0936
Total	381.675063	1587	.240500985	Root MSE =	.4669

eink_d	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
men	.3039971	.0236784	12.839	0.000	.2575528	.3504415
_cons	.2279412	.0179048	12.731	0.000	.1928215	.2630608

Die mittlere „Prozentsatzdifferenz“ der beiden untersuchten Teilpopulationen ergibt sich aus einer „multiplen“ linearen Regression, bei der der Erwerbsstatus als zusätzliche weitere Variable aufgenommen wurde¹.

Source	SS	df	MS	Number of obs =	1588
Model	51.1094254	2	25.5547127	F(2, 1585) =	122.53
Residual	330.565638	1585	.208558762	Prob > F =	0.0000
				R-squared =	0.1339
				Adj R-squared =	0.1328
Total	381.675063	1587	.240500985	Root MSE =	.45668

eink_d	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
men	.226176	.0248921	9.086	0.000	.1773511	.2750008
vollzeit	.3001918	.0351898	8.531	0.000	.2311685	1.3692152
_cons	.0151581	.0304774	0.497	0.619	-.0446222	.0749384

¹Geringfügige Differenzen ergeben sich aufgrund der Rundung der Prozentzahlen in den Tabellen.

Die beiden getrennten Prozentsatzdifferenzen lassen sich durch eine multiple Regression berechnen, in der zusätzlich zum Geschlecht und Einkommen noch ein sogenannter „Interaktionsterm“ aufgenommen wird. Dieser wird durch Multiplikation des Geschlechts mit dem Erwerbsstatus gebildet. Hierdurch entsteht eine Variable welche 1 ist, für alle Fälle die männlich sind und Vollzeit arbeiten und 0 für alle Anderen. Das Ergebnis einer Regression mit dieser Variable lautet:

Source	SS	df	MS			
Model	51.4937997	3	17.1645999	Number of obs =	1588	
Residual	330.181263	1584	.208447767	F(3, 1584) =	82.34	
Total	381.675063	1587	.240500985	Prob > F =	0.0000	
				R-squared =	0.1349	
				Adj R-squared =	0.1333	
				Root MSE =	.45656	

eink_d	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
men	.107628	.0907779	1.186	0.236	-.0704294	.2856854
vollzeit	.2788256	.0385387	7.235	0.000	.2032334	.3544178
menvoll	.1281808	.094394	1.358	0.175	-.0569695	.3133311
_cons	.030303	.0324464	0.934	0.350	-.0333393	.0939454

Der b -Koeffizient des Geschlechts in diesem Modell gibt den Einfluß des Geschlechts an, wenn alle anderen Variablen 0 sind. Im Beispiel handelt es sich dabei um die Teilzeitbeschäftigten. Bei ihnen beträgt der Effekt des Geschlechts demnach .1076. Der Effekt des Geschlechts für die Vollzeitbeschäftigten ist die Summe der b -Koeffizienten des Geschlechts und des Interaktionsterms:

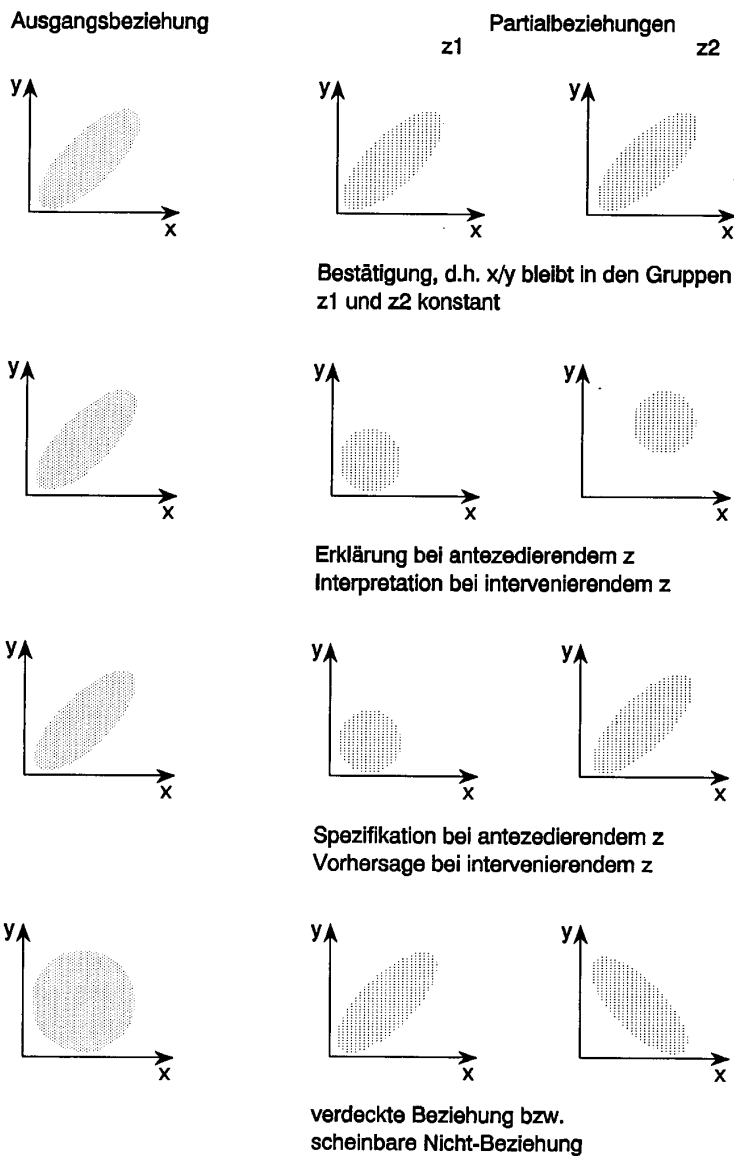
$$.1076 + .1282 = .2358$$

Anders formuliert: Der Effekt des Geschlechts ist für die Vollzeitbeschäftigten um .1282 höher als für die Teilzeitbeschäftigten.

2.3. MÖGLICHE KONSTELLATIONEN BEI DER DRITTVARIABLENKONTROLLE⁹

2.3 Mögliche Konstellationen bei der Drittvariablenkontrolle

Abbildung 5-7: Mögliche Beziehungen zwischen X, Y und Z



2.4 Kontrolle in linearen Regressionsmodellen

Die Koeffizienten der linearen Regression lassen sich im Fall ausschließlich dichotomer Variablen wie Prozentsatzdifferenzen von Kontingenztabellen interpretieren. Der normale Anwendungsfall der Regressionsanalyse sind aber metrische Variablen. Die statistische Kontrolle durch „Konstanthaltung“ wie oben beschrieben ist bei solchen Variablen nicht möglich. Daher soll im folgenden erläutert werden, was statistische Kontrolle im linearen Regressionsmodell bedeutet.

Der einfachste Fall des multiplen linearen Regressionsmodells ist ein Modell mit zwei abhängigen Variablen:

$$Y_i = b_0 + b_1 X_{i1} + b_2 X_{i2} + e_i \quad (2.1)$$

Wie in der einfachen linearen Regression werden die b -Koeffizienten so ausgewählt, dass die Summe der Quadrierten Fehlervariable möglichst klein wird. Eine (unelegante) Art, die b -Koeffizienten zu bestimmen ist die Berechnung folgender einfacher Regressionsmodelle:

$$\text{für } b_1: e_{Y|X_2} = a_1 + b_1 e_{X_1|X_2} + e_i \quad (2.2)$$

$$\text{für } b_2: e_{Y|X_1} = a_2 + b_2 e_{X_2|X_1} + e_i \quad (2.3)$$

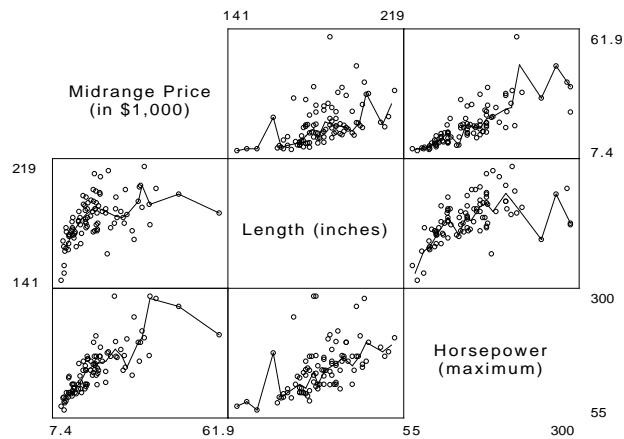
mit $e_{A|B}$ den Residuen einer einfachen Regression der Variable A gegen die Variable B.

Diese Art der Berechnung erfordert die Berechnung einer Serie von einfachen linearen Regressionsmodellen.

Beispiel Gegeben sind Kennwerte von 1993 in den USA neu zugelassenen Autos und zwar

- der mittlere Fahrzeugpreis der Ausstattungsvarianten eines Fahrzeugtyps,
- die Fahrzeuglänge in inch und
- die Stärke des Motors in PS

Daten aus [gopher://jse.stat.ncsu.edu:70/11/jse/data](http://jse.stat.ncsu.edu:70/11/jse/data). Die bivariaten Zusammenhänge zwischen diesen Variablen lassen sich wie folgt darstellen:



Berechnet werden soll der Einfluß der Länge auf den Fahrzeugpreis unter der Kontrolle der PS-Zahl. Hierzu müssen die b -Koeffizienten folgender linearer Regression berechnet werden.

$$\text{Preis} = b_0 + b_{\text{Länge}} \times \text{Länge}_i + b_{\text{PS}} \times \text{PS}_i + e_i \quad (2.4)$$

Die Bestimmung der b -Koeffizienten kann durch folgende Schritte erfolgen:

1. Berechnung der Koeffizienten einer einfachen linearen Regression des Preises gegen die PS-Zahl:

$$\text{Preis} = a_1 + b_1 \times \text{PS}_i + e_i \quad (2.5)$$

2. Berechnung der durch die Schätzung von (2.5) vorhergesagten Preise ($\widehat{\text{Preis}}$)
3. Berechnung der Residuen des Modells aus (2.5)

$$e_{\text{Preis}|\text{PS},i} = \text{Preis}_i - \widehat{\text{Preis}}_i \quad (2.6)$$

4. Berechnung der Koeffizienten einer einfachen linearen Regression der Fahrzeuglänge gegen die PS-Zahl:

$$\text{Länge} = a_2 + b_2 \times \text{PS}_i + e_i \quad (2.7)$$

5. Berechnung der durch die Schätzung von (2.7) vorhergesagten Längen ($\widehat{\text{Länge}}_i$)

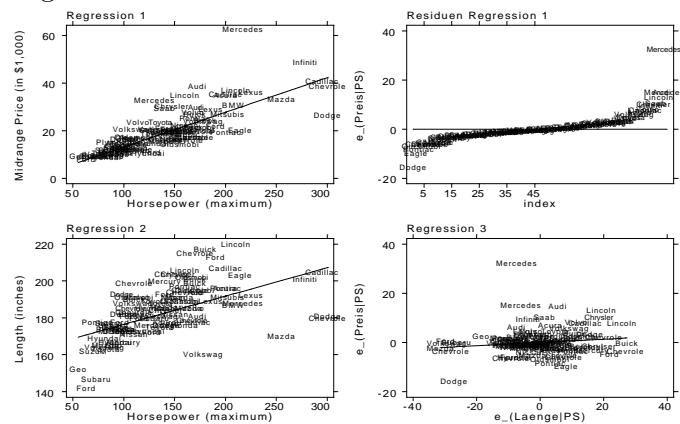
6. Berechnung der Residuen

$$e_{\text{Länge}|PS,i} = \text{Länge}_i - \widehat{\text{Länge}}_i \quad (2.8)$$

7. Berechnung der Koeffizienten einer einfachen Regression der Residuen aus (2.6) gegen die Residuen aus (2.8).

$$e_{\text{Preis}|PS,i} = a_3 + b_{\text{Länge}} \times e_{\text{Länge}|PS,i} + e_i \quad (2.9)$$

Da alle Schritte auf einfachen linearen Regressionen beruhen können sie graphisch dargestellt werden:



Zur Interpretation der Koeffizienten der multiplen Regressionsgleichung muß man sich klarmachen, was die Residuen inhaltlich bedeuten. Die Residuen der Regression des Fahrzeugpreises gegen die PS-Zahl ist die Information des Fahrzeugpreises, die nicht bereits in der PS-Zahl enthalten ist. Man spricht davon, dass die Information über die PS-Zahl aus dem Fahrzeugpreis *herauspartialisiert* wurde. Dasselbe gilt für die Residuen der Regression der Fahrzeuglänge gegen die PS-Zahl.

Der b -Koeffizient der multiplen Regression hat damit folgende Interpretation:

Die Veränderung des — von der PS-Zahl unabhängigen — Fahrzeugpreises für jede Erhöhung der — von der PS-Zahl unabhängigen — Fahrzeuglänge um eine Einheit.

2.5 Etwas Formaler: Lineare Regression mit drei und mehr Variablen

Die b -Koeffizienten des multiplen Regressionsmodells:

$$\hat{Y}_i = b_0 + b_1 X_{i1} + b_2 X_{i2} + \dots + b_{k-1} X_{i,k-1} \quad (2.10)$$

werden so bestimmt, dass die Summe der quadrierten Residuen

$$RSS = \sum (y_i - \hat{Y}_i)^2 = \sum e_i^2 \quad (2.11)$$

kleiner sind als für jeden anderen Set von b -Koeffizienten. Die Bestimmung der b -Koeffizienten über (2.11) wird als *ordinary least squares* (OLS) bezeichnet. Unter der Bedingung, dass die Fehler normal i.i.d. sind besitzt OLS optimale statistische Eigenschaften (BLUE).

Die Koeffizienten können immer analog zur oben beschriebenen Vorgehensweise bestimmt werden. Bei mehr als zwei unabhängigen Variablen ist dies jedoch kaum noch praktikabel. Die Fachliteratur verwendet daher Formeln in Matrixnotation. Das multiple Regressionsmodell aus Gleichung 2 wird in Matrixnotation geschrieben als:

$$\underset{n \times 1}{\mathbf{y}} = \underset{n \times k}{\mathbf{X}} \underset{k \times 1}{\mathbf{b}} + \underset{n \times k}{\mathbf{e}} \quad (2.12)$$

bei n Fällen und $k - 1$ X -Variablen ist \mathbf{y} ein $n \times 1$ Vektor der Y -Variable, \mathbf{X} eine $n \times k$ Matrix der X -Variablen mit einer zusätzlichen Spalte von „lern“, \mathbf{b} ein $k \times 1$ Vektor der geschätzten b -Koeffizienten und \mathbf{e} ein $n \times 1$ Vektor der Residuen.

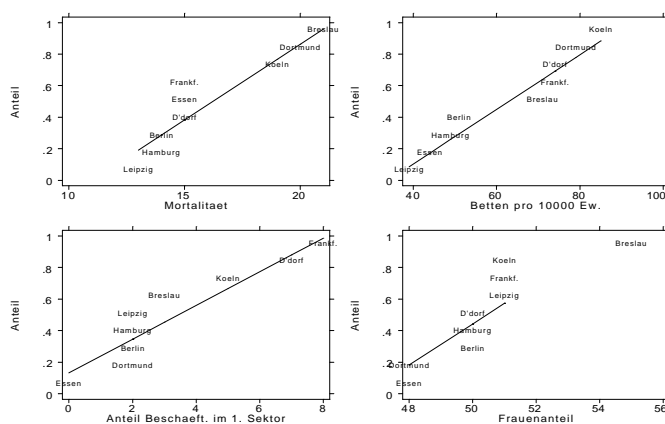
Der Vektor \mathbf{b} dieser Gleichung wird bestimmt durch:

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \quad (2.13)$$

Beispiel Gegeben ist

- die Mortalität (Gestorbene pro 1000 Einwohner)
- der Grad des Ausbaus des Gesundheitswesens indiziert durch die Anzahl der Krankenhausbetten pro 10.000 Einwohner,
- der Modernisierungsgrad, indiziert durch den Prozentsatz der in der Landwirtschaft Beschäftigten
- der Prozentsatz von Frauen in der Bevölkerung

in 9 deutschen Städten um das Jahr 1910. Die Variablen haben folgende Verteilungen:



An der Korrelationsmatrix der Variablen läßt sich ablesen, dass die Mortalität mit dem Frauenanteil, dem Modernisierungsgrad und dem Grad des Ausbaus des Gesundheitswesens ansteigt.

(obs=9)

	mort	pfrau	plandw	pbetten
mort	1.0000			
pfrau	0.3507	1.0000		
plandw	0.0338	0.2786	1.0000	
pbetten	0.7245	0.1966	0.6487	1.0000

Im folgenden sollen die Zusammenhänge jeweils unter Kontrolle der übrigen Variablen betrachtet werden. Hierzu wird die multiple lineare Regression der Mortalität gegen den Grad des Ausbaus des Gesundheitswesens, den Modernisierungsgrad und den Frauenanteil berechnet. Die Daten lassen sich in folgenden Matrizen darstellen:

$$\mathbf{X} = \begin{pmatrix} 1 & 44 & 0 & 48 \\ 1 & 79 & 2 & 48 \\ 1 & 51 & 2 & 50 \\ 1 & 49 & 2 & 50 \\ 1 & 74 & 7 & 50 \\ 1 & 85 & 5 & 51 \\ 1 & 74 & 8 & 51 \\ 1 & 39 & 2 & 51 \\ 1 & 71 & 3 & 55 \end{pmatrix} \quad \mathbf{y} = \begin{pmatrix} 15 \\ 20 \\ 14 \\ 14 \\ 15 \\ 19 \\ 15 \\ 13 \\ 21 \end{pmatrix}$$

Die gesuchten b -Koeffizienten werden durch einsetzen in (2.13) ermittelt:

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} 9 & & & & \\ 566 & 37918 & & & \\ 31 & 2184 & 163 & & \\ 454 & 28607 & 1576 & 22936 & \end{pmatrix}$$

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} 79.84129 & & & & \\ -.03084508 & .0007437 & & & \\ .40866318 & -.00307826 & .03202503 & & \\ -1.5700041 & -.00010551 & -.00645033 & .03169541 & \end{pmatrix}$$

$$\mathbf{X}'\mathbf{y} = \begin{pmatrix} 146 \\ 9473 \\ 505 \\ 7382 \end{pmatrix}$$

$$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \begin{pmatrix} -18.762332 \\ .2082778 \\ -.93922149 \\ .49799915 \end{pmatrix}$$

Der Regressionsoutput eines Datenanalyseprogramms (Stata) lautet:

Source	SS	df	MS			
Model	67.1936427	3	22.3978809	Number of obs =	9	
Residual	2.36191285	5	.472382569	F(3, 5) =	47.41	
Total	69.5555556	8	8.69444444	Prob > F =	0.0004	
				R-squared =	0.9660	
				Adj R-squared =	0.9457	
				Root MSE =	.6873	

mort	Coef.	Std. Err.	t	P> t	Beta
pfrau	.4979992	.1223616	4.070	0.010	.3493146
plandw	-.9392215	.1229962	-7.636	0.001	-.8444153
pbetten	.2082778	.0187433	11.112	0.000	1.203625
_cons	-18.76233	6.141306	-3.055	0.028	.

Hinweis zur Interpretation: Unter Kontrolle der Bettenzahl und des Frauenanteils senkt der Modernisierungsgrad die Sterblichkeit. Der Ausbau der Gesundheitswesens reduziert die Mortalität dagegen nicht.

Die Interpretation der Konstante dieses Modells ist wenig sinnvoll, da kaum von Interesse sein kann, wie hoch die Mortalität in einer Gesellschaft ohne Frauen (und ohne Landwirte und ohne Krankenhausbetten) ist. Die Konstante kann jedoch nur sinnvoll interpretiert werden, wenn die Daten zuvor zentriert werden, d.h. der Mittelwert von jeder unabhängigen Variable abgezogen wird.

Da es sich um Aggregatdaten handelt sind die ausgewiesenen Standardfehler und Signifikanztests ungültig.

2.6 Das allgemeine lineare Modell, oder: Was tun bei kategorialen unabhängigen Variablen?

Das allgemeine lineare Modell integriert elementarstatistische Verfahren, varianzanalytische Verfahren sowie multiple Korrelations- und Regressionsrechnung. Im Kern des Modells steht die multiple lineare Regression. Diese wird im allgemeinen linearen Modell dahingehend erweitert, dass auch kategoriale Variablen berücksichtigt werden können. Außerdem entfällt die Beschränkung auf nur eine abhängige Variable (siehe dazu auch den Auszug aus Holm 1979: 20-24 am Ende dieses Abschnitts). Die folgende Darstellung begrenzt sich auf die Integration kategorialer unabhängiger Variablen in die multiple lineare Regression auf eine abhängige metrische Variable. Durch die Integration der kategorialen Variablen können Fragestellungen der Varianzanalyse (ANOVA) und der Kovarianzanalyse (ANCOVA) im Rahmen der multiplen Regression bearbeitet werden. Nicht behandelt wird die Vorgehensweise bei mehreren abhängigen metrischen Variablen sowie für kategoriale abhängige Variablen. Zur Vorgehensweise bei kategorialen abhängigen Variablen siehe Abschnitt 4.

2.6.1 Codierung einer dichotomen unabhängigen Variablen

Gegeben ist das einfache lineare Regressionsmodell

$$\hat{Y} = b_0 + b_1 X_{i,1} \quad (2.14)$$

Die Variable $X_{i,1}$ ist eine dichotome Variable, d.h. eine Variable mit lediglich zwei Ausprägungen. Wenn die Werte von $X_{i,1}$ so codiert werden, dass eine Ausprägung den Wert 0, die andere Ausprägung den Wert 1 erhält ergeben sich für das Modell in (2.14) folgende vorhergesagten Werte:

- für die Ausprägung 0

$$\begin{aligned} \hat{Y} &= b_0 + b_1 \times 0 \\ \hat{Y} &= b_0 \end{aligned} \quad (2.15)$$

- für die Ausprägung 1

$$\begin{aligned} \hat{Y} &= b_0 + b_1 \times 1 \\ \hat{Y} &= b_0 + b_1 \end{aligned} \quad (2.16)$$

2.6. DAS ALLGEMEINE LINEARE MODELL, ODER: WAS TUN BEI KATEGORIALEN UNABHÄNGIGEN VARIABLEN

Interpretation: Angenommen $X_{i,1}$ sei das Geschlecht mit $X_{i,1} = 0$ für die Männer und $X_{i,1} = 1$ für die Frauen. Die Regressionskonstante des Modells in (2.14) gibt den Vorhersagewert von Y_i wenn alle Variablen im Modell = 0 sind, also für die Männer.

Der Regressionskoeffizient von $X_{i,1}$ gibt wieder, um wieviel größer oder kleiner der Vorhersagewert für die Frauen als für die Männer ist. Regressionskonstante + Regressionskoeffizient ergibt den Vorhersagewert von Y_i für die Frauen.

Der Vorhersagewert für die Männer entspricht dem arith. Mittelwert von Y_i für die Männer, derjenige für die Frauen dem arith. Mittelwert von Y_i für die Frauen. Damit drückt der Regressionskoeffizient auch aus, wie stark sich der Mittelwert der Frauen vom Mittelwert der Männer unterscheidet. Der Signifikanztest des b-Koeffizienten ist damit identisch mit dem herkömmlichen Mittelwertstest.

2.6.2 Codierung einer polytomen nominalskalierten unabhängigen Variablen

Gegeben sei folgende polytome nominalskalierte Variable:

bildung	Freq.	Percent	Cum.
HS/VS	1331	39.85	39.85
Reals.	939	28.11	67.96
FHSReife	93	2.78	70.75
Abitur	436	13.05	83.80
Sonst.	541	16.20	100.00
Total	3340	100.00	

Die Verwendung der Variable Schulbildung in einem Regressionsmodell macht in dieser Form keinen Sinn. Wie beim Geschlecht könnte man aber z.B. das Abitur von allen anderen Ausprägungen der Schulbildung unterscheiden. In Datenanalyseprogrammen lassen sich einfach entsprechende Variablen bilden:

- Stata

```
. gen abi=bildung==4
```
- SPSS

```
. comp abi = 0
. if (abi=1) abi = 1
```

Hierdurch erhält man:

abi	Freq.	Percent	Cum.
0	2904	86.95	86.95
1	436	13.05	100.00
Total	3340	100.00	

Die neue Variable läßt sich wie im ersten Beispiel in die Regression einführen. Insgesamt lassen sich fünf derartige *Kontraste* bilden:

1. Hauptschule/Volksschule vs. Rest
2. Mittlere Reife vs. Rest
3. FHS-Reife vs. Rest
4. Abitur vs. Rest
5. Sonstige vs. Rest

Diese Variablen können wieder auf einfache Weise mit Datenanalyseprogrammen erzeugt werden

- Stata
 - . tab bildung, gen(bil_)
- SPSS
 - . comp bil_1 = 0
 - . if (bildung=1) bil_1 = 1
 - . comp bil_2 = 0
 - . if (bildung=2) bil_2 = 1
 - . comp bil_3 = 0
 - . if (bildung=3) bil_3 = 1
 - . comp bil_4 = 0
 - . if (bildung=4) bil_4 = 1
 - . comp bil_5 = 0
 - . if (bildung=5) bil_5 = 1

Hierdurch erhält man:

```
-> tabulation of bil_1
bildung==HS |
```

2.6. DAS ALLGEMEINE LINEARE MODELL, ODER: WAS TUN BEI KATEGORIALEN UNABHÄNGIGKEITEN

/VS	Freq.	Percent	Cum.
0	2009	60.15	60.15
1	1331	39.85	100.00
Total	3340	100.00	

-> tabulation of bil_2

bildung==Re	Freq.	Percent	Cum.
als.			
0	2401	71.89	71.89
1	939	28.11	100.00
Total	3340	100.00	

-> tabulation of bil_3

bildung==FH	Freq.	Percent	Cum.
SReife			
0	3247	97.22	97.22
1	93	2.78	100.00
Total	3340	100.00	

-> tabulation of bil_4

bildung==Ab	Freq.	Percent	Cum.
itur			
0	2904	86.95	86.95
1	436	13.05	100.00
Total	3340	100.00	

-> tabulation of bil_5

bildung==So	Freq.	Percent	Cum.
nst.			
0	2799	83.80	83.80
1	541	16.20	100.00
Total	3340	100.00	

Alle fünf Variablen sind dichotom und können somit problemlos in das Regressionsmodell aufgenommen werden. Zur Vermeidung von linearen Abhängigkeiten muß jedoch eine der Variablen eliminiert werden. Welche ist dabei arbiträr. Im folgenden wird die erste Variable nicht in das Modell aufgenommen. Sie fungiert als sog. *Referenzkategorie*.

Beispiel Mit den Daten aus Abschnitt 2.2 wird eine lineare Regression des Einkommens gegen das Geschlecht, den Erwerbsstatus und die Bildung berechnet. Der Modellansatz lautet formal:

$$\widehat{\text{Eink}}_i = b_0 + b_1 \text{Man}_i + b_2 \text{Vollzeit}_i + \underbrace{b_3 \text{bil}_2 + \dots + b_6 \text{bil}_6}_{\text{Bildung}}. \quad (2.17)$$

Die Berechnung des Modells ergibt:

Source	SS	df	MS			
Model	2.1053e+09	6	350889886	Number of obs =	1588	
Residual	7.1042e+09	1581	4493510.67	F(6, 1581) =	78.09	
Total	9.2096e+09	1587	5803137.80	Prob > F =	0.0000	
				R-squared =	0.2286	
				Adj R-squared =	0.2257	
				Root MSE =	2119.8	

eink	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
men	1104.815	116.8761	9.453	0.000	875.5663	1334.063
vollzeit	1832.215	164.5809	11.133	0.000	1509.395	2155.035
bil_2	124.2877	133.2269	0.933	0.351	-137.0323	385.6076
bil_3	1682.959	295.2269	5.701	0.000	1103.881	2262.036
bil_4	1302.603	159.9724	8.143	0.000	988.8226	1616.383
bil_5	-126.0901	171.7233	-0.734	0.463	-462.9194	210.7393
_cons	1524.965	158.5208	9.620	0.000	1214.032	1835.898

Interpretation: Der durchschnittliche Verdienst von Realschülern ist um etwa 124,- DM höher als derjenige der Hauptschüler. Der durchschnittliche Verdienst der Abiturienten ist um etwa 1303 DM höher als derjenige der Hauptschüler. Damit verdienen die Abiturienten auch $1303 - 124 = 1179$ DM mehr als die Realschüler. Nochmals $1683 - 1303 = 380$ DM mehr als die Abiturienten verdienen Befragte mit der Fachhochschulreife. Am wenigsten verdienen die Befragte mit Anderer oder Keiner Bildung, nämlich 126 DM weniger als die Hauptschüler, bzw. $(-126) - 1683 = 1809$ DM weniger als die Befragten mit Fachhochschulreife.

Die Konstante des Modells gibt das durchschnittliche Einkommen derjenigen Personen wieder, welche auf allen enthaltenen Variablen die Ausprägung 0 haben. Dies sind alle Teilzeitbeschäftigten Frauen mit Hauptschulabschluß.

Erweiterte Fragestellung: Hat sich die Einkommensungleichheit zwischen Männern und Frauen derart verringert, dass jüngere Frauen heute weniger von der Einkommensungleichheit betroffen sind als ältere?

Diese Frage bedeutet, dass das Geschlecht mit zunehmenden Alter einen immer stärkeren Einfluß, bzw. mit abnehmenden Alter einen immer schwächeren

2.6. DAS ALLGEMEINE LINEARE MODELL, ODER: WAS TUN BEI KATEGORIALEN UNABHÄNGIGEN

ren Einfluß haben sollte. Der Frage kann durch einen Interaktionsterm zwischen Geschlecht und Alter nachgegangen werden. Interaktionsterme werden durch Multiplikation der Variablen die „interagieren“ gebildet². Allerdings sollte man metrische Variablen „zentrieren“ da sonst die „Haupteffekte“ nur schwer interpretiert werden können.

Das entsprechende Regressionsmodell ergibt:

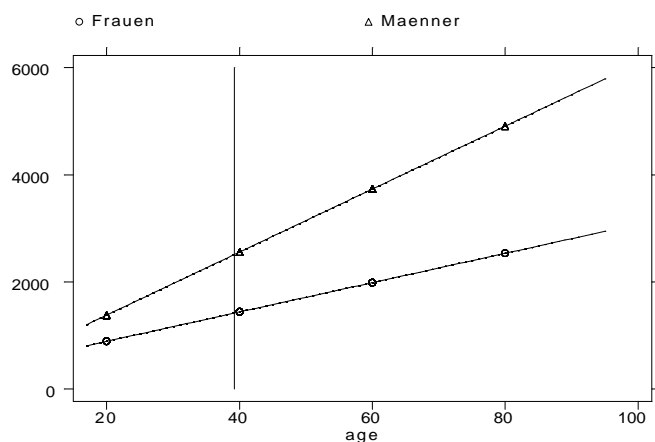
Source	SS	df	MS			
Model	2.4833e+09	8	310412612	Number of obs =	1588	
Residual	6.7263e+09	1579	4259834.58	F(8, 1579) =	72.87	
Total	9.2096e+09	1587	5803137.80	Prob > F =	0.0000	
				R-squared =	0.2696	
				Adj R-squared =	0.2659	
				Root MSE =	2063.9	

eink	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
men	1094.142	114.0015	9.598	0.000	870.5321	1317.753
vollzeit	1869.309	161.6086	11.567	0.000	1552.319	2186.299
bil_2	278.3102	131.2131	2.121	0.034	20.93992	535.6804
bil_3	1827.468	288.1785	6.341	0.000	1262.215	2392.721
bil_4	1402.767	156.1954	8.981	0.000	1096.395	1709.14
bil_5	-67.35257	167.3854	-0.402	0.687	-395.6736	260.9685
age	27.37369	7.733217	3.540	0.000	12.20524	42.54214
menage	31.41018	10.19404	3.081	0.002	11.41489	51.40546
_cons	1418.288	155.8435	9.101	0.000	1112.606	1723.97

Interpretation: Mit jedem Lebensjahr vergrößert sich die Einkommenslücke zwischen Männern und Frauen um die Höhe des Interaktionseffekts, also um DM 31. Bei Personen mit dem Alter 0 verdienen Männer 1094 DM mehr als Frauen. Da die Altersvariable zentriert wurde steht das Alter 0 für das durchschnittliche Alter (der erwerbstätigen Personen), i.e. etwa 39 Jahre. Im Alter von 19, also bei den 20 Jahre jüngeren Personen beträgt die Einkommensungleichheit „nur“ noch $1094 - 20 \times 31 = 474$ DM.

Zur Interpretation von Interaktionseffekten sind Plots der vorhergesagten Werte für unterschiedliche Variablenmuster hilfreich (*Conditional Effects Plot*). Im folgenden Beispiel wurden die vorhergesagten Werte für Teilzeit erwerbstätige Hauptschüler mit unterschiedlichem Geschlecht und Alter berechnet. Der Effekt des Alters ist als Steigung der Linien, der Effekt des Geschlechts als Abstand zwischen den Linien zu erkennen.

²Bei Interaktionen von „Dummy-Variablen“ mit metrischen Variablen wird jede „Dummy-Variable“ mit der metrischen Variable multipliziert.



2.7 Literatur

- Lehrbücher multiple Regression
 - Backhaus, Klaus, B. Erichson, W. Plinke, und R. Weiber, 1994: *Multivariate Analysemethoden. Eine anwendungsorientierte Einführung*. 7. Aufl. Berlin usw.: Springer.
 - Fox, John, 1997: *Applied regression analysis, linear models, and related methods*. Thousand Oaks usw. Sage.
 - Hamilton, Lawrence C., 1992: *Regression with Graphics. A Second Course in Applied Statistics*. Belmont: Wadsworth.
- Hilfreich für das allgemeine Verständnis
 - Hellevik, Ottar, 1984: *Introduction to Causal Analysis. Exploring Survey Data by Crosstabulation*. London: Allen & Unwin.
- Spezialliteratur zur Interpretation von Interaktionseffekten
 - Aiken, Leona S. und Stephen W. West, 1991: *Multiple Regression: Testing and Interpreting Interactions*. Newbury Park usw.: Sage.
 - Kühnel, Steffen M., 1996: Gruppenvergleiche in linearen und logistischen Regressionsmodellen, *ZA-Informationen* 39: 130-160.
- Einführung in Matrix Algebra
 - Nambodiri, Krishnan, 1984: *Matrix Algebra. An Introduction*. Beverly Hills usw.: Sage.
- Eine exemplarische Anwendung

- Falter, Jürgen W, Andreas Link, Jan-Bernd Lohmöller, Johann de Rijke und Siegfried Schumann, 1983: Eine empirische Analyse des Beitrags der Massenerwerbslosigkeit zu den Wahlerfolgen der NSDAP 1932 und 1933, Kölner Zeitschrift für Soziologie und Sozialpsychologie 53: 525-554.

Kapitel 3

Strukturgleichungsmodelle

3.1 Was sind Pfaddiagramme

Pfaddiagramme sind bildhafte Darstellungen eines System simultaner Gleichungen. Sie dienen dazu, die behaupteten Beziehungen zwischen Variablen bildhaft darzustellen. Dabei bedeutet ...

- ein gerader Pfeil: „verursacht“
- ein gebogener Pfeil: „korreliert, nicht verursacht“

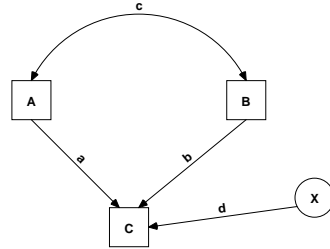
Begriffsdefinitionen

1. *Ursache*: eine Veränderung am Ursprung des Pfeiles führt zu einer Veränderung an der Spitze des Pfeiles
2. *Exogene Variable*: Variable, die keine kausalen Inputs erhält.
3. *Endogene Variable*: Variable, die mindestens einen kausalen Input erhält.

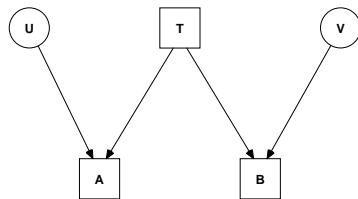
Wichtig

- Alle exogenen Variablen müssen mit gebogenen Pfeilen verbunden werden - es sei denn, die Korrelationen sind 0, bzw. werden als 0 vorausgesetzt.
- Endogene Variablen niemals mit gebogenen Pfeilen verbinden
- Residualpfeile zeigen nicht auf exogene Variablen
- jede endogene Variable muß mit einem Residualpfeil versehen werden

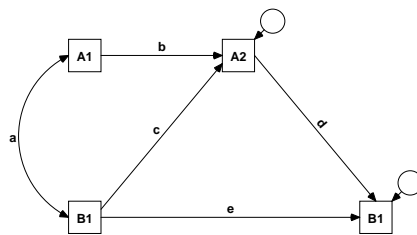
Beispiele



A = Intelligenz Mutter
 B = Intelligenz Vater
 C = Intelligenz Kind
 X = andere Faktoren



A = Intelligenztest A
 B = Intelligenztest B
 T = Intelligenz
 U, V = andere Faktoren



Ereignisse über die Zeit
 (zusätzliche Einflüsse bleiben oft ohne Bezeichnung: „Residualpfeil“)

3.2 Standardisierte Pfadkoeffizienten

Die Aufgabe der Pfadanalyse ist es, den Pfeilen „Gewichte“ zuzuweisen. Je höher das Gewicht eines Pfeiles, desto stärker der kausale Einfluß, den er repräsentiert. Die Zuweisung der Gewichte kann durch multiple Regressionsanalysen erfolgen. Das Gewicht eines Pfeiles entspricht dem standardisierten Regressionskoeffizienten (β -Gewicht) einer multiplen linearen Regression der endogenen Variable gegen *alle* auf sie einwirkenden Variablen.

Im Pfadmodell des Beispiels 3 müssen folgende lineare Regressionsanalysen berechnet werden

für die Pfeile b und c

$$A_2 = b_0 + b_1 A_1 + b_2 B_1 + e \quad b = b_1 \times \frac{s_{A_1}}{s_{A_2}} \quad (3.1)$$

$$c = B_2 \times \frac{s_{B_1}}{s_{A_2}} \quad (3.2)$$

für die Pfeile d und e

$$B_2 = b_0 + b_1 A_2 + b_2 B_1 + e \quad d = b_1 \times \frac{s_{A_2}}{s_{B_3}} \quad (3.3)$$

$$e = B_2 \times \frac{s_{B_1}}{s_{B_3}} \quad (3.4)$$

für den Pfeil a

$$A_1 = b_0 + b_1 B_1 + e \quad a = b_1 \times \frac{s_{B_1}}{s_{A_1}} = r_{A_1 B_1} \quad (3.5)$$

Hinweis: Die Umrechnung der b-Koeffizienten linearer Regressionsanalysen (b_k) in standardisierte Regressionskoeffizienten (β_k) und vice versa erfolgt durch:

$$\beta_k = b_k * \frac{s_{UV}}{s_{AV}} \Leftrightarrow b_k = \beta_k * \frac{s_{AV}}{s_{UV}} \quad (3.6)$$

Bei Berechnung der linearen Regression mit z-standardisierten Variablen ergeben sich die standardisierten Regressionskoeffizienten unmittelbar als b-Koeffizienten.

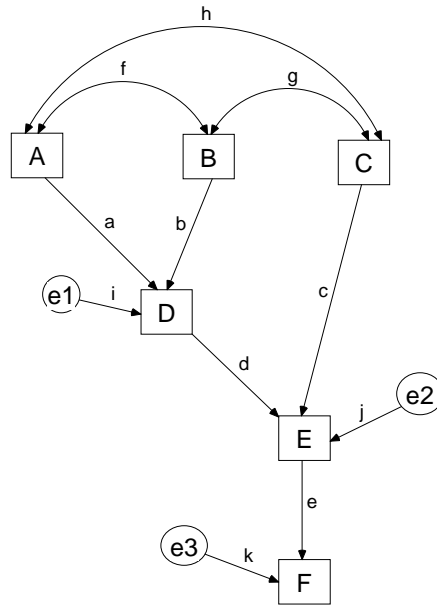
3.3 Wrights Rules

Wenn eine Situation in Form eines wahren Pfaddiagramms präsentiert werden kann, so ist die Korrelation zweier Variablen definiert durch die Summe des Betrags aller *erlaubten Pfade* zwischen diesen Variablen. Ein erlaubter Pfad ist ein Pfad, der folgenden Regeln gehorcht:

- keine Schleifen
- kein Rückwärtsgehen, nachdem man vorwärts gegangen ist
- höchstens ein gebogener Pfeil

Der Betrag eines erlaubten Pfades ergibt sich aus der Multiplikation der Pfadkoeffizienten.

Beispiel



Exogene Variablen:
Endogene Variablen:

$$r_{AD} = a + fb$$

$$r_{AB} = f$$

$$r_{BC} =$$

$$r_{AC} =$$

$$r_{AE} = hc + ad + fbd$$

$$r_{BE} =$$

$$r_{CE} =$$

$$r_{AF} =$$

Übung Bitte setzen Sie die fehlenden Angabe ein.

3.4 Die Berechnung der Pfadkoeffizienten aus der Korrelationsmatrix

Anstatt die Pfadkoeffizienten mit linearen Regressionsmodellen zu berechnen können die Pfadkoeffizienten auch aus der Korrelationsmatrix berechnet werden:

Beispiel Für das erste Pfaddiagramm auf Seite 26 liegt folgende Korrelationsmatrix vor:

	A	B	C
A	1.0	.50	.65
B		1.0	.70
C			1.0

3.5. DIE BERECHNUNG DER PFADKOEFFIZIENTEN AUS DER KOVARIANZMATRIX 29

Durch die Anwendung der Regeln aus Abschnitt 3.3 kann die Korrelationsmatrix in „*Modellparametern*“ ausgedrückt werden:

$$r_{AB} = c = .5 \quad (3.7)$$

$$r_{AC} = a + cb = .65 \quad (3.8)$$

$$r_{BC} = b + ca = .7 \quad (3.9)$$

c in (3.8)

$$.65 = a + .5b$$

c in (3.9)

$$.7 = .5a + b$$

in Matrixnotation

$$\begin{pmatrix} .65 \\ .7 \end{pmatrix} = \begin{pmatrix} 1 & .5 \\ 5 & 1 \end{pmatrix} \times \begin{pmatrix} a \\ b \end{pmatrix}$$

bzw.

$$\mathbf{Y} = \mathbf{Xb}$$

$$\mathbf{b} = \mathbf{X}^{-1}\mathbf{Y}$$

$$= \begin{pmatrix} 1.333 & .667 \\ .667 & 1.333 \end{pmatrix} \times \begin{pmatrix} .65 \\ .7 \end{pmatrix} = \begin{pmatrix} .4 \\ .5 \end{pmatrix}$$

3.5 Die Berechnung der Pfadkoeffizienten aus der Kovarianzmatrix

Statt aus der Korrelationsmatrix lassen sich die Pfadkoeffizienten auch aus der Kovarianzmatrix berechnen. Hierzu müssen die Varianzen und Kovarianzen der Kovarianzmatrix ebenfalls in *Modellparametern* ausgedrückt werden. Modellparameter sind alle Pfadkoeffizienten und alle Varianzen und Kovarianzen der exogenen Variablen.

Um die Kovarianzmatrix in Modellparametern auszudrücken verwendet man elementare Regeln der *Kovarianzalgebra*:

$$COV(X, a) = 0 \quad \text{null rule} \quad (3.10)$$

$$COV(cX, Y) = c * COV(X, Y) \quad \text{constant rule} \quad (3.11)$$

$$COV(X, Y + Z) = COV(X, Y) + COV(X, Z) \quad \text{sum rule} \quad (3.12)$$

Beispiel Die Kovarianz zwischen A und C im ersten Pfaddiagramm auf Seite 26 kann mit Hilfe der Kovarianzalgebra wie folgt ermittelt werden:

$$COV(A, C) = COV(A, aA + bB + dX)$$

sum rule

$$= COV(A, aA) + COV(A, bB) + COV(A, dX)$$

constant rule

$$\begin{aligned} &= aCOV(A, A) + bCOV(A, B) + dCOV(A, X) \\ &= aVAR(A) + bc \end{aligned}$$

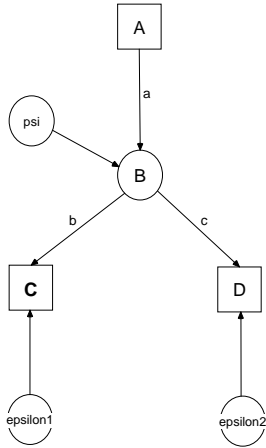
Beachten Sie, dass im Fall von z-Standardisierten Variablen $VAR(A) = 1$ ist und die Anwendung der Kovarianzalgebra dann zum selben Ergebnis wie die Anwendung von „Wright-Rules“ führt.

Die Kovarianz zwischen B und C kann entsprechend entwickelt werden. Danach kann das Gleichungssystem aufgelöst werden und man erhält die Pfadkoeffizienten. Allerdings handelt es sich dann um unstandardisierte Pfadkoeffizienten. Da die Kovarianzmatrix in der Hauptdiagonalen die Varianz enthält können diese jedoch einfach nach (3.6) in standardisierte Pfadkoeffizienten umgerechnet werden. Die Umrechnung von standardisierten Pfadkoeffizienten auf der Basis der Informationen der Korrelationsmatrix ist dagegen nicht möglich. In Strukturgleichungsmodellen wird darum in der Regel die Kovarianzmatrix zur Berechnung der Koeffizienten verwendet.

3.6 Pfadkoeffizienten für Pfadmodelle mit latenten Variablen

Beispiel Gegeben ist folgendes Pfaddiagramm

3.6. PFADKOEFFIZIENTEN FÜR PFADMODELLE MIT LATENTEN VARIABLEN 31



mit der beobachteten Korrelationsmatrix

	A	C	D
A	1.0	.50	.65
C		1.0	.70
D			1.0

Variable B wurde nicht beobachtet!

Die Korrelationsmatrix lässt sich wie folgt in Modellparametern ausdrücken:

$$r_{AC} = ab = .2 \quad (3.13)$$

$$r_{AD} = ac = .24 \quad (3.14)$$

$$r_{CD} = bc = .3 \quad (3.15)$$

Die Auflösung des Gleichungssystems erfolgt durch $\frac{(3.13) \times (3.14)}{(3.15)}$:

$$\begin{aligned} \frac{ab \times bc}{ac} &= b^2 \\ &= \frac{.2 * .3}{.24} = .25 \\ b &= .5 \end{aligned}$$

b in (3.13)

$$\begin{aligned} .5a &= .2 \\ a &= \frac{.2}{.5} \\ &= .4 \end{aligned}$$

a in (3.14)

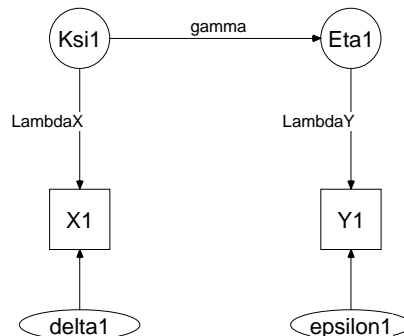
$$\begin{aligned} .4c &= .24 \\ c &= \frac{.24}{.4} \\ &= .6 \end{aligned}$$

Nicht immer können die Koeffizienten in Modellen mit latenten Variablen so einfach berechnet werden. In der Praxis wird daher ein iteratives Verfahren angewandt. Ob die Berechnung der Pfadkoeffizienten möglich ist oder nicht, hängt davon ab, ob das Model *gerade identifiziert*, *überidentifiziert* oder *unteridentifiziert* ist. In der Praxis sind überidentifizierte Modelle die Regel. Bei diesen Modellen wird die Korrelationsmatrix (bzw. die Kovarianzmatrix) nicht exakt reproduziert. Es werden daher diejenigen Koeffizienten ausgewählt, mit denen die Korrelationsmatrix/Kovarianzmatrix möglichst genau reproduziert wird. Maßzahl für die Genauigkeit der Reproduktion ist der Likelihood-Ratio χ^2 .

Unteridentifizierte Modelle sind nicht lösbar.

3.7 Meßfehler und ihre Konsequenzen

Gegeben Sei folgendes Pfadmodell



Das Pfaddiagramm repräsentiert zwei Meßgleichungen (Gleichungen (3.16) und (3.17)) und eine Strukturgleichung Gleichung (3.18):

$$X1 = \lambda X * \xi1 + \delta1 \quad (3.16)$$

$$X2 = \lambda Y * \eta1 + \epsilon \quad (3.17)$$

$$\eta1 = \gamma * \xi + \zeta \quad (3.18)$$

Alle Variablen des Modells sind standardisiert. Wenn

$$\begin{aligned}\gamma &= .5 & \lambda X &= .8 \\ \lambda Y &= .8\end{aligned}$$

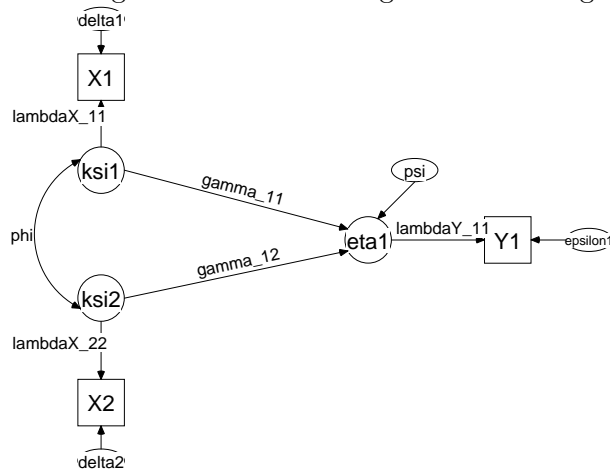
beträgt die beobachtete Korrelation der manifesten Variablen (Wrights Rules):

$$\begin{aligned}r_{X_1 Y_1} &= \lambda X * \gamma * \lambda Y \\ &= r_{X_1 Y_1} = .8 * .5 * .8 \\ &= r_{X_1 Y_1} = .32\end{aligned}$$

Bei Verwendung der beobachteten Werte wird die „wahre“ Korrelation demnach um fast die Hälfte unterschätzt. Die „wahren“ Zusammenhänge sind sehr viel stärker als die beobachteten. Man kann darum argumentieren, dass die Schätzung des Zusammenhangs mit manifesten Variablen konservativ ist. Allerdings beruht diese Argumentation auf zwei Annahmen:

1. Der Fehler in ξ und η ist *zufällig*.
2. Es gibt lediglich 2 latente Variablen

Wenn diese Annahmen nicht zutreffen, z.B. weil δ und ϵ korreliert sind, oder weil mehrere Konstrukte eingeschlossen sind, so sind die Auswirkungen des Meßfehlers schwieriger zu beurteilen. Folgende Abbildung zeigt ein Beispiel:



In nachfolgender Tabelle sind die Koeffizienten unter unterschiedlichen Konstellationen eingetragen. Daran läßt sich ablesen, dass die beobachteten Koeffizienten die „wahren“ Werte auch überschätzen können ($\hat{\gamma}_{12}$ in Zeile 1 und

3). Das Ausmaß der Verzerrung hängt von der Reliabilität der Messung *und* von der Kovarianz der exogenen Variablen ab.

λX_{11}	λX_{22}	λY_{11}	γ_{11}	γ_{12}	ϕ	$\hat{\gamma}_{11}$	$\hat{\gamma}_{12}$	$\hat{\phi}$
.7	.7	.7	.6	0	.5	.28	.08	.25
.6	.9	.7	.86	.54	-.40	.31	.27	-.22
.6	.9	.7	.50	.00	-.60	.16	-.13	-.32

3.8 Probleme linearer Strukturgleichungsmodelle

Die Parameter der Gleichungen von Strukturgleichungsmodellen können durch OLS oder durch eine andere Methode berechnet werden. Wird OLS verwendet wird vorausgesetzt, dass die Variablen *ohne* Fehler gemessen wurde. In Strukturgleichungsmodellen mit latenten Variablen kann diese Annahme aufgegeben werden. Diesem unbestreitbaren Vorteil stehen einige Probleme gegenüber.

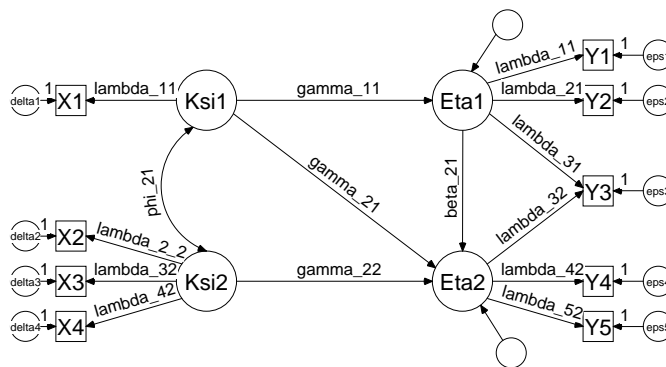
- Die Korrektur der Meßfehler setzt voraus, dass das durch das Pfadmodell hypothetisierte Modell wahr ist.
- Die Meßmodelle werden zusammen mit der eigentlich interessierenden Theorie entwickelt. Es besteht die Gefahr, dass die latenten Variablen so gebildet werden, dass das für die Kerntheorie günstigste Meßmodell erzeugt wird.
- In Strukturgleichungsmodellen mit latenten Variablen werden die Parameter der Gleichungen ausschließlich anhand der *Kovarianzmatrix* bestimmt. Man spricht von der Kovarianzmatrix als *basic building block*. Befürchtungen über Informationsverlust gegenüber der Berechnung mit OLS sind zwar grundsätzlich unberechtigt, da dieselbe Datenreduktion auch bei OLS vorgenommen wird¹. Erst im Rahmen der Residuen- und Modelldiagnostik werden Informationen benötigt, welche in der Kovarianzmatrix nicht enthalten sind. *Vor* der Berechnung von Strukturgleichungsmodellen sollten darum folgenden Problemen Aufmerksamkeit geschenkt werden:
 - *Nichtlinearitäten*: Lösung durch Transformation von Variablen und Berechnung von Σ mit den transformierten Variablen. Schwieriger ist die Verwendung multipler Terme wie z.B. in $Y = a + b_1 * X1 + b_2 * X1^2$.

¹Bei zentrierten Variablen enthält der Term $\mathbf{X}'\mathbf{X}$ in (2.13) die Information über die Variation und Kovariation der Variablen in \mathbf{X} . Die Multiplikation der Elemente von $\mathbf{X}'\mathbf{X}$ mit $1/(n - 1)$ entspricht der Kovarianzmatrix.

- *Outliers*: Lösung wie unter OLS.
- *Verteilungsannahmen*: Die Modelle stehen unter der Annahme multivariater Normalverteilung. Allerdings werden die Parameterschätzungen als *ziemlich robust* gegen Verletzungen dieser Annahme bezeichnet. Zur Korrektur der Signifikanztests liegen einige spezielle Ansätze vor.
- *Kategorisierung metrischer Variablen*: Konzeptuell metrische Variablen — insbesondere Einstellungen — werden oft „kategorisiert“ gemessen, also z.B. auf einer dreier- oder vierer-Skala. Dies führt zu einigen Problemen. In der Praxis sind Variablen mit mindestens fünf Kategorien als „unproblematisch“ anzusehen. Variablen mit weniger Kategorien sollten sparsam eingesetzt werden.
- *kategoriale Variablen*: Konzeptuell kategoriale Variablen können nur als exogene Variablen verwendet werden.

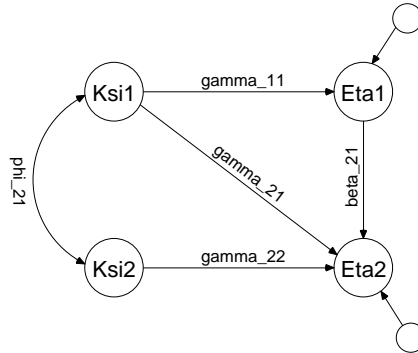
3.9 Grundschemata linearer Strukturgleichungsmodelle

In Pfadmodellen mit latenten Variablen werden latente Variablen mit Kreisen, manifeste Variablen mit Vierecken gekennzeichnet. Folgendes Pfadmodell zeigt ein Beispiel in Lisrel-Notation. Lisrel verwendet griechische Buchstaben für alle latenten Variablen sowie für die Koeffizienten.



Dieses Pfadmodell läßt sich in drei kleinere Einheiten aufteilen. Man unterscheidet das „Strukturmodell“, das Meßmodell auf der ξ -Seite und das Meßmodell auf der η -Seite.

3.9.1 Das Strukturmodell



Das Strukturmodell besteht stets aus den latenten Konstrukten. Die exogenen Variablen heißen ξ , (sprich: Ksi), die endogenen Variablen η (sprich: Eta). Die Fehlerterme im Strukturmodell werden, wie alle Fehlerterme in Strukturgleichungsmodellen, ebenfalls wie latente Variablen behandelt. Sie heißen ζ (sprich: Zeta). Bei mehreren Variablen eines Typs werden die Variablen durchnummeriert (z.B. η_1 und η_2).

Die Pfadkoeffizienten werden ebenfalls mit griechischen Buchstaben bezeichnet. Pfade, welche die exogenen ξ -Variablen mit den endogenen Variablen verbinden heißen γ (sprich: Gamma). Im Strukturmodell aus der Abbildung gibt es drei γ -Koeffizienten. Sie werden am besten mit „Doppelsubscripten“ bezeichnet. Die γ -Koeffizienten der Gleichungen für die erste η -Variable heißen *gamma* eins-eins bzw. γ eins-zwei, je nach dem, ob ξ_1 oder ξ_2 die exogene Variable ist. Der γ -Koeffizient der Gleichungen für die zweite η -Variable heißt *gamma* zwei-zwei. Gäbe es einen Pfad, von ξ_1 nach η_2 , so würde er mit γ_{21} bezeichnet.

Pfade, welche endogene Variablen untereinander verbinden heißen β . In der Abbildung gibt es nur einen β -Koeffizient. In Anwesenheit von mehreren β -Koeffizienten dienen Doppelsubscripten derselben Logik wie oben zur Unterscheidung.

Das Strukturmodell gibt folgendes Gleichungssystem wieder:

$$\begin{aligned}\eta_1 &= \gamma_{11}\xi_1 + \gamma_{12}\xi_2 + \zeta_1 \\ \eta_2 &= \gamma_{22}\xi_2 + \beta_{21}\eta_1 + \zeta_2\end{aligned}\tag{3.19}$$

In Datenanalysepaketen (Simplis, EQS, SAS) mit Skalarprogrammierung werden diese Gleichungen in der jeweiligen Notation zur Programmierung

3.9. GRUNDSHEMA LINEARER STRUKTURGLEICHUNGSMODELLE 37

der Modelle eingegeben. Lisrel verwendet dagegen eine Matrixnotation zur Programmierung der Modelle. Um eine Idee hiervon zu bekommen ist es sinnvoll, obiges Gleichungssystem etwas vollständiger aufzuschreiben:

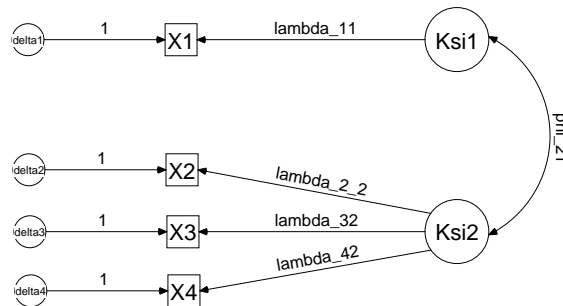
$$\begin{aligned}\eta_1 &= \gamma_{11} * \xi_1 + \gamma_{12} * \xi_2 + 0 * \eta_1 + 0 * \eta_2 + \zeta_1 \\ \eta_2 &= 0 * \xi_1 + \gamma_{22} * \xi_2 + \beta_{21} * \eta_1 + 0 * \eta_2 + \zeta_2\end{aligned}$$

Diese Gleichungen können in Matrizen geschrieben werden:

$$\begin{aligned}\begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} &= \begin{pmatrix} \gamma_{11} & \gamma_{12} \\ 0 & \gamma_{22} \end{pmatrix} * \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ \beta_{21} & 0 \end{pmatrix} * \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} + \begin{pmatrix} \zeta_1 \\ \zeta_2 \end{pmatrix} \\ \boldsymbol{\eta} &= \boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{B}\boldsymbol{\eta} + \boldsymbol{\zeta}\end{aligned}\quad (3.20)$$

In Lisrel verwendet man die Matrizen Γ und B zur Programmierung dieser Gleichungen. Dies geschieht, indem man angibt, welche der Zellen der Matrizen Γ bzw. B 0, und welche „frei“ sind. In unserer Situation würde Lisrel z.B. mitgeteilt, dass in Γ alle Zellen „frei“ sind, bis eben auf die Zelle 2-1.

3.9.2 Das Meßmodell auf der ξ -Seite



Im Meßmodell werden latente Variablen mit ihren manifesten Indikatoren verknüpft. Die Abbildung zeigt das Meßmodell auf der ξ -Seite. Die manifesten Indikatoren der ξ -Variablen heißen X, die Fehlerterme δ . Wie man sieht, werden die Beziehungen derart gestaltet, dass die Indikatoren von den Ausprägungen der latenten Variable abhängt. Dies geschieht, weil man davon ausgeht, dass die manifesten Variablen Messungen des latenten Konzepts darstellen: Zustimmung bzw. Ablehnung zum Statement „Die Regie-

rung sollte Abtreibungen grundsätzlich verbieten“ hängt von der Einstellung zur Abtreibung ab, und nicht etwa umgekehrt.

Das Meßmodell enthält folgendes Gleichungssystem:

$$\begin{aligned} X1 &= \lambda_{11} * \xi_1 + \delta_1 \\ X2 &= \lambda_{22} * \xi_2 + \delta_2 \\ X3 &= \lambda_{32} * \xi_2 + \delta_3 \\ X4 &= \lambda_{42} * \xi_2 + \delta_4 \end{aligned} \tag{3.21}$$

In Datenanalyseprogrammen mit Skalarschreibweise werden diese Gleichungen zur Spezifikation der Modelle eingegeben. Lisrel verwendet auch für das Meßmodell die Matrixnotation. Wie oben können die Gleichungen etwas vollständiger geschrieben werden:

$$\begin{aligned} X1 &= \lambda_{11} * \xi_1 + 0 * \xi_2 + \delta_1 \\ X2 &= 0 * \xi_1 + \lambda_{22} * \xi_2 + \delta_2 \\ X3 &= 0 * \xi_1 + \lambda_{32} * \xi_2 + \delta_3 \\ X4 &= 0 * \xi_1 + \lambda_{42} * \xi_2 + \delta_4 \end{aligned}$$

In Matrixschreibweise:

$$\begin{pmatrix} X1 \\ X2 \\ X3 \\ X4 \end{pmatrix} = \begin{pmatrix} \lambda_{11} & 0 \\ 0 & \lambda_{22} \\ 0 & \lambda_{23} \\ 0 & \lambda_{24} \end{pmatrix} * \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} + \begin{pmatrix} \zeta_1 \\ \zeta_2 \\ \zeta_3 \\ \zeta_4 \end{pmatrix}$$

$$\mathbf{X} = \mathbf{\Lambda}_X \boldsymbol{\xi} + \boldsymbol{\delta} \tag{3.22}$$

Basis der Lisrel-Programmierung des Meßmodells ist die Matrix Λ_X (sprich: Lambda-X). Im vorliegenden Modell würde Lisrel mitgeteilt, dass Λ_X in den Zellen (1,2), (2,1), (3,1), (4,1) „0“ ist.

3.9.3 Das Meßmodell auf der η -Seite

Das Meßmodell auf der η -Seite unterscheidet sich lediglich durch die Nomenklatur vom Meßmodell auf der ξ -Seite. Die manifesten Indikatoren der η -Variablen heißen Y , die Fehlerterme ϵ . Die Koeffizienten heißen ebenfalls λ , die entsprechende Lisrel-Matrix Λ_Y .

3.10 Überblick über die Lisrel Notation

Latente Variablen

ξ	„Ksi“	(Vollkommen) Exogen und keine Fehlervariable	K
η	„Eta“	Endogen	E

Manifeste Variablen

X		Indikatoren für ξ	X
Y		Indikatoren für η	Y

Fehlerterme

δ	„Delta“	Fehler der Meßgleichungen „für“ zwischen X
ϵ	„Epsilon“	Fehler der Meßgleichungen „für“ Y
ζ	„Zeta“	Fehler der Strukturgleichungen „für“ η

Koeffizienten Matrizen (volle Matrizen)

Λ_x	„Lambda-X“	Verbindet ξ und X	LX
Λ_y	„Lambda-Y“	Verbindet η und Y	LY
Γ	„Gamma“	Verbindet ξ und η	GA
B	„Beta“	Verbindet η Variablen untereinander	BE

Varianz-Kovarianz Matrizen (symmetrische Matrizen)

Θ_δ	„Theta- δ “	Varianz-Kovarianz Matrix von δ	TD
Θ_ϵ	„Theta- ϵ “	Varianz-Kovarianz Matrix von η	TE
Φ	„Phi“	Varianz-Kovarianz Matrix der ξ Variablen	PH
Ψ	„Psi“	Varianz-Kovarianz Matrix der ζ Variablen	PS

Groß- und Kleinschrift wird genutzt, um Koeffizienten und ihre Matrizen auseinander zu halten. Die Matrix Γ (großgeschriebenes „Gamma“) enthält die Koeffizienten γ_{ij} (kleingeschriebenes „Gamma“).

Die Gleichungen eines „Lisrel“-Modells lauten in dieser Notation wie folgt:

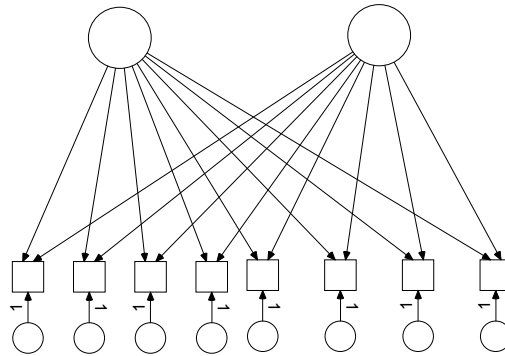
$$\begin{aligned}
 \mathbf{x} &= \mathbf{\Lambda}_x \boldsymbol{\xi} + \boldsymbol{\delta} \\
 \mathbf{y} &= \mathbf{\Lambda}_y \boldsymbol{\eta} + \boldsymbol{\epsilon} \\
 \boldsymbol{\eta} &= \mathbf{B} \boldsymbol{\eta} + \mathbf{\Gamma} \boldsymbol{\xi} + \boldsymbol{\zeta}
 \end{aligned}
 \tag{3.23}$$

3.11 Faktorenanalyse

Anhand des als Meßmodell bezeichneten Teils eines Strukturgleichungsmodells läßt sich die Logik eines weiteren klassischen multivariaten Analyseverfahren — der Faktorenanalyse — aufzeigen. Aufgabe des Meßmodells im Strukturgleichungsmodell ist die Bildung einer messfehler-bereinigten Variablen aus einer bestimmten Anzahl von validen jedoch unterschiedlich reliablen Indikatoren. Das latente Konstrukt wird dabei so konstruiert, dass die Korrelation zwischen den Indikatoren bedeutungslos wird (bzw., in

der Sprache des multiplen Regressionsmodells: der Zusammenhang zwischen zwei Indikatoren verschwindet unter der Kontrolle des latenten Konstrukts). Natürlich gelingt es in der Regel nicht, mit nur einem latenten Konstrukt die Korrelation zwischen den Indikatoren vollständig abzubilden. Die erste Fragestellung einer Faktorenanalyse lautet darum: Wieviel voneinander unabhängige latente Konstrukte — *Faktoren* — sind notwendig, um die Partialkorrelation zwischen den Indikatoren (weitestgehend) zum Verschwinden zu bringen.

Im Sinne eines Strukturgleichungsmodells ausgedrückt hat die Lösung der explorativen Faktorenanalyse folgende Eigenschaften:



- Alle Indikatoren laden auf jeden Faktor
- Alle Fehlerterme sind untereinander unkorreliert
- Die Faktoren — latenten Variablen — sind untereinander unkorreliert

Durch oblique Rotation, können mit der Faktorenanalyse Lösungen erzeugt werden, bei der die latenten Variablen korrelieren. Allerdings können immer nur entweder alle latenten Variablen korrelieren oder nicht korrelieren.

3.12 Literatur

- Bollen, Kenneth A., 1989: Structural Equations with Latent Variables. New York: Wiley.
- Loehlin, John C., 1992: Latent Variable Models. An Introduction to Factor, Path and Structural Analysis. 2. Aufl. Hillsdale u. London: Lawrence Erlbaum.

3.13 Software für SEM's

- Programmierung der Modelle durch Zeichnen des Pfaddiagramms
 - Amos (Zusatzmodul zu SPSS).
- Programmierung der Modelle durch Modellgleichungen in Skalarschreibweise
 - EQS
 - EzPath (Zusatzmodul zu Systat)
 - SAS-Calis (Zusatzmodul zu SAS)
 - Simplis (Bestandteil von Lisrel)
- Programmierung der Modelle durch Modellgleichungen in Matrixnotation
 - Lisrel

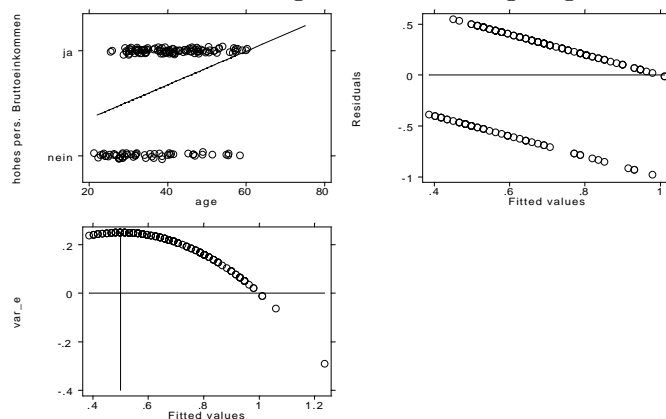
Kapitel 4

Logistische Regression

(Gekürzte und leicht überarbeitete Fassung aus dem Buch Kohler, Ulrich und Frauke Kreuter, Datenanalyse mit Stata, München u. Wien: Oldenbourg, welches voraussichtlich im Jahr 2000 erscheinen wird)

4.1 Warum nicht einfach die lineare Regression?

In Abschnitt 2.2 wurde die lineare Regression verwendet, um den Einfluß des Geschlechts auf das Einkommen festzustellen. Als Einkommen wurde dabei lediglich betrachtet, ob ein Befragter ein hohes Einkommen hat oder nicht. Die Probleme der linearen Regression auf diese abhängige Variable lassen sich am einfachsten mit folgender Abbildung zeigen:



Ein genauer Blick auf die Grafik zeigt Ihnen verschiedene Probleme

1. Die vorhergesagten Werte einer linearen Regression auf eine dichoto-

me abhängige Variable können als Wahrscheinlichkeiten interpretiert werden. Unglücklicherweise kann die Anwendung von OLS zu vorhergesagten Wahrscheinlichkeiten über 1 und unter 0 führen. Im Beispiel ist dies für Personen, die älter als 60 Jahre sind der Fall.

2. Das zweite Problem betrifft eine Annahmen der linearen Regression. Ein dieser Annahmen ist, dass die Fehler der linearen Regression „normal i.i.d.“ sein sollen, d.h. sie sollen nicht, nicht systematisch mit den vorhergesagten Werten zusammenhängen. Zur Überprüfung dieser Annahme verwendet man einen Scatterplot der Residuen (Vorhersagefehler) gegen die vorhergesagten Werte. Dies ist die zweite Grafik in der Abbildung. Sie sehen deutlich, dass Sie systematische Vorhersagefehler gemacht haben.
3. Noch deutlicher wird die Systematik der Fehler in der dritten Graphik. Bei dichotomen abhängigen Variablen können für jeden Vorhersagewert stets nur zwei Residuen auftreten. Die Summe der auftretenden Residuen ist stets 1, d.h. die Fehler sind binomial verteilt. Die Varianz einer binomial-Verteilten Variable ist $p \times (1 - p)$. Berechnet man die Varianz der Fehler und trägt das Ergebnis gegen die vorhergesagten Werte ab, so zeigt sich, dass die Fehlervarianz bei einer vorhergesagten Wahrscheinlichkeit von .5 am größten ist.

4.2 Odds, Odds-Ratio oder: Wetten dass!

Gegeben ist der Datensatz mit Angaben zum sozialen Status, das Geschlecht und das Alter der Passagiere der Titanic, sowie die Angabe darüber, ob sie den Schiffbruch überlebt haben oder nicht.¹ Es soll untersucht werden, ob das in der Schifffahrt geltende Prinzip „Frauen und Kinder zuerst“ bei der Rettung erfolgreich praktiziert wurde. Das heisst, die Kenntnis des Geschlechts und der Altersgruppe sollte eine richtige Zuordnung zu Überlebenden und Gestorbenen ermöglichen. Sicher konnten auch aus diesen Gruppen nicht alle gerettet werden, aber Sie erwarten eine deutlich höhere Überlebenschance für die Frauen und Kinder im Vergleich zu Männern und Erwachsenen.

Den ersten Hinweis auf die Chance, die Frauen gegenüber Männern hatten, erhalten Sie durch eine Kreuztabelle zwischen Geschlecht und dem „Überleben“.

| survived

¹Daten aus gopher://jse.stat.ncsu.edu:70/11/jse/data. Dort auch Angaben zur Herkunft der Daten.

sex	nein	ja	Total
frau	126 26.81	344 73.19	470 100.00
mann	1364 78.80	367 21.20	1731 100.00
Total	1490 67.70	711 32.30	2201 100.00

4.2.1 odds

Die Chance, das Unglück zu überleben, läßt sich berechnen, indem die Zahl der Überlebenden ins Verhältnis zu der Zahl der Gestorbenen gesetzt wird:

$$\frac{711}{1490} = .47$$

Auf die gleiche Überlebenschance kommen Sie, wenn Sie die Anteilswerte (hier die Zeilenprozent) ins Verhältnis setzen²:

$$\frac{.32}{.68} = .47$$

Die Anteilswerte können Sie bei einer dichotomen Variable auch als Wahrscheinlichkeiten interpretieren. Der Anteil der Überlebenden beträgt 32.3%, also konnte der Untergang der Titanic mit einer Wahrscheinlichkeit von .32 überlebt werden. Die Wahrscheinlichkeit zu sterben, beträgt hingegen .68 und ist damit fast doppelt so gross wie die Wahrscheinlichkeit zu überleben. Die Überlebenschance beträgt also knapp 1 : 0.5, die Chance zu sterben ist zwei mal so groß³.

Allgemein können Sie dieses Verhältnis so aufschreiben:

$$Chance_{Leben} = \frac{Wahrscheinlichkeit_{Leben}}{Wahrscheinlichkeit_{Sterben}} \quad (4.1)$$

oder etwas kürzer mit Zeichen anstatt Text:

²Uns interessieren hier nur die ersten beiden Ziffern nach dem Komma. Die Abweichungen danach sind Rundungsfehler.

³Ein Hinweis: nicht immer ist die Umrechnung von Chancen kleiner Null so einfach wie bei 0.5, als die Hälfte von 1. Bei „krummen“ Zahlen können Sie einfach die Zahl mit -1 potenzieren und erhalten dann einen Wert, der Ihnen angibt „um wieviel *Mal* kleiner die Chance ist.“

$$\text{odds} = \theta = \frac{P(y = 1)}{1 - P(y = 1)} \quad (4.2)$$

wobei *odds* die Chance auf ein Überleben bezeichnet.

Die Überlebenschance können Sie nun für die Geschlechter getrennt berechnen. Für die Frauen sieht dieses Verhältnis deutlich positiver aus als für alle Passagiere insgesamt:

$$\frac{344}{126} = 2.73$$

Das heißt, für Frauen stehen die Chancen, zu den Überlebenden zu gehören, 2.7 : 1 . Oder anders formuliert, Frauen gehören 2.71 mal eher zu den Überlebenden, als zu den Nicht-Überlebenden. Die Männer haben hingegen nur eine Chance von

$$\frac{367}{1364} = .269$$

4.2.2 odds-ratio

Interessanter als die Chancen für jede Gruppe getrennt, ist der Vergleich der Chancen von Männern und Frauen. Sprich die Frage, um wieviel kleiner ist die Überlebenschance der Männer *im Vergleich* zu den Frauen? Dazu setzten Sie die „odds“ der Männer ins Verhältnis zu den „odds“ der Frauen:

$$\Omega = \frac{\theta_{\text{Männer}}}{\theta_{\text{Frauen}}} \quad (4.3)$$

Dieses Verhältnis wird „odds ratio“ genannt.

Beispiel Die Überlebenschance eines Mannes ist

$$\frac{.269}{2.73} = .099$$

mal so gross wie die einer Frau, bzw. 10 mal kleiner als die einer Frau.

4.3 Regressionsmodell

In der logistischen Regression wird das sogenannte *Logit* (L), i.e. der Logarithmus der Chance (des odds) als abhängige Variable betrachtet. Wie in der linearen Regression wird versucht, diese abhängige Variable als Linearkombination der unabhängigen Variable auszudrücken. Das logistische Regressionsmodell lautet:

$$\ln\left(\frac{p}{1-p}\right) = \widehat{L} = b_0 + b_1X_{i1} + b_2X_{i2} + \dots + b_{k-1}X_{i,k-1}. \quad (4.4)$$

Die Interpretation der b-Koeffizienten lautet formal: Die logarithmierte Chance steigt um b_1 Einheiten an, wenn man auf der unabhängigen Variable X_{i1} um einen Schritt ansteigt.

Beispiel Das logistische Regressionsmodell für das Titanic-Beispiel ergibt:

Logit estimates	Number of obs	=	2201
	LR chi2(1)	=	434.47
	Prob > chi2	=	0.0000
Log likelihood = -1167.4939	Pseudo R2	=	0.1569

survived	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
sex	-2.317175	.1195885	-19.376	0.000	-2.551564 -2.082786
_cons	1.00436	.104132	9.645	0.000	.8002648 1.208455

Interpretation: Wie im linearen Regressionmodell können die Vorhersagewerte berechnet werden:

- für die Frauen (sex=0)

$$\begin{aligned} \widehat{L} &= 1.00436 + (-2.317175) \times 0 \\ &= 1.00436 \end{aligned}$$

- für die Männer (sex = 1)

$$\begin{aligned} \widehat{L} &= 1.00436 + (-2.317175) \times 1 \\ &= -1.312815 \end{aligned}$$

Die Vorhersagewerte erhalten eine angenehmere Bedeutung, wenn wir sie entlogarithmieren:

$$\begin{aligned} e^{\hat{L}} &= e^{1.00436} \\ &= 2.73 \end{aligned}$$

bzw.

$$\begin{aligned} e^{\hat{L}} &= e^{1.00436+(-2.317175)} \\ &= .269 \end{aligned}$$

Dies sind die bereits bekannten *odds*.

Die Interpretation der Koeffizienten lautet dementsprechend: Männer ($\text{sex}=1$) hatten eine um -2.3 niedrigere logarithmierte Chance die Titanic-Katastrophe zu überleben als Frauen. Diese etwas hölzerne Interpretation kann etwas vereinfacht werden, wenn wir die beide Seiten der Gleichung entlogarithmieren. Wie gesehen wird dadurch aus der abhängigen Variable das *odd*. Was passiert auf der Seite der unabhängigen Variable?

$$\begin{aligned} e^{\hat{L}} &= e^{b_0+b_1X_{i1}+b_2X_{i2}+\dots+b_{k-1}X_{i,k-1}} \\ \theta &= e^{b_0} \times e^{b_1X_{i1}} \times e^{b_2X_{i2}} \times \dots \times e^{b_{k-1}X_{i,k-1}} \end{aligned} \quad (4.5)$$

Jeder Schritt auf der Variable X_{i1} führt zu einer Erhöhung der Chance um des b_1 -fache. Der exponierte b -Koeffizient der logistischen Regression gibt also die *multiplikative* Veränderung wieder, oder, anders ausgedrückt: Der exponierte b -Koeffizient der logistischen Regression ist das *odds-ratio*.

Im Beispiel haben Männer eine um das

$$\begin{aligned} e^{b_{\text{sex}}} &= e^{(-2.317175)} \\ &= .097 \end{aligned}$$

fache kleinere Chance zu überleben als Frauen, oder, anders herum, Männer haben ein $.097^{-1} = 10.31$ mal größeres Riskiko zu sterben als Frauen.

Wie bei der linearen Regression können Sie auch bei der logistischen Regression mehrere unabhängige Variablen aufnehmen.

Beispiel Sie haben nach ihrem letzten Kinobesuch den Verdacht, dass die Regel „Frauen und Kinder zuerst“ nicht immer eingehalten wurde. Feine Herren der ersten Klasse haben sich ihren Platz in den Rettungsbooten erkaufte, auf Kosten von Frauen und Kindern der dritten Klasse. Sie wollen diesen Verdacht erhärten und wollen in Ihre Regression die unabhängigen Variablen Alter und Klasse aufnehmen.

```

Logit estimates                                     Number of obs =      2201
                                                    LR chi2(5)      =      559.40
                                                    Prob > chi2     =      0.0000
Log likelihood = -1105.0306                       Pseudo R2      =      0.2020

```

survived	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
sex	-2.42006	.1404101	-17.236	0.000	-2.695259	-2.144862
age	-1.061542	.2440257	-4.350	0.000	-1.539824	-.5832608
class_2	.8576762	.1573389	5.451	0.000	.5492976	1.166055
class_3	-.1604188	.1737865	-0.923	0.356	-.5010342	.1801966
class_4	-.9200861	.1485865	-6.192	0.000	-1.21131	-.6288619
_cons	2.247704	.2988261	7.522	0.000	1.662015	2.833392

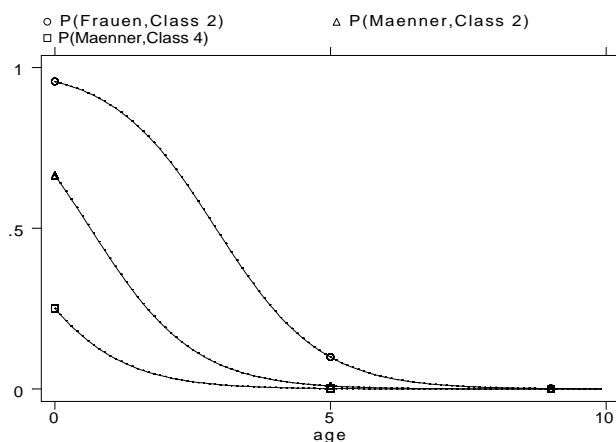
Interpretation: Die Koeffizienten werden wie im vorhergehenden Beispiel interpretiert. Wie man sieht, haben die Personen aus der dritten Klasse (class_4), egal welchen Geschlechts oder Alters, sogar eine schlechtere Chance zu Überleben als die Besatzungsmitglieder (class_1).

Interessant ist der Effekt des Alters. Die Altersvariable ist im Datensatz lediglich dichotom enthalten, d.h. es wird nur zwischen Erwachsenen und Kindern unterschieden. Lassen Sie uns jedoch einen Augenblick so tun, als sei das Alter in 10er-Schritten, von 0 bis 90 im Datensatz enthalten, also 0-10, 10-20, 20-30 usw. In diesem Fall würde man sagen: Alle 10 Jahre älter sinkt die logarithmierte Chance zu überleben um -1.06 Einheiten. Dies entspricht einer Reduktion der Überlebenschance um etwa 1/3 ($e^{-1.06} = .35$)

Beachte: Eine Verdreifachung der Überlebenschance führt nicht zu einer Verdreifachung der Überlebenschance. Die Chancen können aber einfach in Wahrscheinlichkeiten umgerechnet werden:

$$p = \frac{\theta}{1 + \theta} \quad (4.6)$$

Die Interpretation der vorhergesagten Wahrscheinlichkeiten erfolgt am besten graphisch durch *conditional effects plots*:



4.4 Gesamtfit des Modells

Wie bei der linearen Regression besprochen, gilt auch in der logistischen Regression: Prüfen Sie *immer* zuerst, wie gut ihr Modell paßt. Als Test des Gesamtmodells wird ein χ^2 -Wert verwendet. Dieser dient als Test der Nullhypothese: alle Koeffizienten außer der Konstanten sind gleich Null. Der Wert berechnet sich aus der Differenz der Likelihood-Funktion des Modells ohne unabhängige Variablen und des Likelihood-Wertes mit den unabhängigen Variablen.

$$X_{\mathcal{L}}^2 = -2 \times (\ln \mathcal{L}_0 - \ln \mathcal{L}_1) \quad (4.7)$$

mit \mathcal{L}_0 dem Wert der Likelihood-Funktion des Modells ohne unabhängige Variablen und \mathcal{L}_1 dem Likelihood-Wert mit den unabhängigen Variablen.

Im Beispiel ist die Wahrscheinlichkeit für einen χ^2 Wert mit zwei Freiheitsgraden sehr niedrig, so dass die Nullhypothese - der Einfluß des Alters, des Geschlechts und der Klasse ist gleich *Null* - zurückgewiesen werden kann. Allerdings dürfte klar sein, dass kaum einer ernsthaft behaupten wird, Alter, Klasse und Geschlecht habe *keinerlei* Effekt auf die Überlebenswahrscheinlichkeit bei der Titanic-Katastrophe gehabt. Die Zurückweisung der Nullhypothese reicht deshalb in keinster Weise aus, um mit den Ergebnissen zufrieden zu sein.

Ein anderes Maß für die Modellgüte ist Pseudo- R^2 . Es gibt zahlreiche Versionen von Pseudo- R^2 . Im Beispiel wird derjenige von McFadden ausgewiesen. Leider läßt sich Pseudo- R^2 nicht so leicht interpretieren, wie das R^2 der linearen Regression. Je höher je besser ist, in nur ein grober Hinweis:

$$pseudoR^2 = 1 - \frac{\ln \mathcal{L}_1}{\ln \mathcal{L}_0} \quad (4.8)$$

Ein weiterer Test für die Güte des Modells liefert die Klassifikationstabelle, eine Tabelle, die die Zahl korrekt klassifizierter Werte angibt. Dazu wird für alle Fälle, die eine vorhergesagten Wahrscheinlichkeit über 0.5 haben die Ausprägung 1, für alle anderen 0 vorhergesagt. Die Klassifikationen werden mit den tatsächlichen Werten verglichen.

Mit Ihrem Modell sind Sie übrigens in der Lage 77.83% der Fälle korrekt zu klassifizieren. Aber Achtung: auch ohne Kenntnis der unabhängigen Variable sind Sie in der Lage, einige Fälle korrekt zu klassifizieren. Tippen Sie einfach immer auf „gestorben“ und Sie werden meistens richtig liegen. Betrachten Sie also gleichzeitig die Klassifikationstabelle, die sich nach einer logistischen Regression ohne UV's ergibt. Im Beispiel zeigt sich, dass in einem solchen Modell auch schon 67.7% der Fälle richtig klassifiziert wurden.

Einen „goodness-of-fit“-Test für das von uns spezifizierte Modell ist der „Hosmer-Lemeshow-Test“. Ausgangspunkt ist ein Vergleich von beobachteten Fällen in einer Zelle gegenüber der Zahl der vorhergesagten Fälle in einer Zelle. Wobei die Zellen sich aus Kombination der unabhängigen Variablen ergeben („covariate pattern“). Zwei oder mehr Fälle haben das gleiche Muster, wenn sie auf den im Modell verwendeten unabhängigen Variablen die gleichen Werte aufweisen. Wenn die Zahl der Kombinationen nahe an der Zahl der Beobachtungen liegt, wird die Güte des Tests fragwürdig. Hosmer und Lemeshow (1989:135ff.) haben deshalb eine Veränderung des Tests vorgeschlagen, bei dem die Daten nach den vorhergesagten Wahrscheinlichkeiten geordnet werden und in g annähernd gleich grosse Gruppen aufgeteilt werden. Für jede Gruppe wird dann die Häufigkeit der beobachteten Antworten in dieser Gruppe mit der geschätzten verglichen, somit die Differenz zwischen beobachteten und vorhergesagten Häufigkeiten errechnet. Die sich daraus ergebene Prüfgrösse (die Differenz) ist χ^2 verteilt, die Anzahl der Freiheitsgrade ergibt sich der Zahl der Gruppen $g - 2$. Üblich ist eine Unterteilung in 10 Gruppen⁴.

4.5 Test des Einfluß einzelner Variablen

Auch die Koeffizienten einzelner Variablen können gegen Null getestet werden. Die Logik entspricht dem Test des Gesamtmodells. Die Hypothese, dass der Koeffizient einer bestimmten unabhängigen Variablen gleich Null ist,

⁴Achtung: Bei wenigen unabhängigen Variablen mit kategorialer Ausprägung eignet sich diese Test Variante nicht.

wird durch die Differenz der zweifach negativen Log Likelihood-Funktion berechnet, diesmal aber zwischen dem Modell, welche die zu testende Variable enthält ($\ln\mathcal{L}_{mit}$) und dem Modell ohne die zu testende Variable ($\ln\mathcal{L}_{ohne}$). Diese Prüfgröße folgt ebenfalls einer χ^2 Verteilung, wobei die Anzahl der Freiheitsgrade die Differenz der Anzahl der Parameter zwischen den beiden Modellen ist.

$$X_{Diff}^2 = -2(\ln\mathcal{L}_{ohne} - \ln\mathcal{L}_{mit}) \quad (4.9)$$

Wichtig ist, dass beide Modelle auf der gleichen Fallzahl beruhen.

4.6 Zum Maximum-Likelihood-Verfahren

(Die Darstellung folgt S. 28-49 in Andres/Hagenaars/Kühnel (1997), Analyse von Tabellen und kategorialen Daten. Berlin usw.: Springer)

Die Koeffizienten der logistischen Regression werden nicht wie bei der linearen Regression durch OLS bestimmt sondern durch das „Maximum-Likelihood-Verfahren“. In diesem Abschnitt soll dieses Verfahren etwas näher erläutert werden.

Beispiel Angenommen Sie haben drei Populationen, in denen ein bestimmtes Merkmal — z.B. die Parteiidentifikation — unterschiedlich verteilt sind:

Population I	Population II	Population III
CDU SPD	CDU SPD	CDU SPD
CDU SPD	CDU SPD	CDU SPD
CDU SPD	CDU SPD	SPD
CDU	CDU SPD	SPD
CDU	CDU SPD	SPD
CDU		SPD
CDU		SPD

Aus einer dieser Populationen wurde eine Stichprobe mit Zurücklegen vom Umfang 3 gezogen. Die Stichprobe enthält die Merkmalsträger

SPD, SPD, CDU

Aus welcher der drei Populationen stammt die Stichprobe? Die intuitive Antwort lautet: Aus der Population III. Die statistische Antwort verwendet die Verteilungsfunktion der Binomialverteilung für die Antwort. Die Wahrscheinlichkeit in einer Stichprobe mit Zurücklegen vom Umfang n genau f mal ein Merkmal der Ziehungswahrscheinlichkeit π zu ziehen ist

$$P(f|\pi, n) = \binom{n}{f} \pi^f (1 - \pi)^{n-f} \quad (4.10)$$

wobei der Ausdruck $\binom{n}{f}$, sprich n über f , $\frac{n!}{f!(n-f)!}$ bedeutet.

Entsprechend ist die Wahrscheinlichkeit für eine Stichprobe wie die gegebene beim Ziehen aus Population 1

$$\begin{aligned} P(2|.3, 3) &= \binom{3}{2} \cdot .3^2 (1 - .3)^{3-2} \\ &= \frac{3 \times 2 \times 1}{(2 \times 1) \times (3 - 2)!} \times .3^2 \times .7 \\ &= \frac{6}{3} \times .09 \times .7 = .189 \end{aligned}$$

Für die beiden anderen Populationen lauten dieselben Wahrscheinlichkeiten:

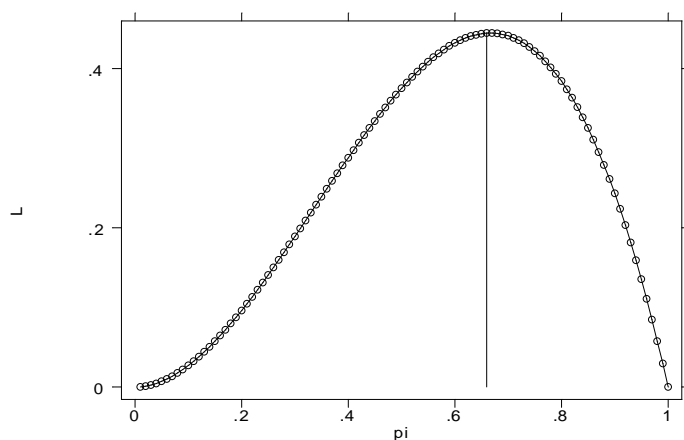
$$\begin{aligned} P(2|.5, 3) &= \binom{3}{2} \cdot .5^2 (1 - .5)^{3-2} = .375 \\ P(2|.8, 3) &= \binom{3}{2} \cdot .8^2 (1 - .8)^{3-2} = .384 \end{aligned}$$

Daraus ergibt sich, dass die Stichprobe wahrscheinlich aus der Population 3 gezogen wurde.

Das war einfach. Leider zeigte das Beispiel aber eine etwas untypische Situation. Normalerweise kennen wir die Häufigkeit eines Merkmals in der Population — π — nicht. Gesucht ist darum der Verteilungsparameter π , *der die gezogene Stichprobe am wahrscheinlichsten macht*. Eine Antwort hierauf findet man durch die *Likelihood-Funktion*. Die Likelihood-Funktion in unserem Beispiel lautet:

$$\mathcal{L}(\pi|f, n) = \binom{n}{f} \pi^f (1 - \pi)^{n-f} \quad (4.11)$$

Ein denkbare Lösung ist es, möglichst viele Werte auszuprobieren. Tut man dies, so kommt man zum Schluß, dass die Stichprobe wahrscheinlich aus einer Population gezogen wurde, in der der Anteil der SPD-Anhänger ca. $\pi = .67$ beträgt:



Genauer ist es allerdings, die erste Ableitung der Likelihood-Funktion auszurechnen und 0 zu setzen.

Dasselbe Verfahren kann man übrigens anwenden, wenn man davon ausgeht, dass es in der Stichprobe zwei Gruppen mit unterschiedlicher Verteilung der Parteiidentifikation gibt. In diesem Fall wird die Likelihood-Funktion

$$\begin{aligned} \mathcal{L}(\pi|f, n, m') &= \binom{n_1}{f_1} \pi_1^{f_1} (1 - \pi_1)^{n_1 - f_1} \times \binom{n_2}{f_2} \pi_2^{f_2} (1 - \pi_2)^{n_2 - f_2} \\ &= \underbrace{\binom{n_1}{f_1} \binom{n_2}{f_2}}_K \pi_1^{f_1} (1 - \pi_1)^{n_1 - f_1} \pi_2^{f_2} (1 - \pi_2)^{n_2 - f_2} \end{aligned} \quad (4.12)$$

Wobei m' anzeigt, dass es sich um eine Modellannahme handelt.

Mit multiplikativen Termen ist schwer zu rechnen. Deshalb bilden wir den Logarithmus, die *Log Likelihood Funktion*:

$$\begin{aligned} \ln \mathcal{L}(\pi|f, n, m') &= \ln(K) + f_1 \ln(\pi_1) + (n_1 - f_1) \ln(1 - \pi_1) \\ &\quad + f_2 \ln(\pi_2) + (n_2 - f_2) \ln(1 - \pi_2) \end{aligned} \quad (4.13)$$

Die erste Ableitung dieser Funktion erbringt eine analytische Lösung für π_1 und π_2 . Dies ist allerdings nicht immer für alle Likelihood-Funktionen der Fall. Häufig können Likelihood-Funktionen nicht direkt gelöst werden und werden darum mit iterativen Verfahren gelöst.

4.7 Literatur

- Agresti, Alan, 1990: *Categorical Data Analysis*. New York usw.: Wiley.
- Aldrich, John und Forest Nelson, 1984: *Linear Probability, Logit and Probit Models*. London: Sage .
- Andreß, Hans-Jürgen, Jacques A. Hagenaars, und Steffen Kühnel, 1997: *Analyse von Tabellen und kategorialen Daten. Log-lineare Modell, latente Klassenanalyse, logistische Regression und GSK-Ansatz*. Berlin usw.: Springer.
- Fox, John, 1997: *Applied regression analysis, linear models, and related methods*. Thousand Oaks usw. Sage. Hamilton, Lawrence C., 1992: *Regression with Graphics. A Second Course in Applied Statistics*. Belmont: Wadsworth.
- Hamilton, Lawrence C., 1992: *Regression with Graphics. A Second Course in Applied Statistics*. Belmont: Wadsworth.
- Long, Scott J., 1997: *Regression models for categorical and limited dependent variables*. London: Sage.

Kapitel 5

Loglineare Modelle

Das Verfahren der logistischen Regression kann auf kategoriale abhängige Variablen mit mehr als 2 Ausprägungen verallgemeinert werden (multinomiale logistische Regression). Liegen nur kategoriale Variablen als unabhängige Variablen vor kann stattdessen das rechentechnisch einfachere loglineare Modell berechnet werden. Ausgangsdaten für loglineare Modelle sind mehrdimensionale Kreuztabellen.

Bei loglinearen Modelle werden die Zellbesetzungen einer mehrdimensionalen Kreuztabellen als abhängige Variable verwendet. Der Logarithmus dieser Zellbesetzungen (G) wird durch ein lineares Modell vorhergesagt. Das saturierte loglineare Modell für eine dreidimensionale Kreuztabelle lautet formal:

$$G_{ijkl}^{ABC} = \theta + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{ik}^{AC} + \lambda_{jk}^{BC} + \lambda_{ijk}^{ABC} \quad (5.1)$$

Wobei die Parameter des Modells unter Dummy-Kodierung durch

$$\begin{aligned} \theta &= G_{000}^{ABC} \\ \lambda_i^A &= G_{i00}^{ABC} - \theta \\ \lambda_{ij}^{AB} &= G_{ij0}^{ABC} - \theta - \lambda_i^A - \lambda_j^B \\ \lambda_{ijk}^{ABC} &= G_{ijk}^{ABC} - \theta - \lambda_i^A - \lambda_j^B - \lambda_k^C - \lambda_{ij}^{AB} - \lambda_{ik}^{AC} - \lambda_{jk}^{BC} \end{aligned} \quad (5.2)$$

bestimmt werden.

Unabhängige Variablen des Modells sind die Variablen, welche die Kreuztabelle aufspannen. Zusammenhangshypothesen werden durch Generierung der entsprechenden Interaktionen modelliert. Gesucht wird ein Modell, dass mit möglichst wenig Zusammenhangshypothesen (Interaktionen) die Zellbesetzungen der Tabelle möglichst gut vorhersagt.

Anhand von (5.2) erkennt man, dass die Parameter durch die Zellhäufigkeiten, bzw. deren Logarithmus unmittelbar determiniert sind. Normalerweise ist man aber nicht an den beobachteten Zellhäufigkeiten, sondern an den Zellhäufigkeiten unter einer bestimmten Modellannahme interessiert. Wir könnten z.B. fragen, wie die Kreuztabelle aussehen würde, wenn Sie nur von den Randverteilungen abhängen würde. Die Zellhäufigkeiten unter dieser Modellannahme kann man mit dem Maximum-Likelihood Verfahren bestimmen. Sind die Zellhäufigkeiten der Modellannahme bekannt ergeben sich hieraus die λ -Parameter.

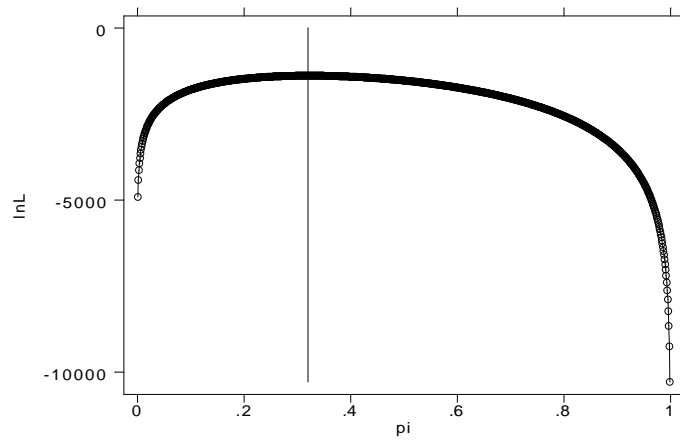
Beispiel Gegeben sind die Titanic-Daten aus dem vorangegangenen Kapitel. Wir beobachten folgende Tabelle:

survived	sex		Total
	frau	mann	
nein	126	1364	1490
ja	344	367	711
Total	470	1731	2201

Wie groß wären die Zellhäufigkeiten, wenn wir annehmen, dass der Anteil der Überlebenden bei Männern und Frauen gleich ist. Da wir nach dem Maximum-Likelihood-Verfahren vorgehen, suchen wir danach, welche Verteilungsparameter π_{men} und π_{women} die gegebenen Zellhäufigkeiten am wahrscheinlichsten machen, unter der Annahme, dass $\pi_{men} = \pi_{women}$. Da das Überleben binomialverteilt ist, können wir die Likelihood-Funktion aus (4.13) verwenden. Wir suchen also diejenigen π_{men} und π_{women} welche (4.13) maximiert. Durch einsetzen erhält man:

$$\begin{aligned} \ln \mathcal{L}(\pi | f, n, m') &= \ln(K) + 344 \ln(\pi) + (470 - 344) \ln(1 - \pi) \\ &\quad + 367 \ln(\pi) + (1731 - 367) \ln(1 - \pi) \end{aligned}$$

Lassen Sie uns einfach ausprobieren. Dabei können wir $\ln(K)$ vergessen, da dieser Teil der Likelihood-Funktion konstant für alle möglichen π ist:



Unter der Annahme gleicher Anteile von Überlebenden Männern und Frauen ist bei unsere Stichprobe eine Verteilungsparameter von ungefähr .323 am wahrscheinlichsten. Wir können den Verteilungsparameter verwenden um die Zelhäufigkeiten für unserer Modellannahme zu bestimmen. Bei gegebener Randverteilung erhalten wir für die Frauen

$$f_{\text{frauen,survived}} = 470 * .323 = 152$$

Überlebende und $470 - 150 = 318$ Gestorbene. Für die Männer erhalten wir:

$$f_{\text{männer,survived}} = 1731 * .323 = 559$$

Überlebende und $1731 - 559 = 1172$ Gestorbene. Dies ergibt folgende Tabelle¹:

survived	sex		Total
	frau	mann	
nein	318	1172	1490
ja	152	559	711
Total	470	1731	2201

Nun können wir die Modellparameter θ und λ gemäß (5.2) berechnen:

¹Dieselben Häufigkeiten erhält man auch durch die Berechnung Unabhängigkeitstabelle des üblichen χ^2 -Tests. Die fiktiven Häufigkeiten unter Unabhängigkeit sind deshalb ein Maximum-Likelihood-Schätzer des Unabhängigkeitsmodells

$$\begin{aligned}\theta &= \ln(318) = 5.76 \\ \lambda_1^{sex} &= \ln(1172) - 5.76 = 1.31 \\ \lambda_1^{surv} &= \ln(152) - 5.76 = -.74 \\ \lambda_{11}^{sex,surv} &= \ln(559) - 5.76 - 1.31 - (-.74) \approx 0\end{aligned}$$

Nur um sicher zu gehen, hier die Lösung eines Datenanalyseprogramms (Stata):

```
Iteration 0: log likelihood = -232.82453
Iteration 1: log likelihood = -232.8116
Iteration 2: log likelihood = -232.8116
```

```
Poisson regression                               Number of obs   =           4
                                                LR chi2(2)      =    1050.10
                                                Prob > chi2     =     0.0000
Log likelihood = -232.8116                    Pseudo R2      =     0.6928
```

	n	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
sex		1.303722	.0520131	25.065	0.000	1.201778	1.405666
survived		-.739859	.0455808	-16.232	0.000	-.8291957	-.6505222
_cons		5.762597	.0484196	119.014	0.000	5.667696	5.857498

Die Interpretation der Modellparameter lautet wie gewohnt: Wenn das Geschlecht um eine Einheit ansteigt — wenn wir also Männer und nicht Frauen betrachten — steigt die logarithmierte Fallzahl gegenüber der Referenzkategorie (die nicht-überlebenden Frauen) um 1.31. Oder, anders ausgedrückt, es gibt $e^{1.31} = 3.70$ mal mehr nicht überlebende Männer als Frauen. Wenn das „Überleben“ um eine Einheit steigt, so sinkt die logarithmierte Fallzahl gegenüber der Referenzkategorie um -.74. Wir haben also $e^{-.74} = .47$ mal so viel Überlebende (Frauen) in der Tabelle als nicht Überlebende. Die Anzahl der Überlebenden Männer, also ein Ansteigen des Geschlechts und des Überlebens, hat keinen weiteren Effekt auf die Fallzahl. Es ist dies lediglich eine Funktion des Anstiegs der Fallzahl der Männer und der insgesamt höheren Sterbewahrscheinlichkeit und damit eine Funktion der Randverteilung. Man kann sagen: Geschlecht und Sterbewahrscheinlichkeit sind voneinander unabhängig. Bevor diese Schlußfolgerung gezogen wird, sollte allerdings der Modellfit betrachtet werden.

5.1 Bestimmung des Modellfits

Der Modellfit wird über den Abstand zwischen den Häufigkeiten der beobachteten Tabelle und dem fiktiven Häufigkeiten unter der jeweiligen Modellannahme bewertet. Üblicherweise wird dazu der Likelihood X^2 verwendet:

$$X_L^2 = 2 \sum_k f_k \ln \left(\frac{f_k}{\hat{F}_k} \right)$$

verwendet, wobei \hat{F}_k die fiktiven Häufigkeiten unter der jeweiligen Modellannahme und f die beobachteten Häufigkeiten darstellen. Manche Datenanalyseprogramme (SPSS) geben gleichzeitig auch Pearson's X^2 aus:

$$X_P^2 = \sum_k \frac{(f_k - \hat{F}_k)^2}{\hat{F}_k} \quad (5.3)$$

Beide X^2 -Werte können gegen die χ^2 -Verteilung getestet werden, wobei man hier daran interessiert ist, eine möglichst gute Anpassung zu erzielen (kleine X^2 -Werte, hohe Wahrscheinlichkeit für die Nullhypothese).

Beispiel Das Unabhängigkeitsmodell der Titanic-Daten hat folgenden Likelihood X^2 :

$$X_L^2 = 2 \times \left[126 \times \ln \left(\frac{126}{318} \right) + 344 \times \ln \left(\frac{344}{152} \right) + 1364 \times \ln \left(\frac{1364}{1172} \right) + 367 \times \ln \left(\frac{367}{559} \right) \right] = 433.64 \quad (5.4)$$

Dies ist ein recht hoher Wert. Die Wahrscheinlichkeit, dass beobachtete Daten und das Modell übereinstimmen ist (bei einem Freiheitsgrad) praktisch 0. Daher sollte man versuchen das Modell zu verbessern.

Wie bei der logistischen Regression kann auch die Differenz der Likelihood-Funktion des Modells ohne unabhängige Variablen und des Likelihood-Wertes mit den unabhängigen Variablen als Maß für den Fit des Modells verwendet werden. Dies ist der LR X^2 in obigen Programoutput. Entsprechend können diese Werte zur Grundlage der diversen Pseudo R^2 -Werte gemacht werden.

5.2 Verbesserungen des Modells

Man kann die Annahme der gleichen Verteilung der Überlebenschance von Männern und Frauen aufgeben, um das Modell zu verbessern ($\pi_{\text{Männer}} \neq \pi_{\text{Frauen}}$). Hierzu wird einfach ein Interaktionsterm in die Gleichung eingeführt. Wie alle Interaktionsterme haben diese die Bedeutung, dass der Effekt der einen Variable von der Ausprägung der anderen Variable abhängt. Hier: der Effekt eines Schritts auf der Variable „Überlebt“ ist für Männer und Frauen unterschiedlich. Dies ist gleichbedeutend mit der Feststellung unterschiedlicher Überlebenschancen.

Wie bei der logistischen Regression dient die Differenz der Likelihood–Werte als Prüfgröße für Verbesserungen des Modells.

$$X_{Diff}^2 = -2(\ln\mathcal{L}_{\text{ohne}} - \ln\mathcal{L}_{\text{mit}}) \quad (5.5)$$

Beispiel Die Aufgabe der Unabhängigkeitsvermutung im loglinearen Modell der Titanic–Daten ergibt folgendes Ergebnis:

```
Iteration 0: log likelihood = -17.741173
Iteration 1: log likelihood = -15.57996
Iteration 2: log likelihood = -15.577177
Iteration 3: log likelihood = -15.577177
```

```
Poisson regression                                Number of obs =          4
LR chi2(3)                                       =        1484.57
Prob > chi2                                       =          0.0000
Pseudo R2                                        =          0.9794
Log likelihood = -15.577177
```

n	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
sex	2.381895	.0931109	25.581	0.000	2.199401	2.564389
survived	1.00436	.1041321	9.645	0.000	.8002647	1.208455
sexsurv	-2.317175	.1195885	-19.376	0.000	-2.551564	-2.082786
_cons	4.836282	.0890871	54.287	0.000	4.661674	5.010889

Im Unabhängigkeitsmodell betrug $\ln\mathcal{L}_{\text{ohne}} = -232.18$. Im verbesserten Modell ist $\ln\mathcal{L}_{\text{mit}} = -15.57$. Daraus ergibt sich

$$X_{\text{Diff}}^2 = -2(-232.18 - (-15.57)) = 433 \quad (5.6)$$

Die Differenz ist so groß, dass sie fast sicher nicht 0 ist². Wir können also

²Es ist übrigens derselbe Wert wie der X_L^2 im Unabhängigkeitsmodell. Dies ist kein

davon ausgehen, dass sich die Überlebenschancen von Männern und Frauen unterscheiden.

Noch ein Wort zum Koeffizient des Interaktionsterms. Die formale Interpretation wäre: Die logarithmierte Fallzahl des Überlebens reduziert sich bei den Männern ist um 2.3 weniger als bei den Frauen. Eine sinnvollere Interpretation ergibt sich, wenn wir aus dem loglinearen Modell eine Logit-Gleichung formen. Wir könnten z.B. an der logarithmierten Chance zu Überleben interessiert sein. Mit dem loglinearen Modell läßt sich dieses Logit für die Männer durch

$$\begin{aligned} L_{\text{Männer}} &= G_{11}^{\text{sex,surv}} - G_{10}^{\text{sex,surv}} \\ &= (\theta + \lambda_{\text{sex}} + \lambda_{\text{surv}} + \lambda_{\text{sex,surv}}) - (\theta + \lambda_{\text{sex}}) \\ &= \lambda_{\text{surv}} + \lambda_{\text{sex,surv}} \end{aligned} \quad (5.7)$$

Das entsprechende Logit für die Frauen ist:

$$L_{\text{Frauen}} = \lambda_{\text{surv}} + \lambda_{\text{sex,surv}} \quad (5.8)$$

Die logarithmierte Chance der Männer ist demnach um $\lambda_{\text{sex,surv}}$ kleiner als die der Frauen. Der Koeffizient hat also dieselbe Interpretation wie der Geschlechtseffekt des Logit-Modells auf Seite 47 ist. Er hat im übrigen auch denselben Wert.

5.3 Ein multivariates Beispiel

Gegeben ist folgende Tabelle aus dem Titanic-Datensatz:

```

-----+-----
age, sex | survived
and class | nein   ja
-----+-----
jung     |
frau     |
         | 0 |
         | 1 |         1
         | 2 |         13
         | 3 |    17    14
-----+-----
jung     |

```

Zufall. Denn X_L^2 im aktuellen Modell ist genau 0, d.h. wir haben die Zellhäufigkeiten exakt reproduziert.

```

mann      |
          |
          0 |
          1 |           5
          2 |          11
          3 |         35   13
-----+-----
alt
frau      |
          |
          0 |         3   20
          1 |         4  140
          2 |        13   80
          3 |        89   76
-----+-----
alt
mann      |
          |
          0 |       670  192
          1 |       118   57
          2 |       154   14
          3 |       387   75
-----+-----

```

Ein loglineares Modell dieser Tabelle ist etwas schwerer zu berechnen, da nicht alle Variablen binomial verteilt sind (Klasse). Deshalb muß eine andere Verteilungsannahme verwendet werden — die Multinomial, Produkt-Multinomial oder die Poisson-Verteilung. Wir sollten uns hierum nicht kümmern und die Arbeit einem Datenanalyseprogramm überlassen. Hier ist das Ergebnis des Unabhängigkeitsmodells:

```

Poisson regression                                Number of obs =      24
                                                LR chi2(6)      = 2643.96
                                                Prob > chi2     =  0.0000
Log likelihood = -585.10175                    Pseudo R2      =  0.6932

```

```

-----+-----
          n |      Coef.  Std. Err.      z    P>|z|      [95% Conf. Interval]
-----+-----
survived |  -.7755565  .0457275   -16.960  0.000   - .8651807   -.6859324
sex      |   1.303722  .0520131   25.065  0.000    1.201778    1.405666
age      |   2.001222  .101319   19.752  0.000    1.802641    2.199804
class2   |  -1.043496  .0649989   -16.054  0.000   -1.170891   -.9161001
class3   |  -1.174832  .06824     -17.216  0.000   -1.30858    -1.041084
class4   |  -.3527548  .0518838    -6.799  0.000   -.4544451   -.2510644
_cons    |   2.861688  .1161767   24.632  0.000    2.633986    3.08939
-----+-----

```

Das Modell gibt nur darüber Auskunft, dass mehr Männer als Frauen, mehr alte als junge und mehr Besatzungsmitglieder als jede Passagiere einer bestimmten Klasse an Bord sind. Dies ist nicht sonderlich interessant, ist aber immerhin mehr als nichts ($LRX^2 = 2644$).

Haben Geschlecht, Alter und Klasse einen Einfluß auf die Überlebenswahrscheinlichkeit? Dies kann mit einem Modell, das alle Zwei–Wegs–Interaktionseffekte enthält beantwortet werden:

```
Poisson regression                               Number of obs   =      24
                                                LR chi2(17)     =    3599.38
                                                Prob > chi2     =      0.0000
Log likelihood = -107.39063                    Pseudo R2      =      0.9437
```

n	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
survived	1.701681	.3329435	5.111	0.000	1.049123	2.354238
sex	4.860522	.3528853	13.774	0.000	4.16888	5.552165
age	4.016384	.5324094	7.544	0.000	2.972881	5.059888
class2	1.108557	.2606141	4.254	0.000	.5977625	1.619351
class3	3.714516	.543737	6.831	0.000	2.648811	4.780221
class4	4.349522	.5262514	8.265	0.000	3.318089	5.380956
cl3age	-2.127791	.4734647	-4.494	0.000	-3.055765	-1.199818
cl4age	-1.771056	.4641749	-3.815	0.000	-2.680822	-.8612904
cl2sex	-2.951249	.250822	-11.766	0.000	-3.442851	-2.459647
cl3sex	-3.143253	.2631536	-11.945	0.000	-3.659025	-2.627481
cl4sex	-3.077534	.2449579	-12.564	0.000	-3.557643	-2.597426
agesex	-.0364887	.244036	-0.150	0.881	-.5147905	.4418131
suvcl2	.8497454	.1576442	5.390	0.000	.5407684	1.158722
suvcl3	-.3023453	.1785027	-1.694	0.090	-.6522042	.0475136
suvcl4	-.8417544	.1484866	-5.669	0.000	-1.132783	-.550726
suvage	-.5114478	.2839172	-1.801	0.072	-1.067915	.0450198
suvsex	-2.424243	.1405017	-17.254	0.000	-2.699622	-2.148865
_cons	-2.336675	.5901384	-3.960	0.000	-3.493325	-1.180025

Die Verbesserung des Modells durch die Zusammenhangshypothesen beträgt:

$$X_{\text{Diff}}^2 = -2(-585 - (-107)) = 956$$

Dies ist bei 5 Freiheitsgraden (fünf zusätzlichen Koeffizienten) eine signifikante Verbesserung. Allerdings zeigt der X_L^2 von 87 mit 6 Freiheitsgraden (nicht im Output), dass noch Raum für Verbesserungen des Modells vorhanden ist. Dennoch können wir schon einmal feststellen, dass sich die Koeffizienten der Interaktionsterme mit dem „Überleben“ des Loglinearen Modells nicht von den Koeffizienten der logistischen Regression auf Seite 49 unterscheiden. Wir können diese Koeffizienten so interpretieren, als sie das Überleben die abhängige Variable eines Logit–Modells.

Was könnte man noch verbessern? Vielleicht hat das Geschlecht in den einzelnen Klassen eine Unterschiedliche Auswirkung auf die Überlebenswahrscheinlichkeit?

scheinlichkeit. Dies entspricht den Dreiwegs-Interaktionseffekten zwischen Klasse, Geschlecht und „Überleben“:

```
Poisson regression                                Number of obs =          24
                                                LR chi2(20)   =       3664.81
                                                Prob > chi2   =          0.0000
Log likelihood = -74.677561                    Pseudo R2     =          0.9608
```

	n	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
<output omitted>						
suvcl2		1.636486	.8003559	2.045	0.041	.0678171 3.205155
suvcl3		-.1514236	.687893	-0.220	0.826	-1.499669 1.196822
suvcl4		-2.121745	.6363603	-3.334	0.001	-3.368989 -.8745021
suvage		-.5301487	.2648613	-2.002	0.045	-1.049267 -.0110301
suvsex		-3.146902	.6245272	-5.039	0.000	-4.370953 -1.922852
susecl2		-1.062859	.8196262	-1.297	0.195	-2.669297 .5435783
susecl3		-.6647462	.7258655	-0.916	0.360	-2.087416 .7579241
susecl4		1.735867	.6515598	2.664	0.008	.4588329 3.0129
_cons		-2.962383	.7699923	-3.847	0.000	-4.471541 -1.453226

Tatsächlich: Die Männer der 1. Klasse haben eine kleinere Überlebenswahrscheinlichkeit als ihre Frauen, während die Männer der 3. Klasse eine höhere Überlebenswahrscheinlichkeit haben als ihre Frauen. Mit anderen Worten: Hätte sich Leonardo di Caprio in ein Mädchen der dritten Klasse verliebt, müßte die Geschichte anders herum ausgehen!

5.4 Literatur

- Agresti, Alan, 1990: *Categorical Data Analysis*. New York usw.: Wiley.
- Andreß, Hans-Jürgen, Jacques A. Hagenaars, und Steffen Kühnel, 1997: *Analyse von Tabellen und kategorialen Daten. Log-lineare Modell, latente Klassenanalyse, logistische Regression und GSK-Ansatz*. Berlin usw.: Springer.
- DeMaris, Alfred, 1992: *Logit, Modeling. Practical Applications*. Newbury Park usw.: Sage.
- Fienberg, Stephen E., 1980: *The Analysis of Cross-Classified Categorical Data*. 2. Aufl. Cambridge, MA: MIT Press.
- Knoke, D. und P. J. Burke, 1980: *Log-linear Models*. Sage University Paper 53. Beverly Hills: Sage.

- Wickens, Thomas D., 1989: Multiway Contingency Tables Analyses for the Social Sciences. Hillsdale usw.: Erlbaum.

Kapitel 6

Ereignisdatenanalyse

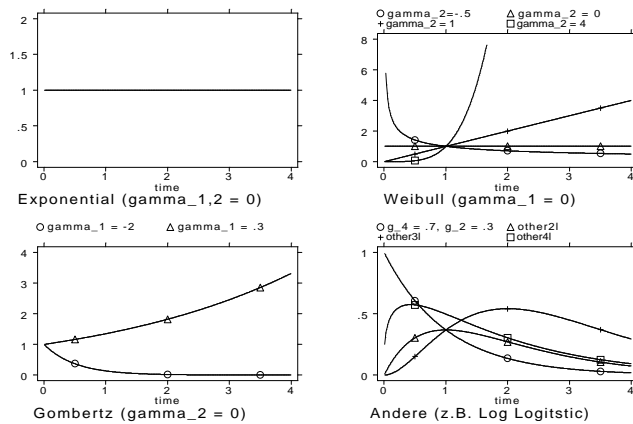
Ereignisdatenanalysen folgen dem bekannten Muster „linearer“ Modelle wie die bisher behandelten Ansätze. Abhängige Variable ist nunmehr jedoch nicht ein bestimmter Wert (bzw. eine Transformation eines Wertes) sondern eine Funktion: die „Hazardrate“ oder „Übergangsrate“:

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t + \Delta t > T \geq t | T \geq t)}{\Delta t} \quad (6.1)$$

mit T dem Zeitraum den es braucht, bis eine Veränderung eines bestimmten Merkmals eintritt. Der Ausdruck im Zähler von (6.1) bedeutet dann die Wahrscheinlichkeit, dass ein Ereignis im Zeitintervall Δt auftritt, unter der Bedingung, dass es bisher noch nicht aufgetreten ist. Die Hazardrate bezeichnet damit das momentane Risiko einen bestimmten „Zustand“ zu verlassen. In Parametrischen Modelle wird versucht diese Hazardrate in Abhängigkeit von Kovariaten und der Zeit zu modellieren:

$$\ln(h_{jk}(t)) = \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki} + \gamma_1 t + \gamma_2 \ln(t) \quad (6.2)$$

(6.2) besteht aus zwei Teilen. Erstens einem Teil, der wie die übliche Regressionsgleichung aussieht und zweitens einem zusätzlichen Teil, der die Abhängigkeit der Hazardrate von der Zeit enthält. Bei der Anwendung solcher Modell muß man sich zunächst darüber im klaren sein, welcher Zusammenhang zwischen der Hazardrate und der Zeit besteht. Nachfolgende Graphik enthält einige typische Verläufe.



Typisches Beispiel ist das Risiko den Zustand der „Lebendigkeit“ zu verlassen und in den Zustand „Todsein“ zu wechseln. Hierfür wird man vielleicht eine Weibull-Funktion mit $\gamma_2 > 1$ unterstellen. Andere Beispiele wären das Risiko einer Scheidung (Log logistische Rate, erst ansteigend, danach zurückgehend), einer Heirat (ebenfalls Log logistisch), einer Geburt usw.

Ist der Verlauf der Hazardfunktion bekannt, bzw. als bekannt vorausgesetzt kann man den Effekt von Kovariaten in der üblichen Weise bestimmen. Positive Effekte erhöhen das (logarithmierte) momentane Risiko einen bestimmten Zustand zu verlassen, Negative Effekte schmälern es.

Neben parametrischen Modellen gibt es auch nicht-parametrische Modelle, bei denen der Verlauf der Hazardfunktion unbekannt bleibt. Bekannt ist das Proportional-Hazard-Modell von Cox („Cox-Regression“). In diesem bleibt die „Base-line-Hazardfunktion“ unbekannt und es wird modelliert, wie sich diese unbekannte Hazardfunktion für unterschiedliche Ausprägungen der Kovariaten verschiebt. Nachteil dieses Modells ist allerdings, dass alle Effekte der Kovariaten proportional sind, d.h. unabhängig von der Zeit.

6.1 Literatur

- Blossfeld, Hans-Peter und Götz Rohwer, 1995: Techniques of Event History Modelling. Hillsdale: Lawrence Erlbaum.
- Diekmann, Andreas Mitter Peter, 1984: Methoden zur Analyse von Zeitverläufen. Stuttgart: Teubner.

Kapitel 7

Skalierungs- und Klassifikationsverfahren

7.1 Nochmal: Faktorenanalyse (hier: Hauptkomponentenanalyse)

Ziel von Faktorenanalysen ist die Reduktion der Zahl von Variablen eines Datensatzes. Der Begriff „Faktorenanalyse“ bezeichnet dabei keinesfalls ein einheitliches Verfahren, sondern eine Gruppe von Verfahren, die das Ziel der Datenreduktion erreichen. Häufig angewandt wird die Hauptkomponentenanalyse (PCA). Dies ist im Prinzip eine lineare Transformation von p gemeinsamen beobachteten Variablen in p unkorrelierte Variablen — den Hauptkomponenten. Dabei werden die Hauptkomponenten so bestimmt, dass der erste Faktor das Maximum der gemeinsamen Varianz der Variablen aufweist usw.

Beispiel Für die Planeten unseres Sonnensystems liegen Angaben zur Anzahl der Ringe, der Distanz zur Sonne, dem Radius, der Masse, der Dichte und der Anzahl von Monde vor¹. Eine Hauptkomponentenanalyse gibt folgende Lösung:

(principal components; 6 components retained)				
Component	Eigenvalue	Difference	Proportion	Cumulative
1	4.61054	3.44197	0.7684	0.7684
2	1.16857	1.05780	0.1948	0.9632
3	0.11077	0.03823	0.0185	0.9816

¹Daten aus Kap. 12 in Hamilton, Lawrence C., (1993): Statistics with Stata 3. Belmont: Duxbury Press.

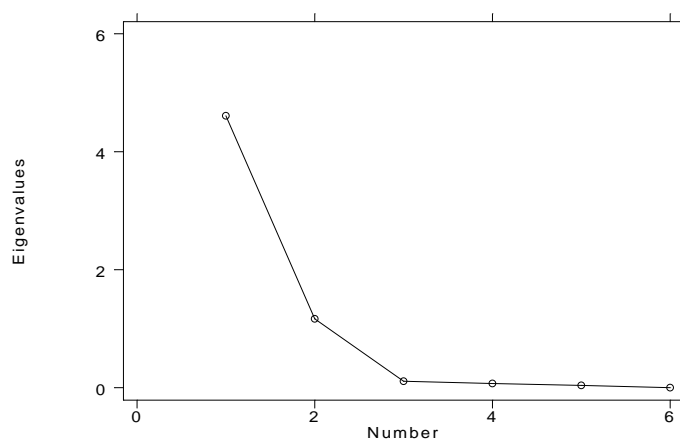
72KAPITEL 7. SKALIERUNGS- UND KLASSIFIKATIONSVERFAHREN

4	0.07254	0.03503	0.0121	0.9937		
5	0.03751	0.03745	0.0063	1.0000		
6	0.00006	.	0.0000	1.0000		
Variable	1	2	3	4	5	6
rings	0.45585	0.07489	0.27379	-0.10933	-0.83593	0.03004
logdsun	0.31415	-0.65523	0.57756	-0.18063	0.32482	-0.01642
lograd	0.42895	0.34890	-0.07231	-0.27164	0.30302	0.72349
logmass	0.38737	0.50679	0.11095	-0.21492	0.29709	-0.66810
logdens	-0.39372	0.43175	0.75622	0.23598	0.04690	0.16976
logmoons	0.45216	-0.00947	-0.04529	0.88304	0.11593	0.01495

Mit der Hauptkomponentenanalyse wird zunächst noch keine Datenreduktion erreicht. Aus p Variablen erhält man p voneinander unabhängige Hauptkomponenten (bzw. Faktoren). Darum beinhalten alle Typen von Faktorenanalysen Verfahren zur Bestimmung einer „hinreichenden“ Faktorenzahl:

- Das *Kaiserkriterium* bestimmt, dass nur diejenigen Faktoren verwendet werden sollen, welche Eigenwerte über 1 haben.
- Mit einem Screeplot können diejenigen Faktoren extrahiert werden, welche einen großen Zuwachs an Erklärungskraft bringen.
- Man kann soviel Faktoren extrahieren, dass man alle beteiligten Variablen gut mit den Faktoren erklären kann (*Kommunalitäten*).

Im vorliegenden Fall legt das Kaiserkriterium ein Lösung mit 2 Faktoren nah. Nach dem Screeplot wird man eher zu einer Lösung mit 3 Faktoren tendieren:



Bei Extraktion der ersten beiden Faktoren erhalten wir eine Angabe der Kommunalitäten (bzw. der Uniqueness = 1 - Communality).

7.1. NOCHMAL: FAKTORENANALYSE (HIER: HAUPTKOMPONENTENANALYSE)73

Factor Loadings			
Variable	1	2	Uniqueness
rings	0.97881	0.08096	0.03538
logdist	0.67455	-0.70831	0.04328
lograd	0.92105	0.37716	0.00941
logmass	0.83176	0.54784	0.00805
logdens	-0.84540	0.46673	0.06747
logmoons	0.97088	-0.01024	0.05730

Das zweite Problem betrifft die Interpretation der Bedeutung der Faktoren. Dies erfolgt zumeist durch Beurteilung der Ladungsmatrix. Dabei ist meist eine Rotation der Lösung vonnöten. Bei der Varimax-Rotation werden die Hauptkomponenten so gedreht, dass die Variablen entweder sehr hoch oder sehr niedrig auf die Faktoren laden. Die Orthogonalität der Hauptkomponenten bleibt dabei erhalten:

(varimax rotation)

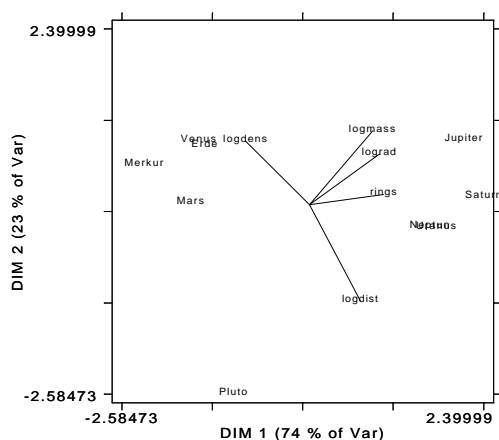
Rotated Factor Loadings			
Variable	1	2	Uniqueness
rings	-0.74746	0.63712	0.03538
logdsun	0.02680	0.97776	0.04328
lograd	-0.91682	0.38734	0.00941
logmass	-0.97492	0.20369	0.00805
logdens	0.26498	-0.92861	0.06747
logmoons	-0.67719	0.69578	0.05730

Bei der Promax-Rotation wird die Forderung der Unabhängigkeit der Faktoren aufgegeben:

(promax rotation)

Rotated Factor Loadings			
Variable	1	2	Uniqueness
rings	-0.70861	0.41692	0.03538
logdsun	0.23521	1.07127	0.04328
lograd	-0.95181	0.08352	0.00941
logmass	-1.05577	-0.13720	0.00805
logdens	0.10385	-0.91068	0.06747
logmoons	-0.61713	0.50643	0.05730

Ein recht mächtiges Verfahren ist die graphische Darstellung der Variablen und Beobachtungen im Raum der ersten beiden Hauptkomponenten (sog. Biplot, hier: GH-Biplot):



Die Position der Variablen entlang der X, bzw. Y Achse gibt die Ladung der Variablen auf die jeweilige Hauptkomponente wieder. Der Winkel der Vektoren approximiert die Korrelation der Variablen. Die Lage der Datenpunkte zu den Vektoren zeigt die vorhergesagte Position der Fälle auf den Variablen. Die Lage der Datenpunkte zu den Achsen zeigt die relative Position der Fälle auf den Hauptkomponenten. Die Lage der Punkte zueinander approximiert die „euklidische Distanz“ der Fälle. Damit dient der Biplot erfüllt diese Graphik auch Aufgaben der Clusteranalyse.

7.2 Hierarchische Clusteranalyse

Das Ziel einer Clusteranalyse ist die Zusammenfassung von Einzelobjekten zu Gruppen („Clustern“). Die Objekte innerhalb der Cluster sollen einander ähnlich sein, die Objekte in unterschiedlichen Clustern dagegen unähnlich. Ausgangspunkt von Clusternanalysen ist eine Distanzmatrix, i.d.R. die euklidische Distanz oder die quadrierte euklidische Distanz:

$$d_{ij} = \sqrt{\sum_{k=1}^K (x_{ik} - x_{jk})^2} \quad (7.1)$$

für alle $i, j = 1, \dots, n$. Dabei ist x_{ik} der Wert der i -ten Beobachtung auf der k -ten Variable und x_{jk} der Wert der j -ten Beobachtung auf derselben Variable.

Beispiel Gegeben sind die Planetendaten aus Abschnitt 7.1. Die Variablen haben folgende Werte:

rings logdsun loggrad logmass logdens logmoons

Merkur	0	4.058717	7.799344	54.15338	1.690096	0
Venus	0	4.683981	8.707813	56.84513	1.658228	0
Erde	0	5.007965	8.76061	57.05046	1.708378	1
Mars	0	5.428907	8.130942	54.81887	1.371181	1.693147
Jupiter	1	6.657112	11.18303	62.81165	.2700271	3.772589
Saturn	1	7.26333	11.0021	61.60592	-.3710637	3.833213
Uranus	1	7.961928	10.17141	59.72334	.1739534	3.70805
Neptun	1	8.411077	10.11658	59.89677	.5068176	3.079442
Pluto	0	8.682708	7.34601	50.75218	.1823216	1

Da die Variablen in unterschiedlichen Maßeinheiten gemessen wurden, müssen die Variablen zunächst standardisiert werden. Danach erhalten wir:

	srrings	slogdsun	slograd	slogmass	slogdens	slogmoon
Merkur	-.8432741	-1.393821	-1.025757	-.8675001	1.106278	-1.249954
Venus	-.8432741	-1.031151	-.3817934	-.173387	1.06672	-1.249954
Erde	-.8432741	-.8432322	-.344369	-.1204393	1.128971	-.6279639
Mars	-.8432741	-.5990747	-.7907051	-.6958917	.7104034	-.1968333
Jupiter	1.054093	.1133168	1.372751	1.365178	-.6564759	1.096558
Saturn	1.054093	.464939	1.244499	1.054261	-1.452272	1.134266
Uranus	1.054093	.8701445	.6556713	.5688066	-.7757338	1.056416
Neptun	1.054093	1.130663	.6168036	.6135278	-.3625442	.6654279
Pluto	-.8432741	1.288216	-1.3471	-1.744556	-.7653461	-.6279639

Nun kann die euklidische Distanz berechnet werden, z.B. zwischen Merkur und Venus:

$$d_{1,2} = \sqrt{(-.843 - (-.843))^2 + (-.139 - (-1.031))^2 + \dots + (-1.250 - (-1.250))^2} = 1.01$$

Berechnet man die euklidische Distanz für alle $i, j = 1, \dots, n$ erhält man folgende Distanzmatrix:

	Merkur	Venus	Erde	Mars	Jupiter	Saturn	Uranus	Neptun	Pluto
Merkur	0.00								
Venus	1.01	0.00							
Erde	1.31	0.66	0.00						
Mars	1.41	1.36	0.98	0.00					
Jupiter	5.02	4.34	3.98	4.07	0.00				
Saturn	5.30	4.69	4.36	4.29	0.93	0.00			
Uranus	4.74	4.19	3.80	3.64	1.32	1.10	0.00		
Neptun	4.55	3.95	3.59	3.49	1.56	1.56	0.63	0.00	
Pluto	3.46	3.54	3.43	2.71	5.00	4.73	4.00	3.86	0.00

Aus der Distanzmatrix werden nun zunächst die Fälle mit der geringsten Distanz ermittelt. Hier handelt es sich um Uranus und Neptun ($d_{8,7} = .63$). Diese beiden Fälle werden darum zu einem „Cluster“ verschmolzen.

76 KAPITEL 7. SKALIERUNGS- UND KLASSIFIKATIONSVERFAHREN

Nach der Verschmelzung von Uranus und Neptun zu einem „Cluster“ müssen die Distanzen dieses Clusters zu den übrigen Clustern ermittelt werden. Dafür liegen unterschiedliche Verfahren vor. Hier soll zunächst das „*Complete-Linkage*“-Verfahren angewandt werden. Danach werden die am weitesten entfernten Objekte innerhalb eines Clusters zur Beurteilung der Entfernung zu den anderen Objekten verwendet. Mit der Complete-Linkage-Methode ergibt sich folgende neue Distanzmatrix:

	Merkur	Venus	Erde	Mars	Jupiter	Saturn	UrNep	Pluto
Merkur	0.00							
Venus	1.01	0.00						
Erde	1.31	0.66<-	0.00					
Mars	1.41	1.36	0.98	0.00				
Jupiter	5.02	4.34	3.98	4.07	0.00			
Saturn	5.30	4.69	4.36	4.29	0.93	0.00		
UrNep	4.74	4.19	3.80	3.64	1.56	1.56	0.00	
Pluto	3.46	3.54	3.43	2.71	5.00	4.73	4.00	0.00

In der neuen Distanzmatrix ist Erde und Venus das nächste Verschmelzungspaar. Nach der Verschmelzung und Anwendung von Complete-Linkage ergibt sich:

	Merkur	VeEr	Mars	Jupiter	Saturn	UrNep	Pluto
Merkur	0.00						
VeEr	1.31	0.00					
Mars	1.41	1.36	0.00				
Jupiter	5.02	4.34	4.07	0.00			
Saturn	5.30	4.69	4.29	0.93<-	0.00		
UrNep	4.74	4.19	3.64	1.56	1.56	0.00	
Pluto	3.46	3.54	2.71	5.00	4.73	4.00	0.00

USW.:

	Merkur	VeEr	Mars	SaJu	UrNep	Pluto
Merkur	0.00					
VeEr	1.31<-	0.00				
Mars	1.41	1.36	0.00			
SaJu	5.30	4.69	4.29	0.00		
UrNep	4.74	4.19	3.64	1.56	0.00	
Pluto	3.46	3.54	2.71	5.00	4.00	0.00

	MeVeErd	Mars	SaJu	UrNep	Pluto
MeVeErd	0.00				
Mars	1.41<-	0.00			
SaJu	5.30	4.29	0.00		
UrNep	4.74	3.64	1.56	0.00	
Pluto	3.54	2.71	5.00	4.00	0.00

	MeVeErMa	SaJu	UrNep	Pluto
MeVeErMa	0.00			
SaJu	5.30	0.00		
UrNep	4.74	1.56<-	0.00	
Pluto	3.54	5.00	4.00	0.00

	MeVeErMa	SaJuUrNe	Pluto
MeVeErMa	0.00		
SaJuUrNe	5.30	0.00	
Pluto	3.54<-	5.00	0.00

	MeVeErMaPl	SaJuUrNe
MeVeErMaPl	0.00	
SaJuUrNe	5.30<-	0.00

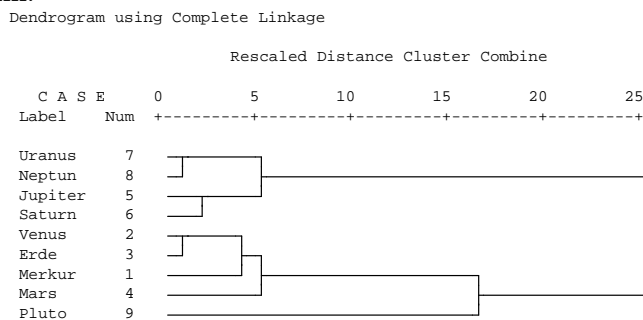
Die Verschmelzung von MeVeErMaPl mit SaJuUrNe führt zum Ende der Prozedur, da nunmehr nur ein Cluster vorliegt.

Wie das Beispiel zeigt ergibt sich aus der Clusteranalyse selbst kein Hinweis darüber, wieviele Cluster vorliegen. Ein Hinweis über die Anzahl der Cluster kann das Verschmelzungsschema bzw. dessen graphische Darstellung, das Dendrogramm, liefern:

Verschmelzungsschema:

Neptun	Uranus	8	.63	.00
Venus	Erde	7	.66	.03
Jupiter	Saturn	6	.93	.17
Merkur	Venus-Erde	5	1.31	.38
Mars	Merkur-Venus-Erde	4	1.41	.10
Uranus-Neptun	Saturn-Jupiter	3	1.56	.14
Pluto	Merkur-Venus-Erde-Mars	2	3.54	1.98
Saturn-Jupiter-Uranus-Neptun	Merkur-Venus-Erde-Mars-Pluto	1	5.30	1.86

Dendrogramm:



Sinnvoll erscheint eine Lösung mit 2 Clustern und dem Planeten Pluto als Außreißer. Diesselbe Lösung wurde auch durch den Biplot nahegelegt.

Varianten der Clusteranalyse verwenden unterschiedliche Distanzmaße und unterschiedliche Fusionskriterien. Die divergierenden Distanzmaße entstammen der sog. „Minowski-r-Metriken“:

$$d_{ij} = \left[\sum_{k=1}^K (x_{ik} - x_{jk})^r \right]^{\frac{1}{r}} \quad (7.2)$$

Andere Fusionskriterien sind:

- Single - Linkage: Fusioniert werden die Cluster, die den nächsten Nachbar aufweisen („nearest neighbour“)
- Complete - Linkage: s.o.
- Average - Linkage: Fusioniert werden die Cluster, bei denen der Durchschnitt der Distanzen aller Objekte eines Clusters am kleinsten ist.
- Medianverfahren: Fusion derjenigen Cluster, deren quadriertes, euklidischer Centroidabstand minimal ist (Clustercentroid=durchschnittliche Merkmalsausprägungen aller Objekte eines Clusters)
- Ward-Verfahren: Fusion derjenigen Cluster, mit deren Fusion die geringste Erhöhung der gesamten Fehlerquadratsumme einhergeht.

7.3 Multidimensionale Skalierung

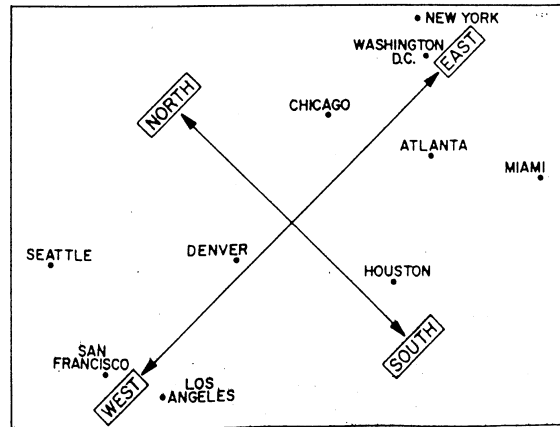
Die Aufgabe einer MDS kann am einfachsten anhand von Positionen von Städten auf Landkarten verdeutlicht werden. Die Entfernungen zwischen den Städten können in eine Distanzmatrix geschrieben werden:

CITIES	ATLA	CHIC	DENV	HOUS	L.A.	MIAMI	N.Y.	S.F.	SEAT	WASH D.C.
ATLANTA		587	1212	701	1936	604	748	2139	2182	543
CHICAGO	587		920	940	1745	1188	713	1858	1737	597
DENVER	1212	920		879	831	1726	1631	949	1021	1494
HOUSTON	701	940	879		1374	968	1420	1645	1891	1220
LOS ANGELES	1936	1745	831	1374		2339	2451	347	959	2300
MIAMI	604	1188	1726	968	2339		1092	2594	2734	923
NEW YORK	748	713	1631	1420	2451	1092		2571	2408	205
SAN FRANCISCO	2139	1858	949	1645	347	2594	2571		678	2442
SEATTLE	2182	1737	1021	1891	959	2734	2408	678		2329
WASHINGTON D.C.	543	597	1494	1220	2300	923	205	2442	2329	

(B) AIRLINE DISTANCES BETWEEN TEN U.S. CITIES

Die Aufgabe der MDS besteht nun darin, die Punktekonfiguration der Distanzmatrix in einen niederdimensionalen Raum derart darzustellen, dass

die Abstände zwischen den Punkten in diesem Raum den ursprünglichen Distanzen so ähnlich wie möglich werden. Eine Anwendung der MDS auf die obige Distanzmatrix ergibt z.B. folgende Darstellung:



In der Regel besteht das Problem einer MDS darin, dass die Zahl der Dimensionen, die der Distanzmatrix zugrundeliegt nicht bekannt ist. Darüber hinaus kann muß bei sozialwissenschaftlichen Daten meist mit Fehlern bei der Datenerhebung rechnen. In der Praxis entscheidet man sich darum zunächst vorläufig für eine bestimmte Anzahl von Dimensionen. Im Raum dieser Dimensionen wird dann eine Anfangskonfiguration bestimmt. Danach wird ein Abweichungsmaß definiert, welches sodann mit numerischen Standardtechniken minimiert wird. Die Konfiguration der Objekte am Ende des Minimierungsvorganges ist die gesuchte MDS-Lösung.

Das Abweichungsmaß der MDS ist definiert durch

$$\text{STRESS}_1 = \sqrt{\frac{\sum (\delta - d)^2}{\sum d^2}} \quad (7.3)$$

wobei über alle Distanzen summiert wird. d sind die Distanzen in der MDS-Lösung, δ modifizierte Distanzen der Ausgangsmatrix.

Die Zahl der Dimensionen wird durch den Vergleich der STRESS-Werte von MDS-Lösungen unterschiedlicher Dimensionalität festgelegt. Üblich sind Plots der STRESS-Werte gegen die Zahl der Dimensionen (vergleichbar den scree-Plot bei der Faktorenanalyse)

7.4 Literatur

- Hauptkomponentenanalyse

80 KAPITEL 7. SKALIERUNGS- UND KLASSIFIKATIONSVERFAHREN

- Dunteman, George H., 1989: Principal Component Analysis. Newbury Park usw.: Sage.
- Fox, John, 1997: Applied regression analysis, linear models, and related methods. Thousand Oaks usw. Sage.
- Hamilton, Lawrence C., 1992: Regression with Graphics. A Second Course in Applied Statistics. Belmont: Wadsworth.
- Clusteranalyse
 - Aldenderfer, M. S. und R. K. Blashfield, 1984: Cluster Analysis. Newbury Park usw.: Sage.
 - Bacher, Johann, 1994: Clusteranalyse: München und Wien.
- MDS
 - Bacher, Johann, 1989: Einführung in die Logik der Skalierung.
 - Kruskal, Joseph B. und Myron Wish, 1978: Multidimensional Scaling. Beverly Hill u. London: Sage.