



ALM

Allgemeines lineares Modell II

Teil 2

Logit- und Probit-Analyse
Multivariate Analysen
Hierarchische Regression
Messwiederholungen
Nichtlineare Regression
Homogenitäts-Test (**Johann Bacher**)

Kurt Holm

Almo Statistik-System

www.almo-statistik.de

holm@almo-statistik.de

kurt.holm@jku.at

2014

Autor: em. Prof. Dr. Kurt Holm, Universität Linz, Österreich

Nachfolgend wird häufig auf das Dokument **P0** Bezug genommen. Dabei handelt es sich um das Almo-Dokument "Arbeiten mit Almo.PDF". Es kann in Almo heruntergeladen werden

Weitere Almo-Dokumente

Die folgenden Dokumente können alle von der Handbuchseite in www.almo-statistik.de heruntergeladen werden

0. Arbeiten mit Almo.PDF (1 MB)
- 1a. Eindimensionale Tabellierung.PDF (1,8 MB)
- 1b. Zwei- und drei-dimensionale Tabellierung.PDF (1.1 MB)
2. Beliebig-dimensionale Tabellierung.PDF (1.7 MB)
3. Nicht-parametrische Verfahren.PDF (0.9 MB)
4. Kanonische Analysen.PDF (1.8 MB)
Diskriminanzanalyse.PDF (1.8 MB)
enthält: Kanonische Korrelation, Diskriminanzanalyse, bivariate Korrespondenzanalyse, optimale Skalierung
5. Korrelation.PDF (1.4 MB)
6. Allgemeine multiple Korrespondenzanalyse.PDF (1.5 MB)
7. Allgemeines ordinales Rasch-Modell.PDF (0.6 MB)
- 7a. Wie man mit Almo ein Rasch-Modell rechnet.PDF (0.2 MB)
8. Tests auf Mittelwertsdifferenz, t-Test.PDF (1,6 MB)
9. Logitanalyse.pdf (1,2MB) enthält Logit- und Probitanalyse
- 9b. Bootstrap bei Logit- und Probitanalyse.pdf
10. Koeffizienten der Logitanalyse.PDF (0,06 MB)
11. Daten-Fusion.PDF (1,1 MB)
12. Daten-Imputation.PDF (1,3 MB)
13. ALM Allgemeines Lineares Modell.PDF (2.3 MB)
- 13a. ALM Allgemeines Lineares Modell II.PDF (2.7 MB)
- 13b. Bootstrap bei Allgemeinem Linearem Modell III.PDF
14. Ereignisanalyse: Sterbetafel-Methode, Kaplan-Meier-Schätzer, Cox-Regression.PDF (1,5 MB)
15. Faktorenanalyse.PDF (1,6 MB)
- 15a. Bootstrap bei Faktorenanalyse.PDF
16. Konfirmatorische Faktorenanalyse.PDF (0,3 MB)
17. Clusteranalyse.PDF (3 MB)
18. Pisa 2012 Almo-Daten und Analyse-Programme.PDF (17 KB)
19. Guttman- und Mokken-Skalierung.PFD (0.8 MB)
20. Latent Structure Analysis.PDF (1 MB)
21. Statistische Algorithmen in C (80 KB)
22. Conjoint-Analyse (PDF 0,8 MB)
23. Ausreisser entdecken (PDF 170 KB)
24. Statistische Datenanalyse Teil I, Data Mining I
25. Statistische Datenanalyse Teil II, Data Mining II
26. Statistische Datenanalyse Teil III, Arbeiten mit Almo-Datenanalyse-System
27. Mehrfachantworten. Tabellierung von Fragen mit Mehrfachantworten
28. Metrische multidimensionale Skalierung (MDS) (0,4 MB)
29. Metrisches multidimensionales Unfolding (MDU) (0,6 MB)
30. Nicht-metrische multidimensionale Skalierung (MDS) (0,4 MB)
31. Pfadanalyse.PDF (0,7 MB)
32. Datei-Operationen mit Almo (1,1 MB)
33. Wählerstromanalyse und Wahlhochrechnung (1,6 MB)
34. Soziometrie. Auswertung soziometrischer Daten (0,5 MB)
35. Konfidenzintervall und p-Wert beim Bootstrap-Verfahren (200 KB)

Inhaltsverzeichnis

P20.9.3.3 Logit- und Probit-Analyse als Kleinste-Quadrate-Schätzung	5
P20.9.4 Multivariate Analyse: Mehrere abhängige quantitative/ordinale Variable.....	18
P20.9.4.1 Kalkül der multivariaten Analyse	18
P20.9.4.2 Beispiel einer multivariaten Analyse	27
P20.9.5 Multivariate Analyse: Eine nominale Variable als abhängige Variable.....	35
P20.9.5.1 Tabellenanalyse mit polytomen Variablen.....	35
P20.9.5.2 Die Diskriminanzanalyse (lineares Wahrscheinlichkeitsmodell)	42
P20.9.5.3 Prognostizierte Wahrscheinlichkeiten in der linearen Wahrscheinlichkeitsanalyse ..	46
P20.9.5.4 Diskriminanzanalyse mit unabhängigen quantitativen und nominalen Variablen (lineares Wahrscheinlichkeitsmodell)	50
P20.9.7 Allgemeines Lineares Modell mit Rangvariablen	58
P20.9.7.1 Zum Begriff der "Rangvariablen.....	58
P20.9.8 Analysen mit vielen unabhängigen nominalen Variablen.....	59
P20.10 Hierarchische Regression	60
P20.10.1 Gruppenweise hierarchische Regression.....	62
P20.10.2 Regression von Polynomen als hierarchische Regression	64
P20.10.3 Gründe für die Hierarchisierung	65
P20.11 Gleichrangige Gruppen bei den Kovarianten: Partielle multiple Korrelation	65
P20.12 Hierarchische u. gleichrangige Gruppen bei quantitativen u. ordinalen Kovarianten	67
P20.13 Frei gewählte Hierarchie bei den nominalen Variablen	70
P20.13.1 Kovarianzanalyse und multivariate Analyse mit frei gewählter Hierarchie	72
P20.14 Bildung von Interaktionsvariablen.....	73
P20.14.1 Einbeziehung von nur einigen Interaktionsvariablen	75
P20.14.2 Leere Zellen und Interaktionen.....	77
P20.14.3 Zelleneffekte	84
P20.15 Die Verwendung der PARTIAL-Anweisung.....	86
P20.15.1 Die einseitige Anpassung von Kovarianten und nominalen Variablen in der Kovarianzanalyse.....	89
P20.16 Unvollständige, geschachtelte und hierarchische Versuchspläne	91
P20.16.1 Unvollständige Versuchspläne	91
P20.16.2 Geschachtelte Variable	91
P20.16.3 Hierarchische Versuchspläne	93
P20.17 Analysen mit wiederholten Messungen	95
P20.17.0 Uni- und multivariater Ansatz	95
P20.17.0.1 Der multivariate Ansatz	96
P20.17.1 Univariater Ansatz	100
P20.17.1.1 Exkurs: Zuverlässigkeit und wiederholte Messungen:	107
P20.17.6 Aggregatdaten-Analyse (z.B. Wahlanalyse)	108

P20.18 Johann Bacher: Interaktionen zwischen nominalen und quantitativen/ordinalen Variablen	113
P20.19 Johann Bacher: Überprüfung der Varianzhomogenität	115
P20.19.1 Modelle der Varianzheterogenität	115
P20.19.2 Dateneingabe.....	117
P20.19.3 Interpretation der Ergebnisse.....	119
P20.20 Johann Bacher: Homogenität der Regressionskoeffizienten	120
P20.20.1 Test auf Homogenität der Regressionskoeffizienten	120
P20.20.2 Modellspezifikationen: Modelle mit unterschiedlichen Kovariaten je Merkmalskombination der nominalen Variable(n).....	124
P20.21 Nichtlineare Regression mit Prog20	126
P20.21.1 Die parabolische und hyperbolische Funktion	126
P20.21.1.1 Eigenschaften der parabolischen Funktion.....	127
P20.21.1.2 Eigenschaften der hyperbolischen Funktion.....	127
P20.21.1.3 Schätzung der Parameter durch lineare Regression	127
P20.21.1.4 Eingabe in Almo-Maskenprogramm.....	127
P20.21.1.5 Das selbst geschriebene Almo-Programm.....	131
P20.21.1.6 Das Ergebnis.....	131
P20.21.2 Die Exponential-Funktion	134
P20.21.3 Die Gompertz-Kurve	136
P20.21.4 Die logistische Funktion	137
P20.21.5 Literatur	139
Schlagwortverzeichnis.....	137
Literatur zum Allgemeinen Linearen Modell	141

P20.9.3.3 Logit- und Probit-Analyse (Kleinste-Quadrate-Schätzung)

Die Logit und die Probit-Analyse werden angewendet, wenn die abhängige Variable nominal ist. Dabei gilt für die Probit-Analyse, dass die abhängige nominale Variable dichotom sein muss.

Almo ermöglicht es, im Rahmen des Allgemeinen Linearen Modells, eine auf der Kleinste-Quadrate-Schätzung beruhende Logit- oder Probit-Analyse zu rechnen. Diese Kleinste-Quadrate-Schätzung wird in der Literatur auch gelegentlich "Minimum-Chi-Quadrate-Schätzung" genannt. Siehe etwa Aldrich/Nelson (1984, S.66).

In der (vorallem sozialwissenschaftlichen) Forschungspraxis wird eher die **Maximum-Likelihood-Schätzung** der Logit-Probit-Analyse vorgezogen. Sie ist in Almo durch die Programm-Masken Prog22m und Prog22mb realisiert. Ausführlich beschrieben wird sie im Handbuch Teil 4 Fortgeschrittene Verfahren, Abschnitt P22 und im separaten PDF-Dokument "Logit-Analyse.PDF".

Die Kleinste-Quadrate-Schätzung erbringt andere (manchmal erheblich andere) Ergebnisse als die nach der Maximum-Likelihood-Schätzung gerechnete Analyse. Der Benutzer rechne die beiden Beispiel-Programme Logit_3.Alm und Arm102a.Alm. Beide rechnen eine Logitanalyse mit denselben Daten; das erste mit Kleinster-Quadrate-Schätzung, das zweite mit ML-Schätzung. Sie finden diese Programme durch Klick auf das Menü „Almo/Liste aller Almo-Programme“.

Betrachten wir ein Beispiel: Die unabhängigen nominalen Variablen seien Geschlecht (2 Ausprägungen) und Beruf (3 Ausprägungen). Die unabhängige quantitative Variable sei das Einkommen (V5), das in 3 Ausprägungen unterteilt sei. Die abhängige dichotomische Variable sei Kaufabsicht (ja, nein).

Wir rechnen folgendes Almo-Programm:

Eingabe in Maskenprogramm

Prog20mj.Msk Kurzprogramm
Allgemeines lineares Modell: Logit- oder Probit-Analyse
Kleinste-Quadrate-Schätzung

Unabhängige Variable: nominal und quantitativ
Abhängige Variable : nominal

Beispiel: Der Einfluss der nominalen Variablen Geschlecht (U1) und Beruf (U3) sowie der quantitativen Variablen Einkommen (U5) auf die nominale Variable U10 (Kauf:Ja,Nein) soll ermittelt werden

Selbstverstaendlich kann die Logit- bzw. Probit-Analyse auch nur mit einer unabhaengigen nominalen oder quantitativen Variablen gerechnet werden.

Die abhaengige nominale Variable koennte auch mehr als 2 Auspraegungen besitzen, z.B. "Kauf:Mercedes,UW,Opel,Ford".

Almo rechnet in diesem Falle eine "multivariate" Analyse und gibt dabei auch Koeffizienten aus (wie etwa Wilks Lambda) die bei Logit- und Probit-Analyse ueblicherweise nicht ausgegeben werden

Siehe auch die Maximum-Likelihood-Schaetzung des Logit- und Probit-Modells in Prog22m und die Diskriminanzanalyse mit Prog20mi, sowie die kanonische Diskriminanzanalyse mit Prog29m3

Siehe Handbuch, Abschnitt P20.9.3.2

Was ist ein Kurzprogramm ? -->
Bedienung -->

- 1
Vereinbare Variable= 22 ;
- 2 Option: Weitere Vereinbarungen - nur wenn Almo dazu auffordert
- 3
 "C:\Almo7\TESTDAT\Uarnamen.nam"
 zeige = Namensdatei in Output zeigen
leer = nicht

Freie Namensfelder **Hilfe**

unabhängige nominale Variable

unabhängige quantitative Variable

abhängige nominale Variable. Sie darf auch mehr als 2 Ausprägungen besitzen

4 BEACHTEN: Als Nummern für die Dummies der abhängigen nominalen Variablen müssen freie, sonst nicht verwendete Variablennummern verwendet werden **Hilfe**

Die abhäng. nominale Variable "Kauf" wird im Programmverlauf in Dummies aufgelöst. Wir nennen sie kurz "KaufJa" u. "KaufNein". Für diese Dummies werden die freien, nicht durch eine andere Analyse-Variablen belegte Variablen-Nummern 21 und 22 verwendet

erzeuge zusätzliche Namensfelder

Datei aus der gelesen wird **Hilfe**

bei Datei-Problemen

Format der Daten **Hilfe**

der Datensatz enthält diese Variablen
Bei Format DIREKT schreiben Sie: alle_U

Wenn Dateiformat FIX oder Nicht-Standard-FREI **Hilfe**

Modell

Logit oder Probit

Analyse-Variablen: Unabhängige Variable **Hilfe**

unabhängige nominale Variable

unabhängige quantitative Variable

9

Analyse-Variabile: Unabhängige Variable

unabhängige nominale Variable Hilfe

Geschlecht, Beruf

unabhängige quantitative Variable

Einkomm

10

Analyse-Variabile: Abhängige nominale Variable Hilfe

Kauf

Kauf Ja, Kauf Nein

Dummies der abhängigen nominalen Variablen

11

Option: Ein- und Ausschliessen von Untersuchungseinheiten

12

Loesche wieder diese Box

Umkodierungen und Kein-Wert-Angaben Hilfe

Umkodierungen Hilfe
Kein_Wert-Angabe Hilfe

Einkomm(1:3=1; 3:6=2; 6:9=3)

erzeuge zusätzliche Felder für Umkodierungen / Kein_Wert-Angaben

Kontrollieren, ob Umkodierung so erfolgt wie gewünscht Hilfe

diese Variablen ...

... aus diesen Datensätzen vor und nach der Umkodierung zur Kontrolle anzeigen

13

Unabh. quant. Variable: Diskret oder kontinuierlich (Option 40)

1

1=die Ausprägungen der unabh. quant. Var. werden für die Bildung der Merkmalskombinationen mitverwendet (optimal)

BEACHTEN: Die unabh. quant. Variablen müssen in diesem Falle ganzzahlig sein. Ist dies nicht der Fall, dann werden sie von ALMO zwangsweise umkodiert

2=die unabh. quant. Variablen werden als Kovariate in das Modell aufgenommen. In diesem Falle dürfen die unabh. quant. Variablen beliebig kontinuierlich sein (problematisch)

- 14
- Optionen**

möglich sind folgende Verfahren
 = w_squares_of_means
 = fitting_constants
 das sequentielle Verfahren ist ungewöhnlich

<Option 42>

1=die Logit- bzw. Probit-transformierten
 Daten zum Anschauen ausgeben
 0=nicht ausgeben

<Option 41>

nur bei Logit-Analyse
 und wenn abhängige Variable dichotom
 0=gewöhnliche Kleinste-Quadrate
 1=gewichtete Kleinste-Quadrate
- 15
- Grafik-Optionen**

Almo = Almo-Grafik ausgeben
 Excel = Tabelle für Excel-Graphik ausgeben
 Stanford= Tabelle für Stanford-Graphics
 0 = nichts
- 16
- Almo-Programmtext in Ergebnisliste zeigen oder nicht ?**

zeige = Almo-Programmtext zeigen

Editfeld leer = nicht zeigen

Der Rechengang der Logit- und Probit-Analyse (Kleinste-Quadrate-Schätzung)

Almo erstellt im 1. Anfang-Ende-Block des selbst geschriebenen Programms aus den Kombinationen der Ausprägungen der unabhängigen Variablen folgende Häufigkeitstabelle:
Häufigkeitstabelle

Geschlecht	Beruf	Einkommen	Kauf	
			Ja	Nein
1	1	1	0	3
1	1	2	0	5
1	1	3	0	0
1	2	1	4	4
1	2	2	4	4
1	2	3	2	0
1	3	1	2	5
1	3	2	0	1
1	3	3	0	0
2	1	1	0	2
2	1	2	1	2
2	1	3	2	1
2	2	1	3	2
2	2	2	3	1
2	2	3	2	0
2	3	1	3	0
2	3	2	1	3
2	3	3	1	0

Sätze zur Logit- und Probit-Analyse

1. Die Häufigkeitstabelle wird dadurch gebildet, dass die unabhängigen nominalen und eventuell auch die unabhängigen quantitativen Variablen in Ihren Ausprägungen kombiniert werden.
2. Wird in der Optionsbox "Unabhängige quantitative Variable: Diskret oder kontinuierlich" **Option 40 = 1** gesetzt, wie das oben in unserem Almo-Programm geschehen ist, dann werden die unabhängigen quantitativen Variablen mit Ihren Ausprägungen für die Bildung der Ausprägungskombinationen der Häufigkeitstabelle verwendet. Sie werden aber trotzdem als quantitative behandelt, d.h. es wird nur ein Regressionskoeffizient für sie berechnet - während die unabhängigen nominalen Variablen in Dummies aufgelöst werden, für die je ein Regressionskoeffizient ermittelt wird.
3. Option 40 kann nur dann =1 gesetzt werden, wenn die unabhängigen Variablen ganzzahlig sind und sich in Einer-Schritten von ihrer Werte-Untergrenze zu Ihrer Obergrenze bewegen. Sind sie das nicht, dann müssen sie entsprechend umkodiert werden. Geschieht dies nicht, dann werden die aufsteigenden Zahlenwerte der Variablen in die Ganzzahlen 1,2,3,..... von Almo zwangsweise programmintern umkodiert.
4. Option 40 kann nur dann =1 gesetzt werden, wenn dadurch nicht zu viele Ausprägungskombinationen entstehen, d.h. wenn dadurch nicht eine Häufigkeitstabelle entsteht, die zu viele unbesetzte Zellen besitzt.
5. Wird **Option 40 = 2** gesetzt, dann werden die unabhängigen quantitativen Variablen nicht für die Bildung der Häufigkeitstabelle herangezogen. Der 1. Anfang-Ende-Block liefert dann folgende Häufigkeitstabelle:

Häufigkeitstabelle 2

Geschlecht	Beruf	Einkommen	Kauf	
			Ja	Nein
1	1	1.6250	0	8
1	2	1.6667	10	8
1	3	1.1250	2	6
2	1	2.1250	3	5
2	2	1.7273	8	3
2	3	1.7500	5	3

Für die quantitative Variable "Einkommen" werden die Mittelwerte je Ausprägungskombination von Geschlecht und Beruf ermittelt und ausgegeben. Man wird Option 40=2 setzen, wenn die unabhängigen Variablen zu viele Ausprägungen besitzen und eine Zusammenfassung nicht gerechtfertigt erscheint. Wird nur eine unabhängige nominale Variable verwendet, dann entsteht in der Regel ein Modell mit 0 Freiheitsgraden (oder sogar einer negativen Zahl von Freiheitsgraden), wenn eine oder mehrere quantitative Variable unter Option 40=2 hinzugefügt werden. Das Modell ist also nicht mehr rechenbar.

6. Im Verlauf des Kalküls der Logit- bzw. Probitanalyse wird dann jede Zeile der obigen Häufigkeitstabelle als ein Datenvektor verstanden und dem spezifischen Kalkül des ALM in Prog20 unterworfen wird.
7. Die Häufigkeiten werden zunächst in Anteilswerte umgerechnet. In der 7. Zeile (der 1. Häufigkeitstabelle) haben wir KaufJa=2 und KaufNein=5. Das ergibt $p_1=2/7=0.2857$ und $p_2=5/7=0.7143$. Der 2. Anteilswert p_2 ist redundant (da $p_2=1-p_1$).
9. a. Der Ansatz der Logit-Analyse ist nun folgender (siehe Hartung/Elpelt, 1984, S. 131 ff.):

$$(1) p_1 = \frac{1}{1 + e^{-(k_0 * 1 + a_1 * A1 + b_1 * B1 + b_2 * B2 + c_1 * \text{Einkommen})}}$$

Durch Umformen erhält man

$$(1a) \ln\left(\frac{p_1}{1-p_1}\right) = k_0 * 1 + a_1 * A1 + b_1 * B1 + b_2 * B2 + c_1 * \text{Einkommen}$$

- b. Für die Probitanalyse lautet der Ansatz:

$$(2) p_1 = \text{Phi}(k_0 * 1 + a_1 * A1 + b_1 * B1 + b_2 * B2 + c_1 * \text{Einkommen})$$

Durch Umformen erhält man

$$(2a) \text{InvPhi}(p_1) = k_0 * 1 + a_1 * A1 + b_1 * B1 + b_2 * B2 + c_1 * \text{Einkommen}$$

p_1 = Anteilswert für KaufJa
 e = Exponentialfunktion
 k_0 = Effekt der Konstanten
 A_1 = 1. Dummy für Geschlecht (0 oder 1)
 a_1 = Regressionskoeffizient für 1. Dummy des Geschlechts
 B_1 = 1. Dummy für Beruf (0 oder 1)
 B_2 = 2. Dummy für Beruf (0 oder 1)
 b_1 = Regressionskoeffizient für 1. Dummy des Berufs
 b_2 = Regressionskoeffizient für 2. Dummy des Berufs
 c = Regressionskoeffizient für Einkommen
 Φ = Verteilungsfunktion der Standard-Normalverteilung
 InvPhi = deren Umkehrfunktion

c. Wir rechnen nun mit den Gleichungen 1a und 2a, ermitteln also für die abhängige Variable KaufJa

bei der Logitanalyse: $\ln(p_1/p_2)$

bei der Probitanalyse: $\text{InvPhi}(p_1)$

InvPhi ist die Inverse der Normalverteilung. Bei der Probitanalyse wird p als Fläche unter der Normalverteilungskurve verstanden. $\text{InvPhi}(p)$ ist dann der entsprechende z-Wert.

Der Zweck der Logit- bzw. Probit-Transformation ist es zu gewährleisten, dass durch die rechte Seite der Regressionsgleichung ein p-Wert prognostiziert wird, der auf den Bereich 0 bis 1 beschränkt ist.

10. Die nominalen Variablen Geschlecht und Beruf werden in (nicht-redundante) Dummies aufgelöst. Die Zeilen der 1. Häufigkeitstabelle werden also letztendlich in folgende Datenvektoren transformiert (Werte von KaufJa und KaufNein aus Logit-Analyse) :

A1	B1	B2	Einkommen	Konstante	Kauf Ja	Kauf Nein
1	1	0	1	1	-4.59512	0
1	1	0	2	1	-4.59512	0
KW	KW	KW	KW	KW	KW	KW
1	0	1	1	1	0	0
1	0	1	2	1	0	0
1	0	1	3	1	-4.59512	0
1	0	0	1	1	-0.91629	0
...

A1= 1.Dummy von Geschlecht (2.Dummy ist redundant)

B1, B2= 1. und 2. Dummy von Beruf (3.Dummy ist redundant)

- Wird in der Optionsbox "Weitere Optionen" **Option 42 = 1** gesetzt, dann werden diese Datenvektoren von Almo zum Anschauen ausgegeben. Sie können sich diese im Editor auch "herausschneiden" und mit Ihnen eine Regressionsanalyse rechnen, die dann dieselben Ergebnisse erbringt.
- Die Konstante wird nur benötigt, wenn Option 42=1. Wir kommen darauf noch zurück. Sie hat, da sie nur aus 1.0 besteht, bei Option42=0 keinen Einfluss.
- Die redundante Variable KaufNein wird von Almo noch "mitgeschleppt".
- Wenn, wie in der 1. Zeile (der 1. Häufigkeitstabelle) eine Häufigkeit =0 ist, dann ist das Logit oder Probit nicht berechenbar. Almo setzt in diesem Falle für die leere Zelle: 0.01, für die (einzige) besetzte Zelle: 0.99

- e. Wenn, wie in der 3. Zeile, beide Häufigkeiten =0 sind dann setzt Almo den gesamten Datenvektor auf Kein_Wert (KW). Er wird damit aus der Analyse ausgeschlossen.
- f. Der Wert der Variablen KaufJa der obigen 7. Zeilen ergibt sich aus
 $\ln(0.2857/0.7143) = -0.91629$ bei der Logit-Analyse
 $\text{InvPhi}(0.2857) = -0.56569$ bei der Probit-Analyse

Für die (redundante) Variable KaufNein entsteht

$$\ln(0.7143/0.7143) = 0 \quad \text{bei der Logit-Analyse}$$

$$\text{InvPhi}(0.7143) = 0.56569 \quad \text{bei der Probit-Analyse}$$

Bei gleichen Häufigkeiten für KaufJa und KaufNein, wie in Zeile 4 der -.
Häufigkeitstabelle entsteht ein Logit von

$$\ln(4/4) = \ln(1) = 0$$

- 11. Jeder transformierte Datenvektor wird in Programm 20 eingelesen und dem üblichen Kalkül des Allgemeinen linearen Modells unterworfen.

Als wesentliche **Ergebnisse** gibt Almo aus:

Koeffizienten fuer quantitat./ordinale Variable aus univariater Analyse
hinsichtlich der abhaeng. Var. V11 KaufJa

Variable	Regr. koef.	Standard fehler	95% Konfidenzbereich nach		erklärte Streuung	part. Korrel.
			oben	u.unten		
V5 Einkommen	1.5131	0.7871	1.7323	21.1350	0.502	

Variable	F-Wert	Signifikanz p	(1-p)100	df1	df2	Test-staerke
V5 Einkommen	3.696	0.082	91.81	1	11	0.4166

=====
Regressionskoeffizienten der Dummies
(=Kontraste bezueglich der letzten (auf 0 gesetzten) Dummy

		KaufJa
		V11
Geschlecht	männlich A1	-2.3477
Beruf	Arbeiter B1	-3.2731
Beruf	Angestel B2	1.1985

Koeffizienten fuer Konstante

hinsichtlich der abh.Variablen V11 KaufJa
Effekt (Regressionskoeffizient) -2.869083

hinsichtlich der abh.Variablen V12 KaufNein
Effekt (Regressionskoeffizient) 0.000000

Koeffizienten der Dummies
hinsichtlich der abh. Var. V11 KaufJa

Effekte von A Geschlecht

	Effekte	Standard- fehler	erklärte Streuung	partielle Korrelat.	t-Wert	Signifikanz p	(1-p)100	Test- staerke
A1	-1.3206	0.6946	20.6681	-0.4973	1.9012	0.0850	91.50%	0.4091
A2	1.0271	0.5402	20.6681	0.4973	1.9012	0.0850	91.50%	0.4091

Kontraste von A Geschlecht

	Kontraste	Standard- fehler	erklärte Streuung	t-Wert	Signifikanz p	(1-p)100	Test- staerke
A1 - A2	-2.3477	1.2348	20.6681	1.9012	0.0849	91.51%	0.4091

Effekte von B Beruf

	Effekte	Standard- fehler	erklärte Streuung	partielle Korrelat.	t-Wert	Signifikanz p	(1-p)100	Test- staerke
B1	-2.6997	0.8905	52.5583	-0.6747	3.0318	0.0114	98.86%	0.7861
B2	1.7719	0.7837	29.2270	0.5633	2.2608	0.0457	95.43%	0.5383
B3	0.5734	0.8905	2.3711	0.1906	0.6440	0.5311	46.89%	0.0901

Kontraste von B Beruf

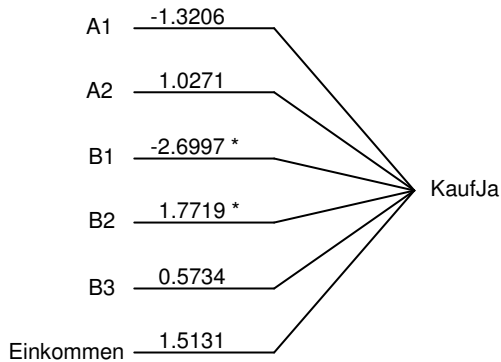
	Kontraste	Standard- fehler	erklärte Streuung	t-Wert	Signifikanz p	(1-p)100	Test- staerke
B1 - B2	-4.4716	1.4643	53.3204	3.0537	0.0109	98.91%	0.7919
B1 - B3	-3.2731	1.5124	26.7832	2.1642	0.0540	94.60%	0.5034
B2 - B3	1.1985	1.4643	3.8303	0.8184	0.4296	57.04%	0.1155

Zusammenfassung

Streuungsquelle	Streuung	F-Wert	df	Signifik. (1-p)100	Korrel Koeff.	Test- staerke
Gesamtstreuung	173.2835					
Fehlerstreuung	62.8989		11			
alle unabh. Var. zusammen	110.3846	4.8261	4	98.2850	0.7981	0.8241
quant./ordin. Var. zusammen	21.1350	3.6962	1	91.8106	0.5015	0.4188
nominale Variable zusammen	71.3022	4.1565	3	96.6409	0.7289	0.6968
V5 Einkommen	21.1350	3.6962	1	91.8106	0.5015	0.4166
V1 Geschlecht	20.6681	3.6145	1	91.5081	0.4973	0.4113
V3 Beruf	56.0102	4.8976	2	97.0385	0.6863	0.6809

Almo zeichnet noch folgendes Flussdiagramm der Regressionskoeffizienten und Effekte:

Effekte und Regressionskoeffizienten
 A Geschlecht: A1=männlich A2=weiblich
 B Beruf: B1=Arbeiter B2=Angestellter B3=Selbständiger



Ein Stern bedeutet: Mit 95% signifikant (p=0.05)

12. Die Ergebnisse können in obige Gleichung 1 eingesetzt werden:

$$p_1 = 1 / (1 + e^{*- (k_0 * 1 + a_1 * A_1 + b_1 * B_1 + b_2 * B_2 + c_1 * \text{Einkommen})})$$

$$p_1 = 1 / (1 + e^{*- (-2.8691 * 1 + \dots \text{Konstante} - 2.3477 * A_1 + \dots \text{Geschlecht} - 3.2731 * B_1 + 1.1985 * B_2 + \dots \text{Beruf} + 1.5131 * \text{Einkommen})}) \dots \text{Einkommen}$$

Für die 4. Zeile der Häufigkeitstabelle (mit A1=1 B1=0 B2=1 und Einkommen=2) erhalten wir p=0.7950. Der empirische Anteilswert ist p=0.5. Der prognostizierte Wert weicht also weit vom empirischen ab, was natürlich an unseren Testdaten liegt, die viel zu schwach besetzte Häufigkeiten besitzen.

13. Die Logit- nicht jedoch die Probit-Analyse kann auch gerechnet werden, wenn die abhängige Variable mehr als 2 Ausprägungen besitzt, z. B. Kaufabsicht: Opel, Mercedes, VW, Ford, Sonstige. Die letzten Ausprägung m wird als Bezugsgruppe verwendet. Die Logits werden in dann folgender Weise gebildet:

$$\ln(p_i/p_m)$$

Es wird also jeweils das Verhältnis des i-ten Anteilswerts p_i zum letzten Anteilswert p_m gebildet.

Für jede Ausprägung der abhängigen Variablen erhalten wir einen eigenen Satz von Koeffizienten. Die letzte Ausprägung ist wieder redundant. Almo liefert außerdem die Koeffizienten, die es bei der multivariaten Analyse ermittelt, also z.B. generalisierte Streuungen, Wilks Lambda etc. Diese Koeffizienten sind zu negieren. Sie werden bei der Logit- und Probit-Analyse üblicherweise nicht berechnet.

14. Bei der Logit- und Probit-Analyse tritt eine modellbedingte Varianzheterogenität auf.

Diese kann in der Optionsbox "witere Optionen" durch

Option 41 = 1;

beseitigt werden. Also rechnet dann eine "gewichtete Kleinste-Quadrate-Lösung". Es wird die Standardabweichung errechnet (siehe Maddala, 1983, S. 30,31):

$$s = \sqrt{\frac{1}{n * p_1 * p_2}} \quad \text{beim Logit-Modell}$$

$$s = \sqrt{\frac{p_1 * p_2}{n * \Phi(p_1)}} \quad \text{beim Probit-Modell}$$

n= Gesamt-Häufigkeit in der jeweiligen Zeile in der Häufigkeitstabelle

p₂= 1-p₁

Phi(p₁)= Ordinatenwert der Verteilungsfunktion der Standard-Normalverteilung

Für die 7. Zeile der obigen 1. Häufigkeitstabelle ergibt sich z.B. für das Logit-Modell:

$$s = \sqrt{\frac{1}{7 * 0.2857 * (1 - 0.2857)}} = 0.8367$$

für das Probit-Modell:

$$s = \sqrt{\frac{0.2857 * (1 - 0.2857)}{7 * \Phi(0.2857) * \Phi(0.2857)}} = 0.5023$$

wobei

$$\Phi(0.2857)=0.34$$

Diese Standardabweichung s wird für jeden Datenvektor (siehe oben Punkt 10) separat berechnet. Jede Variable des Datenvektors (auch die Konstante) wird dann mit diesem s dividiert.

Der Benutzer kann die Standardabweichung sehr einfach ermitteln. Werden mit Option 42=1 die Logit- bzw. Probittransformierten und gewichteten Datensätze ausgegeben, dann ergibt sich s als Kehrwert des für die Konstante ausgegebenen Wertes.

Die Konstante muss (an beliebiger Stelle) als unabhängige quantitative Variable eingeführt werden.

Bei kleinen Stichproben (wie bei unseren Testdaten) sollte eine gewichtete Analyse nicht gerechnet werden. Der multiple Korrelationskoeffizient sollte nicht interpretiert werden.

15. Kumulative Logits (**Option 44=1**)

Betrachten wir folgendes Beispiel:

Altersgruppe	Schwere des Verkehrsunfalls		
	leicht	mittel	schwer
1	61	28	7
2	68	23	13
3	58	40	12
4	53	38	16

Es soll ermittelt werden, ob zwischen Alter und Schwere eines Verkehrsunfalls ein Zusammenhang besteht. Das Alter ist in 4 Stufen eingeteilt. Die Häufigkeiten der Unfälle verschiedener Schwere sind in der Tabelle angegeben.

Die abhängigen Variablen lassen sich hier als Ausprägungen der ordinalen Variablen der Unfallschwere interpretieren: Die Unfallschwere nimmt zu von "leicht" über "mittel" nach "schwer".

Durch die Verwendung kumulativer Logits kann der ordinale Charakter der abhängigen Variablen berücksichtigt werden. Kumulative Logits werden gemäß der Formel gebildet:

$$L_i = \ln\left(\frac{1-S_i}{S_i}\right)$$

$$S_i = P_1 + P_2 + \dots + P_i$$

L_i = Logit der abhängigen Variablen i

S_i = Summe der Anteilswerte der 1. bis zur i -ten abhängigen Variablen

Dieselbe Wirkung erreicht man, wenn man mit "gewöhnlichen" Logits 2 Analysen rechnet.

1. Eine Analyse, bei der "mittel" und "schwer" zusammengefügt werden.
2. Und eine Analyse bei der "leicht" und "mittel" zusammengefügt werden.

Bei dieser Vorgehensweise kann man dann auch eine "gewichtete" Logitanalyse rechnen.

Literatur:

Aldrich/Nelson: Linear Probability, Logit and Probit Models,
Sage Publications 1984, S. 68 ff.

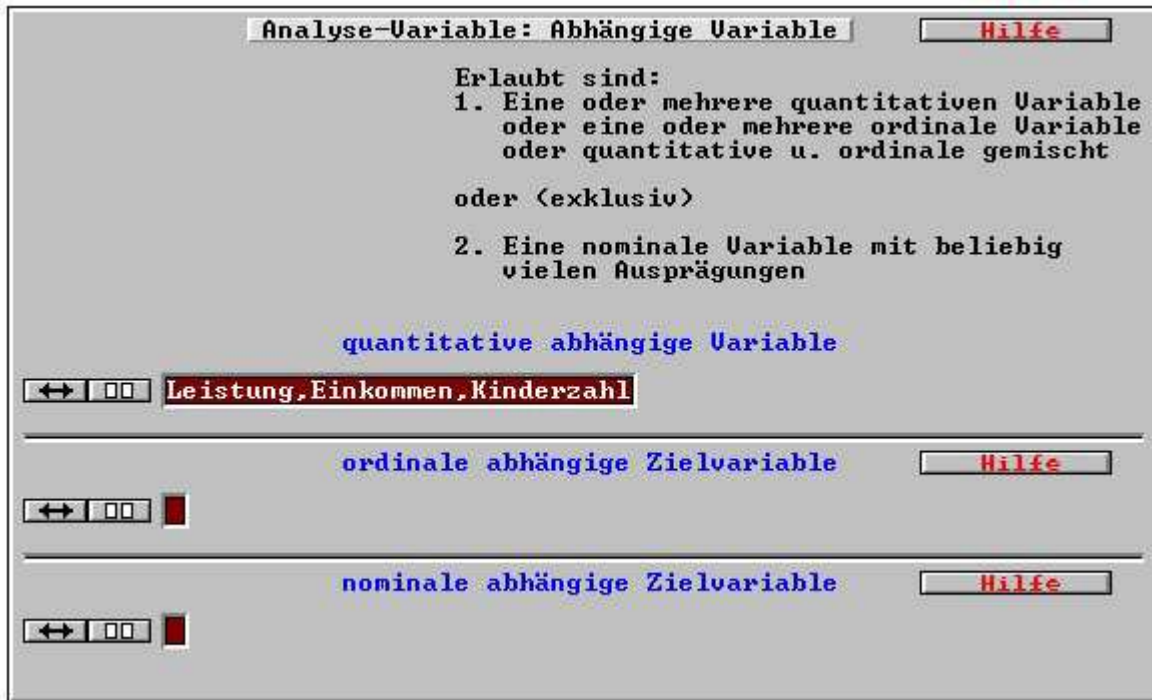
Hartung/Elpelt: Multivariate Statistik, 1984, S.128ff.

G.S. Maddala: Limited-dependent and qualitative variables in econometrics
Cambridge University Press, 1983

P20.9.4 Multivariate Analyse: Mehrere abhängige quantitative/ordinale Variable

Bei der multivariaten Analyse haben wir mehrere abhängige Variable. Dabei können diese (1) quantitativ bzw. ordinal oder (2) Dummies einer nominalen Variablen sein.

Im Maskenprogramm Prog20mx oder Prog20mo könnte die Box „Analyse-Variable: Abhängige Variable“ etwa folgendermaßen ausgefüllt sein.



Bei der multivariaten Analyse ist es empfehlenswert, als Streuungsmatrix KOVARIANZ oder KORRELATION zu verwenden. Bei QUADRATSUMME entstehen in der Regel generalisierte Streuungen mit sehr großen Zahlenwerten.

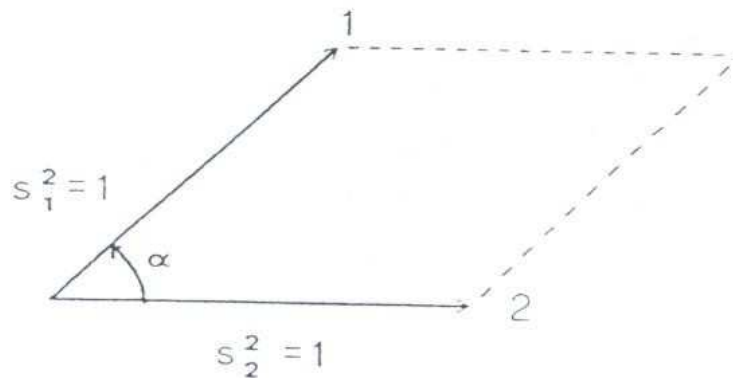
P20.9.4.1 Kalkül der multivariaten Analyse

Zum Begriff der "generalisierten Varianz"

Betrachten wir die Kovarianzmatrix von zwei standardisierten Variablen. Sie ist selbstverständlich gleich der Korrelationsmatrix, wobei $s_1^2 = s_2^2 = 1.0$

$$\begin{matrix} & \begin{matrix} 1 & 2 \end{matrix} \\ \begin{matrix} 1 \\ 2 \end{matrix} & \begin{bmatrix} s_1^2 & r_{12} \\ r_{21} & s_2^2 \end{bmatrix} \end{matrix}$$

Wir wissen, dass die Korrelation zwischen zwei Variablen geometrisch als 2 Vektoren, die durch den Winkel α (errechnet aus $\cos \alpha = r_{12}$) getrennt sind, dargestellt werden kann. Wir können nun die beiden Vektoren zu einem Parallelogramm erweitern.



Wilks schlug nun vor, die Fläche dieses Parallelogramms als Varianz der Menge der Variablen (1 2) zu begreifen. In unserem 2-Variablen-Beispiel ist die

$$\text{Fläche} = s_1^2 \cdot s_2^2 - r_{12} \cdot r_{21} = 1 - r_{12}^2$$

also gleich dem Varianzanteil, der nicht Kovarianz ist. Wilks nannte diese Art der Varianz "generalized variance". Die Fläche des obigen Parallelogramms ist identisch mit der Determinante obiger Matrix.

Dieser Gedanke lässt sich auf m Variable verallgemeinern: Die Varianz der Menge der Variablen (1 2 . . . m) ist gleich dem Rauminhalt des m -dimensionalen Parallelepipeds. Dieser ist gleich der Determinante aus der Kovarianzmatrix der m Variablen. Die Seiten des Parallelepipeds werden durch die Varianzen $s_1^2, s_2^2, \dots, s_i^2, s_j^2, \dots, s_m^2$ gebildet. Die Winkel zwischen den Seiten werden aus den Kovarianzen r_{ij} zwischen den entsprechenden Variablen gebildet. Dabei muss nicht notwendigerweise wie bei unserer seitherigen Darstellung, von standardisierten Variablen ausgegangen werden.

Als Streuungsmaße können dann auch Abweichungsquadratsummen verwendet werden. D.h. die Determinante kann auch aus der Quadratsummenmatrix berechnet werden.

Der Kalkül

Wir gehen von der Streuungsmatrix \mathbf{Q} aus, die in folgende Submatrizen unterteilt ist.

	a_1^*	.	.	x_m	y_1	.	.	y_w
a_1^*	\mathbf{Q}_{xx}				\mathbf{Q}_{xy}			
.								
x_m	\mathbf{Q}_{yx}				\mathbf{Q}_{yy}			
y_1								
.								
.								
y_w								

$a_1^* \dots x_m$ unabhängige Variable (Dummies und/oder Interaktionsdummies und/oder Kovariate)

m^* Zahl der unabhängigen Variablen. Sie ergibt sich aus m_A (Zahl der Dummies) plus m_x (der Zahl der Kovarianten).

$y_1 \dots y_w$ abhängige Variable (quantitative Variable oder 0-1 kodierte Nominaldummies einer nominalen Variablen)

w Zahl der abhängigen Variablen

Die Daten können in der Form von (1) Rohwerten, (2) Abweichungswerten, (3) standardisierten Werten vorliegen. Bei (2) ist \mathbf{Q} die Quadratsummen- (oder nach Division mit N , die Kovarianzmatrix). Bei (3) ist \mathbf{Q} (nach Division mit N) die Korrelationsmatrix. Bei (1) ist \mathbf{Q} die Kreuzproduktmatrix. In diesem Falle muss jedoch für die Konstantenvariable eine zusätzliche Spalte/Zeile hinzugefügt werden. Wir wollen im Folgenden, der Einfachheit halber, von Abweichungswerten (oder standardisierten Werten) ausgehen.

1) Zuerst bilden wir die generalisierte Gesamtstreuung der abhängigen Variablen $y_1 \dots y_w$. Diese ist gleich der Determinanten aus \mathbf{Q}_{yy} . Wir symbolisieren das durch $\det(\mathbf{Q}_{yy})$.

2) Dann werden die Regressionskoeffizienten gebildet.

$$(1) \mathbf{B} = \mathbf{Q}_{xx}^{-1} \cdot \mathbf{Q}_{xy}$$

$\mathbf{B} = m^* \cdot w$ -Matrix der Regressionskoeffizienten

Wir bilden also die w Regressionsgleichungen:

$$\begin{aligned} y'_1 &= \beta_{11} * a_1 + \dots + \beta_{1m} * x_m \\ y'_2 &= \beta_{21} * a_1 + \dots + \beta_{2m} * x_m \\ &\vdots \\ y'_w &= \beta_{w1} * a_1 + \dots + \beta_{wm} * x_m \end{aligned}$$

Die Koeffizienten jeder dieser Gleichungen hätten wir auch durch w gewohnte univariate Analysen ermitteln können.

- 3) Die durch die unabhängigen Variablen $a_1 \dots x_m$ in den abhängigen Variablen erklärte Streuung ist

$$(2) \mathbf{V}_{\text{mult}} = \mathbf{Q}_{yx} \cdot \mathbf{B}$$

\mathbf{V}_{mult} ist eine $w \cdot w$ -Matrix.

Im Diagonalglied ii von \mathbf{V}_{mult} sind die erklärten Streuungen hinsichtlich der abhängigen Variablen y_i - wie wir sie auch in einer gewohnten univariaten Analyse errechnen - enthalten. Im Außendiagonalglied ik ist die Streuung angegeben, die die unabhängigen Variablen von der gemeinsamen Streuung (Kovarianz) der abhängigen Variablen y_i und y_k erklären.

- 4) Die verbleibende Fehlerstreuung ist dann

$$(3) \mathbf{WM} = \mathbf{Q}_{yy} - \mathbf{V}_{\text{mult}}$$

\mathbf{WM} ist eine $w \cdot w$ -Matrix.

- 5) a) Es wird die generalisierte Fehlerstreuung berechnet. Sie ist $\det(\mathbf{WM})$

- b) Wir ermitteln die durch die Gesamtmenge der unabhängigen Variablen erklärte generalisierte Streuung. Sie ist

$$(4) SS_{\text{mult}} = \det(\mathbf{Q}_{yy}) - \det(\mathbf{WM})$$

Es gilt die Gleichung

$$(5) \det(\mathbf{Q}_{yy}) = SS_{\text{mult}} + \det(\mathbf{WM})$$

Bei der multivariaten Regressionsanalyse werden die Regressionskoeffizienten \mathbf{B} so gewählt, dass die Spur der Matrix \mathbf{WM} ein Minimum wird.

- 6) Es wird - hinsichtlich der Gesamtmenge der unabhängigen Variablen - das Wilks'sche Lambda ermittelt. Wir bezeichnen es mit Λ_{mult} .

$$(6) \Lambda_{\text{mult}} = \frac{\det(\mathbf{WM})}{\det(\mathbf{Q}_{yy})}$$

Die Bedeutung dieses Wilks'schen Lambda werden wir in Punkt 9, 10, 13 erläutern.

- 7) Wir wollen ermitteln, welchen Varianzanteil eine Untermenge der unabhängigen Variablen in den abhängigen Variablen erklärt. Diese Untermenge bestehe aus den unabhängigen Variablen $x_1 \dots x_k$. Diese Untermenge kann selbstverständlich auch nur aus einer unabhängigen Variablen bestehen. Die Untermenge kann im konkreten Falle auch aus den Dummies $a_1 \dots a_{p-1}$ der unabhängigen nominalen Variablen A (oder denen einer anderen nominalen Variablen oder denen einer Interaktion) bestehen.
- Es wird so verfahren, dass diese Untermenge $x_1 \dots x_k$ herausgenommen wird. Dann wird eine neue, zweite multivariate Analyse gerechnet. Wir erhalten dabei die neue Fehlermatrix $\mathbf{WM}_{-(x_1 \dots x_k)}$ (=Fehlermatrix einer Analyse ohne $x_1 \dots x_k$). Die einzelnen Glieder von $\mathbf{WM}_{-(x_1 \dots x_k)}$ sind größer als die entsprechenden Glieder von \mathbf{WM} - weil durch die Herausnahme von $x_1 \dots x_k$ weniger erklärt wird, also mehr Fehler verbleibt. Die durch die Untermenge $x_1 \dots x_k$ erklärte Streuung $\mathbf{V}_{(x_1 \dots x_k)}$ ist dann

$$(7) \mathbf{V}_{(x_1 \dots x_k)} = \mathbf{WM}_{-(x_1 \dots x_k)} - \mathbf{WM}$$

$\mathbf{V}_{(x_1 \dots x_k)}$ ist eine $w \cdot w$ -Matrix.

($x_1 \dots x_k$) Damit kennzeichnen wir eine beliebige Untermenge von unabhängigen Variablen. Sie kann aus den Dummies der nominalen Variablen, den Interaktionsdummies und den Kovarianten bestehen.

- 8) Wir berechnen nun die Reduktion der generalisierten Fehlerstreuung aus der 1. und der 2. Analyse (ohne $x_1 \dots x_k$) Wir wollen diese kurz "generalisierte Fehlerreduktion" oder die durch die Variablenmenge $x_1 \dots x_k$ "generalisierte erklärte Streuung" nennen. Sie ist

$$(8a) SS_{(x_1 \dots x_k)} = \det(\mathbf{WM}_{-(x_1 \dots x_k)}) - \det(\mathbf{WM})$$

Da gemäß Gleichung 7

$$(8b) \mathbf{WM}_{-(x_1 \dots x_k)} = \mathbf{WM} + \mathbf{V}_{(x_1 \dots x_k)}$$

ist, könnten wir auch die "generalisierte Fehlerreduktion" bzw. die "generalisierte erklärte Streuung" in folgender Weise ausdrücken:

$$(8c) SS_{(x_1 \dots x_k)} = \det(\mathbf{WM} + \mathbf{V}_{(x_1 \dots x_k)}) - \det(\mathbf{WM})$$

Bestünde die Untermenge aus den Hauptdummies $a_1 \dots a_{p-1}$ einer nominalen Variablen A, dann wäre nun $SS_{(x_1 \dots x_k)} = SS_{a_1 \dots a_{p-1}}$, und wir hätten die durch die nominale Variable A erklärte generalisierte Streuung gefunden.

Es gilt

$$\det(\mathbf{V}_{(x_1 \dots x_k)}) \neq \det(\mathbf{WM}_{-(x_1 \dots x_k)}) - \det(\mathbf{WM}).$$

Die Matrizen sind addierbar, nicht jedoch die Determinanten dieser Matrizen. Die "generalisierte erklärte Fehlerstreuung" darf also nicht als Determinante der Matrix $\mathbf{V}_{(x_1 \dots x_k)}$ der erklärten Streuungen errechnet werden. Es wäre deswegen im Grunde genommen besser, nicht von "generalisierter erklärter Streuung", sondern nur von der durch die Variablenmenge $x_1 \dots x_k$ verursachten "generalisierten Fehlerreduktion" zu sprechen. Wir wollen aber trotzdem den anderen Begriff - seiner Anschaulichkeit wegen - beibehalten. Die Größe $\det(\mathbf{V}_{(x_1 \dots x_k)})$ wird in der multivariaten Analyse nicht benötigt. Wir

müssen noch darauf hinweisen, dass die generalisierten erklärten Streuungen sich nicht zu SS_{mult} addieren, auch nicht, wenn die unabhängigen Variablen unkorreliert sind. Trotz dieses "Gebrechens" sind die "generalisierten erklärten Streuungen" verwendbar, wenn es darum geht, die Wirkungsstärke von unabhängigen Variablen (oder von Untermengen dieser) durch den PRE-Korrelationskoeffizienten (siehe Abschnitt P20.6.3 und die nachfolgenden Gleichungen) miteinander zu vergleichen.

- 9) Es wird - hinsichtlich der Untermenge $x_i \dots x_k$ - das Wilks'sche Lambda ermittelt. Es ist

$$(9) \Lambda_{(x_i \dots x_k)} = \frac{\det(\mathbf{WM})}{\det(\mathbf{WM}_{-(x_i \dots x_k)})}$$

Das Wilks'sche Λ gibt die Relation zweier generalisierter Fehlerstreuungen an. Es bewegt sich zwischen 0 und 1.

- 10) Das Wilks'sche Λ wird nun verwendet, um die "generalisierte erklärte Streuung" bzw. die PRE-Korrelationskoeffizienten auf ihre Signifikanz zu prüfen. Rao gelang es, Λ zu einem F-Wert zu transformieren (siehe dazu Bock, 1975, S.153).

Wir verwenden folgende Notation.

- w Zahl der abhängigen Variable
- r Zahl der unabhängigen Variablen der Untermenge $x_i \dots x_k$, deren Erklärungsfähigkeit getestet werden soll
- t ist $\min(w, r)$, d.h. t ist die kleinere Zahl von w und r
- m^* Gesamtzahl der unabhängigen Variablen $a_1 \dots a_m$. Sie ergibt sich aus m_A (der Zahl der Dummies) plus m_x (der Zahl der Kovarianten).
- N Probandenzahl

$$(10) F = \frac{1 - \Lambda^{1/s}}{\Lambda^{1/s}} \cdot \frac{df_2}{df_1}$$

Die Zahl der Freiheitsgrade ist

$$(11) df_1 = w \cdot r$$

$$(12) df_2 = f \cdot s - 1$$

Dabei ist

$$(13) s = \sqrt{\frac{w^2 \cdot r^2 - 4}{w^2 + r^2 - 5}}$$

$$(14) l = \frac{1}{2}(w \cdot r - 2)$$

$$(15) f = N - 1 - m^* - \frac{1}{2}(w - r + 1)$$

Ist $w \cdot r = 2$ dann wird $s = 1$ gesetzt

- 11) **Pillais Spur.** Neben Wilks Lambda kann auch noch "Pillais" Spur und die "Hotelling-Lawley" Spur zur Signifikanzprüfung verwendet werden. Pillais Spur ergibt sich in folgender Weise:

- a. Für alle unabhängigen Variablen

$$(16) P = \text{Spur}(\mathbf{V}_{\text{mult}} \cdot \mathbf{WM}^{-1})$$

P = Pillais Spur

\mathbf{V}_{mult} ist gemäß Gleichung 2 bestimmt

\mathbf{WM} ist gemäß Gleichung 3 bestimmt.

b. Für die Untermenge $x_i \dots x_k$ von unabhängigen Variablen

$$(17) P = \text{Spur}(\mathbf{V}_{(x_i \dots x_k)} \cdot \mathbf{WM}_{-(x_i \dots x_k)}^{-1})$$

P = Pillais Spur

$\mathbf{V}_{(x_i \dots x_k)}$ gemäß Gleichung 7

$\mathbf{WM}_{-(x_i \dots x_k)}$ gemäß Gleichung 8b.

Die Transformation zu einem (approximativen) F-Wert geschieht in folgender Weise)

$$(18) F = \frac{t \cdot P}{t - P} \cdot \frac{df_2}{df_1}$$

dabei ist

$$(19) t = \min(w, r)$$

t ist die kleinere Zahl von w bzw. r

$$(20) df_1 = l \cdot t$$

$$(21) df_2 = t \cdot (N - 1 - m^* - w + t)$$

dabei ist

$$(22) l = \max(w, r)$$

l ist die größere Zahl von w bzw. r .

Alle Größen, w , r , N , m^* sind in 11.) definiert.

12) Die Hotelling-Lawley Spur (gelegentlich auch nur "Hotellings Spur" genannt) ergibt sich gemäß

a. Für alle unabhängigen Variablen

$$(23) H = \text{Spur}(\mathbf{WM}^{-1} \cdot \mathbf{V}_{\text{mult}})$$

H = Hotelling-Lawley Spur

\mathbf{WM} gemäß Gleichung 3

\mathbf{V}_{mult} gemäß Gleichung 2

- b. Für die Untermenge $x_1 \dots x_k$ von unabhängigen Variablen

$$(24) H = \text{Spur} (\mathbf{W}\mathbf{M}_{-(x_1 \dots x_k)}^{-1} \cdot \mathbf{V}_{(x_1 \dots x_k)})$$

H = Hotelling-Lawley Spur
 $\mathbf{W}\mathbf{M}_{-(x_1 \dots x_k)}$ gemäß Gleichung 8b
 $\mathbf{V}_{(x_1 \dots x_k)}$ gemäß Gleichung 7

Der Unterschied zu Pillai's Spur besteht darin, dass bei der Hotelling-Lawley-Spur die Matrix \mathbf{V} der erklärten Streuungen postmultipliziert wird mit $\mathbf{W}\mathbf{M}_{-(x_1 \dots x_k)}^{-1}$, bei Pillai's Spur prämultipliziert.

Die Transformation zu einem (approximativen) F-Wert geschieht in folgender Weise

$$(25) F = \frac{H}{t^2} \cdot \frac{df_2}{df_1}$$

dabei ist
 t gemäß (19) bestimmt

$$(26) df_1 = 1 + t + 1$$

$$(27) df_2 = t \cdot (N - m^* - w - 2) + 2$$

dabei ist
(28) $l = \text{abs}(w - r) - 1$

Alle Größen w , t , N , m^* sind in Punkt 10 definiert.

- 13) Aus allen 3 Kriterien kann ein Korrelationskoeffizient gewonnen werden.
a. Aus Wilks Lambda wird folgender (quadrierter) Korrelationskoeffizient abgeleitet.

$$(29) r_w^2 = \frac{\det(\mathbf{W}\mathbf{M}_{-(x_1 \dots x_k)})^u - \det(\mathbf{W}\mathbf{M})^u}{\det(\mathbf{W}\mathbf{M}_{-(x_1 \dots x_k)})^u}$$

wobei $u = 1/t$ und t in (19) definiert ist
Gleichung (29) kann unter Verwendung von Gleichung 9 umgeformt werden zu

$$(30) r_w^2 = 1 - \Lambda_{(x_1 \dots x_k)}^u$$

$$(31) r_w = \sqrt{r_w^2}$$

r_w = Wilks Korrelation

- b) Auch aus Pillais Spur kann ein Korrelationskoeffizienten abgeleitet werden. Wir nennen ihn „Pillais Korrelation“.

$$(32) r_p = \sqrt{\frac{P}{t}}$$

r_p = Pillais Korrelation

t = Zahl der unabhängigen Variablen oder Zahl der abhängigen Variablen; die kleiner Zahl wird verwendet

P = Pillais Spur gemäß (16) oder (17)

Bei der Darstellung der kanonischen Korrelationsanalyse im Handbuch "Teil4: Fortgeschrittene Verfahren", Abschnitt P29.1.2 (bzw. im PDF-Dokument "Kanonische Analyse") werden wir zeigen, dass die quadrierte „Pillais Korrelation“ dem Durchschnitt aus den quadrierten kanonischen Korrelationen entspricht. Bestehen die unabhängigen und die abhängigen Variablen aus den Dummies je einer nominalpolytomen Variablen, dann entspricht „Pillais Korrelation“ dem Cramer’schen V (aus der Tabellenanalyse), siehe dazu den nachfolgenden Abschnitt P20.9.5.1.

c) Auch aus der Hotelling-Lawley-Spur kann ein Korrelationskoeffizient gewonnen werden.

$$(33) \quad r_H = \sqrt{\frac{H \cdot t}{H \cdot t + t^2}}$$

H ist gemäß (23) bzw. (24) und t gemäß (19) bestimmt

14. In Abschnitt P20.6.3 haben wir dargestellt, dass die Korrelationskoeffizienten, die im Rahmen des univariaten Allgemeinen Linearen Modells berechnet werden, PRE-Koeffizienten sind. „PRE“ heißt „proportional reduction of error“. Die Formel des PRE-Koeffizienten lautet:

$$(34) \quad \text{PRE}_{(x_1 \dots x_k)}^2 = \frac{\text{Fehler in Analyse ohne } x_1 \dots x_k - \text{Fehler in Analyse mit } x_1 \dots x_k}{\text{Fehler in Analyse ohne } x_1 \dots x_k}$$

$$= \frac{\text{durch } x_1 \dots x_k \text{ erklärte Streuung}}{\text{Fehler in Analyse ohne } x_1 \dots x_k}$$

Im Zähler des Quotienten steht die durch die Untermenge $x_1 \dots x_k$ verursachte Fehlerreduktion. Wir haben diese auch als die durch $x_1 \dots x_k$ „erklärte Streuung“ bezeichnet.

Wenn wir nun in der multivariaten Analyse Matrizen einsetzen, dann erhalten wir

$$(35) \quad \frac{\mathbf{WM}_{-(x_1 \dots x_k)} - \mathbf{WM}}{\mathbf{WM}_{-(x_1 \dots x_k)}} = \frac{\mathbf{V}_{(x_1 \dots x_k)}}{\mathbf{WM}_{-(x_1 \dots x_k)}}$$

Für Wilks Korrelation r_w (siehe Gleichung 29) werden die Determinante der auf der linken Gleichungsseite stehenden Matrizen verwendet (die noch mit $1/t$ potenziert werden).

Für Pillais und Hotellings Spur wird die Spur des Matrix-Quotienten auf der rechten Seite verwendet. Siehe Gleichung 17 und 24. Diese Spur wird dann noch mit t in besonderer Weise gewichtet.

Die drei Korrelationskoeffizienten sind also strukturgleich zum PRE-Korrelationskoeffizienten.

15. Vergleich von Wilks, Pillais und Hotellings Korrelation

Die drei Koeffizienten sind gleich, wenn $t = 1$ ist, das heißt, wenn auf Seiten der unabhängigen oder der abhängigen Variablen nur eine Variable vorhanden ist. Sie sind dann identisch mit der üblichen multiplen Korrelation.

Ist t größer 1, dann sind die 3 Koeffizienten etwas verschieden voneinander. In diesem Falle würden wir Pillais Korrelation vorziehen, da sie (quadriert) sich, wie oben in 13b oder in P20.9.5.1 dargestellt, als „durchschnittliche (quadrierte) kanonische Korrelation“ interpretieren lässt.

Im Almo-Prog20 wird demzufolge so verfahren:

Alle Korrelationskoeffizienten im Falle der multivariaten Analyse werden standardmäßig als „Pillais Korrelation“ errechnet. Wird eine entsprechende Option gesetzt, dann werden zusätzlich Wilks und Hotellings Korrelation ermittelt.

Pillais und Wilks Korrelation können auch im Rahmen der kanonischen Korrelationsanalyse entwickelt werden. Siehe dazu Handbuch Teil4, P29.

Pillais Korrelation erweist sich dort als die Wurzel aus dem Mittelwert aller quadrierten kanonischen Korrelationen.

Wilks Lambda wird im Rahmen der kanonischen Korrelationsanalyse wie wir sie in Prog29 rechnen für jeden einzelnen kanonischen Faktor berechnet. Wilks Lambda, wie wir es aus der multivariaten Analyse im Rahmen des Allgemeinen Linearen Modells erhalten, ist identisch mit Wilks Lambda für den 1. kanonischen Faktor, wie wir es im Rahmen der kanonischen Korrelationsanalyse erhalten. Die anderen kanonischen Faktoren werden nicht berücksichtigt. Das ist der Grund, warum Pillais Korrelation der Wilks'schen Korrelation vorzuziehen ist. Pillais Korrelation berücksichtigt alle kanonischen Faktoren.

Die Zahl der kanonischen Faktoren, die man bei einer kanonischen Korrelationsanalyse erhält, ist gleich t , d.h. der kleineren Zahl der unabhängigen bzw. der abhängigen Variablen.

Anmerkung: Befindet sich bei einer multivariaten Analyse mit Prog20 auf Seiten der unabhängigen und auf Seiten der abhängigen Variablen je eine (in Dummies aufgelöste) nominale polytome Variable, dann ist Pillais Korrelation identisch mit Cramers V , wie es im Rahmen einer 2-dimensionalen Tabellenanalyse mit den Makenprogrammen Prog10m1 oder Prog10m2 errechnet wird.

P20.9.4.2 Beispiel einer multivariaten Analyse

Wir wollen eine multivariate Analyse mit 2 abhängigen quantitativen Variablen, y_1 und y_2 und 2 unabhängigen nominalen Variablen A und B rechnen.

Dazu können wir problemlos das Maskenprogramm Prog20mo aus Abschnitt P20.8.1 verwenden.

Wir haben die Dialog-Boxen von Prog20mo entsprechend ausgefüllt und das Programm unter dem Namen "Multvar2.Alm" als Beispiel-Programm gespeichert. Sie finden Multvar2.Alm durch Klick auf das Menü "Almo / Liste aller Almo-Programme".

Die Boxen in Multvar2.Alm, die wir erläutern müssen, sind folgende:

Box "Freie Namensfelder"

Freie Namensfelder **Hilfe**

erzeuge zusätzliche Namensfelder

Box "Datei aus der gelesen wird"

Datei aus der gelesen wird **Hilfe**

bei Datei-Problemen

Format der Daten **Hilfe**

der Datensatz enthält diese Variablen
Bei Format DIREKT schreiben Sie: alle_U

Die Daten sind folgende:

	FaktA	FaktB	Y1	Y2
	1	1	23.5	7
	1	1	23.7	8
	1	2	28.7	9
	2	1	8.9	5
	2	2	5.6	1
	2	2	8.9	4
	3	1	10.3	4
	3	1	12.5	2
	3	2	13.6	5
	3	2	14.6	6

Box "Analyse-Variable: Abhängige Variable"

Analyse-Variable: Abhängige Variable **Hilfe**

Erlaubt sind:

- Eine oder mehrere quantitativen Variable oder eine oder mehrere ordinale Variable oder quantitative u. ordinale gemischt

oder (exklusiv)

- Eine nominale Variable mit beliebig vielen Ausprägungen

quantitative abhängige Variable

ordinale abhängige Zielvariable **Hilfe**

nominale abhängige Zielvariable **Hilfe**

Box "Analyse-Variable: Unabhängige Variable"

Analyse-Variable: Unabhängige Variable Hilfe

nominale unabhängige Variable Hilfe

FaktA, FaktB

2 Hilfe

Interaktionen x. Ordnung zwischen den nominalen unabhängigen Variablen bilden oder einige ausgewählte Interaktionen bilden
0 =keine Interaktionen bilden

FaktA, FaktB

paarweise Vergleiche (Kontraste) für die nominalen unabhängigen Variablen rechnen

quantitative unabhängige Variable Hilfe

ordinale unabhängige Variable Hilfe

Box "Option Streuungsmatrix"

Option: Streuungsmatrix

Die Optionsbox wird geöffnet:

Lösche wieder diese Box

Kovarianz

Streuungsmatrix

Folgende Streuungsmatrizen können analysiert werden: Hilfe

- = Korrelation
- = Quasi_Korrelation
- = Kovarianz
- = Quadratsumme
- = Kreuzprodukt
- = d_Kreuzprodukt

Im Verlauf des Kalküls werden Determinanten aus Submatrizen der Streuungsmatrix errechnet. Dabei können sehr hohe Zahlenwerte entstehen - mit der Folge, dass die Rechengenauigkeit reduziert wird und dass die generalisierten erklärten Streuungen extrem hohe Zahlenwerte annehmen.

Es ist deswegen sinnvoll, nicht die Quadratsummen-Matrix zu verwenden, obwohl das im Prinzip möglich wäre, sondern die Kovarianz-Matrix.

Auch ein "selbst geschriebenes Programm" in der Almo-Syntax ist als Beispielprogramm "Multvar.alm" vorhanden. Sie finden es durch Klick auf das Menü „Almo/Liste aller Almo-Programme“.

Almo liefert folgende Ergebnisse (die wir hier nicht vollständig ausgeben):

Kovarianzmatrix

	A1	A2	B1	A1B1	A2B1	Y1	Y2
A1	0.6900	0.3900	0.1000	0.1100	-0.0100	3.9930	1.2100
A2	0.3900	0.6900	-0.1000	0.0100	-0.1100	-1.2570	-0.1900
B1	0.1000	-0.1000	1.0000	-0.1000	-0.1000	0.7500	0.1000
A1 B1	0.1100	0.0100	-0.1000	0.6900	0.4100	0.8870	0.5900
A2 B1	-0.0100	-0.1100	-0.1000	0.4100	0.6900	1.4830	1.0100
Y1	3.9930	-1.2570	0.7500	0.8870	1.4830	52.8861	15.3370
Y2	1.2100	-0.1900	0.1000	0.5900	1.0100	15.3370	5.6900

Alle im Folgenden angegebenen Streuungen und erklärte Streuungen sind Varianzen

=====

Ergebnisse aus multivariater Analyse

generalisierte Gesamtstreuung 65.698340

=====

Koeffizienten fuer Gesamt-Modell

Durch alle unabh. Variable
erklärte generalisierte Streuung 65.181690
generalisierte Fehlerstreuung 0.516650
multipler Korrelat.koeff. 0.996060

Wilks Lambda 0.007864
F-Wert f. erklärte Streuung 6.165973
Freiheitsgrade Nenner = 10
 Zaehler= 6

Signifikanz: p 0.019690
Signifikanz: (1-p)*100 98.031001 %
Teststaerke von F 0.867270

Pillais Spur 1.480050
F-Wert f. erklärte Streuung 2.277219
Freiheitsgrade Nenner = 10
 Zaehler= 8

Signifikanz: p 0.128122
Signifikanz: (1-p)*100 87.187849 %
Teststaerke von F 0.518213

multiple Korrelation (aus Pillais Spur) 0.860247
quadriert 0.740025

=====

******* Erläuterung:** Wenn mehrere abhängige quantitative / ordinale Variable vorhanden sind oder wenn die abhängige Variable eine nominale ist (die Almo-intern in Dummies aufgelöst wurde) dann wird der Korrelationskoeffizient berechnet nach der Formel:

$$(1) \text{ Korrelation} = \text{Wurzel} (\text{Pillais Spur} / t)$$

t = Zahl der unabhängigen Variablen oder der abhängigen Variablen; die kleinere Zahl wird verwendet

Wir nennen diesen Koeffizienten kurz "**Pillais Korrelation**"

Wird im Maskenprogramm Prog20mo die Optionsbox "Spezielle Programm-Optionen" geöffnet und im 1. Eingabefeld "1" eingesetzt, dann können zusätzlich noch berechnet werden

Wilks Korrelation
Hotellings Korrelation

Die 3 Koeffizienten sind etwas verschieden. Sie sind gleich, wenn auf Seiten der unabhängigen Variablen nur eine Variable vorhanden ist. Sind mehrere vorhanden, dann ist Pillais Korrelation vorzuziehen, wie wir anschließend zeigen werden.

In Prog20 wird demzufolge standardmäßig so verfahren:

- a. Die multiple Korrelation für das Gesamtmodell wird als Pillais Korrelation berechnet.
- b. Die partielle multiple Korrelation für die Gruppe der unabhängigen quantitativen Variablen sowie für die Gruppe der unabhängigen nominalen Variablen wird als Pillais Korrelation berechnet.
- c. Die partielle multiple Korrelation für hierarchische oder gleichrangige Gruppen der unabhängigen quantitativen Variablen sowie für die hierarchischen Gruppen der unabhängigen nominalen Variablen wird als Pillais Korrelation berechnet. Zum Begriff der "hierarchischen" bzw. "gleichrangigen" Gruppen siehe die nachfolgenden Abschnitte P20.10, P20.11 und P20.12.
- d. Die Korrelation für eine unabhängige nominale Variable wird als Pillais Korrelation berechnet.
Ist die unabhängige nominale Variable dichotom, dann würden die 2 anderen Korrelationskoeffizienten den gleichen Wert ergeben. Ist sie polytom dann würden sich für die beiden anderen ein etwas anderer Wert ergeben.
- e. Die Korrelation für eine unabhängige quantitative / ordinale Variable wird als Pillais Korrelation berechnet. Sie könnte auch als Wilks oder Hotellings Korrelation berechnet werden. Die sich ergebenden Wert sind gleich.
- f. Genauso für die einzelnen Dummies einer unabhängigen nominalen Variablen oder Interaktionsvariablen.

Pillais und Wilks Korrelation können auch im Rahmen der kanonischen Korrelationsanalyse entwickelt werden. Siehe Handbuch, Teil 4, Abschnitt P29 oder das PDF-Dokument "Kanonische Analysen".

Pillais Korrelation (quadriert) ist der Mittelwert aus allen quadrierten kanonischen Korrelationen.

Wilks Lambda wird im Rahmen der kanonischen Korrelationsanalyse für jeden einzelnen kanonischen Faktor berechnet. Wilks Lambda, wie wir es aus der multivariaten Analyse im Rahmen des Allgemeinen Linearen Modells (Prog20) erhalten, ist identisch mit Wilks Lambda für den 1. kanonischen Faktor, wie wir es im Rahmen der kanonischen Korrelationsanalyse (Prog29) erhalten.

Die anderen kanonischen Faktoren werden nicht berücksichtigt. Das ist der Grund, warum Pillais Korrelation vorzuziehen ist. Pillais Korrelation berücksichtigt alle kanonischen Faktoren.

Anmerkung: Befindet sich bei einer multivariaten Analyse mit Prog20 auf Seiten der unabhängigen und auf Seiten der abhängigen Variablen je eine (in Dummies aufgelöste) nominale polytome Variable, dann ist Pillais Korrelation identisch mit **Cramers V**, wie es im Rahmen einer 2-dimensionalen Tabellenanalyse mit Prog10 errechnet wird. Siehe dazu Handbuch, Teil 3, Abschnitt P10.

Koeffizienten fuer Konstante

hinsichtlich der abh.Variablen V3 Y1
Effekt (Regressionskoeffizient) 15.658333

hinsichtlich der abh.Variablen V4 Y2
Effekt (Regressionskoeffizient) 5.416667

```
=====
Koeffizienten fuer Variable V1 FaktA
erklaerte generalisierte Streuung 34.737261
-----
Wilks Lambda 0.014655
F-Wert f. erklarte Streuung 10.890724
Freiheitsgrade Nenner = 4
Zaehler= 6
Signifikanz: p 0.007409
Signifikanz: (1-p)*100 99.259117 %
Teststaerke von F 0.957134
-----
Pillais Spur 1.109830
F-Wert f. erklarte Streuung 2.493521
Freiheitsgrade Nenner = 4
Zaehler= 8
Signifikanz: p 0.126311
Signifikanz: (1-p)*100 87.368884 %
Teststaerke von F 0.451820
-----
Korrelation (aus Pillais Spur) 0.744926
quadriert 0.554915
=====
```

```
=====
Koeffizienten fuer Variable V2 FaktB
erklaerte generalisierte Streuung 0.601819
-----
Wilks Lambda 0.461926
F-Wert f. erklarte Streuung 1.747272
Freiheitsgrade Nenner = 2
Zaehler= 3
Signifikanz: p 0.313891
Signifikanz: (1-p)*100 68.610907 %
Teststaerke von F 0.167795
-----
Pillais Spur 0.538074
F-Wert f. erklarte Streuung 1.747272
Freiheitsgrade Nenner = 2
Zaehler= 3
Signifikanz: p 0.313891
Signifikanz: (1-p)*100 68.610907 %
Teststaerke von F 0.167830
-----
Korrelation (aus Pillais Spur) 0.733535
quadriert 0.538074
=====
```

```
=====
Koeffizienten fuer Variable : Interaktion V1*V2
erklaerte generalisierte Streuung 1.852432
-----
Wilks Lambda 0.218080
```

```

F-Wert f. erklarte Streuung          1.712056
Freiheitsgrade Nenner = 4
          Zaehler= 6
Signifikanz: p                        0.264229
Signifikanz: (1-p)*100                73.577053 %
Teststaerke von F                    0.276028
-----
Pillais Spur                          0.996849
F-Wert f. erklarte Streuung          1.987437
Freiheitsgrade Nenner = 4
          Zaehler= 8
Signifikanz: p                        0.189256
Signifikanz: (1-p)*100                81.074375 %
Teststaerke von F                    0.366939
-----
Korrelation (aus Pillais Spur)        0.705992
quadriert                             0.498425
=====

```

"Multivariate Effekte"

Generalisierte Streuung in den abhaengigen Variablen
die erklart wird durch die Dummies von A FaktA

	part. Korrel	erkl.gen Streuung	Wilks Lambda	F-Wert	df1	df2	Signifikanz p	(1-p)100	Test- staerke
A1	0.9913	29.3806	0.0173	85.3013	2	3	0.002	99.78	1.0000
A2	0.9839	15.6669	0.0319	45.4860	2	3	0.005	99.46	0.9974
A3	0.9156	2.6803	0.1616	7.7818	2	3	0.065	93.50	0.5314

"Multivariate Effekte"

Generalisierte Streuung in den abhaengigen Variablen
die erklart wird durch die Dummies von B FaktB

	part. Korrel	erkl.gen Streuung	Wilks Lambda	F-Wert	df1	df2	Signifikanz p	(1-p)100	Test- staerke
B1	0.7335	0.6018	0.4619	1.7473	2	3	0.314	68.61	0.1678
B2	0.7335	0.6018	0.4619	1.7473	2	3	0.314	68.61	0.1678

"Multivariate Effekte"

Generalisierte Streuung in den abhaengigen Variablen
die erklart wird durch die Dummies von AB

	part. Korrel	erkl.gen Streuung	Wilks Lambda	F-Wert	df1	df2	Signifikanz p	(1-p)100	Test- staerke
A1 B1	0.7361	0.6111	0.4581	1.7743	2	3	0.310	69.00	0.1697
A1 B2	0.7361	0.6111	0.4581	1.7743	2	3	0.310	69.00	0.1697
A2 B1	0.8247	1.0987	0.3198	3.1900	2	3	0.181	81.88	0.2646
A2 B2	0.8247	1.0987	0.3198	3.1900	2	3	0.181	81.88	0.2646
A3 B1	0.6408	0.3600	0.5894	1.0451	2	3	0.454	54.64	0.1201
A3 B2	0.6408	0.3600	0.5894	1.0451	2	3	0.454	54.64	0.1201

Zusammenfassung
aus multivariater Analyse

Streuungsquelle	generalisierte Streuung	Wilks Lambda	Korrel Koeff.	F-Wert	df	Signifikanz p	(1-p)100
Gesamtstreuung	65.6983						
Fehlerstreuung	0.5166				6		
alle unabh. Var. zusammen	65.1817	0.0079	0.8602	6.1660	10	0.0197	98.0310
V1 FaktA	34.7373	0.0147	0.7449	10.8907	4	0.0074	99.2591
V2 FaktB	0.6018	0.4619	0.7335	1.7473	2	0.3139	68.6109
V1*V2	1.8524	0.2181	0.7060	1.7121	4	0.2642	73.5771

Multiple Korrelation aus univariater Analyse
hinsichtlich der abhaengigen Variablen V3 Y1

```
-----
Fehlerstreuung                                0.838500
Durch alle unabhaeng. Variablen erklarte Streuung 52.047600
Multiples Bestimmtheitsmass                   0.984145
Multiple Korrelation                           0.992041
F-Wert f. erklarte Streuung                   49.657818
Freiheitsgrade Nenner = 5
                Zaehler= 4
Signifikanz: p                                0.002621
Signifikanz: (1-p)*100                        99.737861 %
Teststaerke von F                             0.999992
```

Multiple Korrelation aus univariater Analyse
hinsichtlich der abhaengigen Variablen V4 Y2

```
-----
Fehlerstreuung                                0.750000
Durch alle unabhaeng. Variablen erklarte Streuung 4.940000
Multiples Bestimmtheitsmass                   0.868190
Multiple Korrelation                           0.931767
F-Wert f. erklarte Streuung                   5.269333
Freiheitsgrade Nenner = 5
                Zaehler= 4
Signifikanz: p                                0.067825
Signifikanz: (1-p)*100                        93.217469 %
Teststaerke von F                             0.566538
```

Hatten wir im Maskenprogramm Multvar2.Alm die Box "Spezielle Programm-Optionen" geoffnet und im 2. Eingabefeld "1" eingesetzt, dann wurde Almo noch berechnen:

1. zusatzlich noch die Hotelling-Lawleys Spur
2. die Matrix der in den abhangigen Variablen erklarten Streuungen

Schneiden wir aus der oben abgebildeten Kovarianzmatrix die Submatrix der Kovarianzen zwischen den abhangigen Variablen Y1 und Y2 heraus.

	Y1	Y2
Y1	52.8861	15.3370
Y2	15.3370	5.6900

Die in den abhaengigen Variablen durch die unabhangige nominale Variable FaktA erklarte Streuung ist

	Y1	Y2
Y1	47.9108	12.5900
Y1	12.5900	3.4000

Von der Gesamt-Varianz von Y1 (sie steht im Diagonalglied oben links und betragt 52.8861) wird 47.9108 durch Fakt A erklart. Von der Gesamt-Varianz von Y2 (= 5.69) wird 3.4 erklart. Von der Kovarianz von Y1 und Y2 (sie betragt 15.337) wird 12.59 durch Fakt A erklart.

```
Hotelling-Lawley Spur                          58.741293
Korrelation (aus Hotelling-Lawley Spur)         0.983399
quadriert                                       0.967073
F-Wert f. erklarte Streuung                   29.370647
```

```

Freiheitsgrade Nenner = 4
                  Zaehler= 4
Signifikanz: p      0.005162
Signifikanz: (1-p)*100 99.483755 %
Teststaerke von F    0.997228

```

P20.9.5 Multivariate Analyse: Eine nominale Variable als abhängige Variable

Als unabhängige Variable können gemischt quantitative, ordinale und nominale Variable eingesetzt werden. Wir wollen etwas ausführlicher zwei spezielle Fälle betrachten, den der Tabellenanalyse mit polytomer Variabler und den der multivariaten Diskriminanzanalyse.

P20.9.5.1 Tabellenanalyse mit polytomen Variablen

Wir wollen den Zusammenhang zwischen 2 polytomen nominalen Variablen betrachten. Die unabhängige nominale Variable sei V1; die abhängige nominale Variable sei V2. Die Häufigkeiten werden als V3 eingelesen und als Gewichtungvariable verwendet. Wir wollen dieses Mal das allgemeine lineare Modell auf die Kovarianzmatrix anwenden.

Wir verwenden das Maskenprogramm Prog20mo. Die Boxen dieses Programms haben wir entsprechend ausgefüllt und dann das Programm unter dem Namen "ZweiPol2.Alm" als Beispiel-Programm gespeichert. Sie finden **ZweiPol2.Alm** durch Klick auf das Menü "Almo / Liste aller Almo-Programme".

Auch ein selbst geschriebenes Syntaxprogramm ist unter dem Namen "ZweiPoly.Alm" als Beispielprogramm vorhanden.

Die Daten, die es zu analysieren gilt, sind als 2-dimensionale Tabelle dargestellt folgende:

		Variable Z		
		Z1	Z2	Z3
Variable A	A1	7	6	19
	A2	31	15	46
	A3	27	8	15

Für Prog20mo wird diese Tabelle in Form folgender Datenmatrix gebraucht:

A	Z	Häufigkeit
1	1	7
1	2	6
1	3	19
2	1	31
2	2	15
2	3	46
3	1	27
3	2	8
3	3	15

Die 3. Variable in der Datenmatrix enthält die Häufigkeit je Zelle der Tabelle.

Die Boxen in ZweiPol2.Alm, die wir erläutern müssen, sind folgende:

Box "Freie Namensfelder"

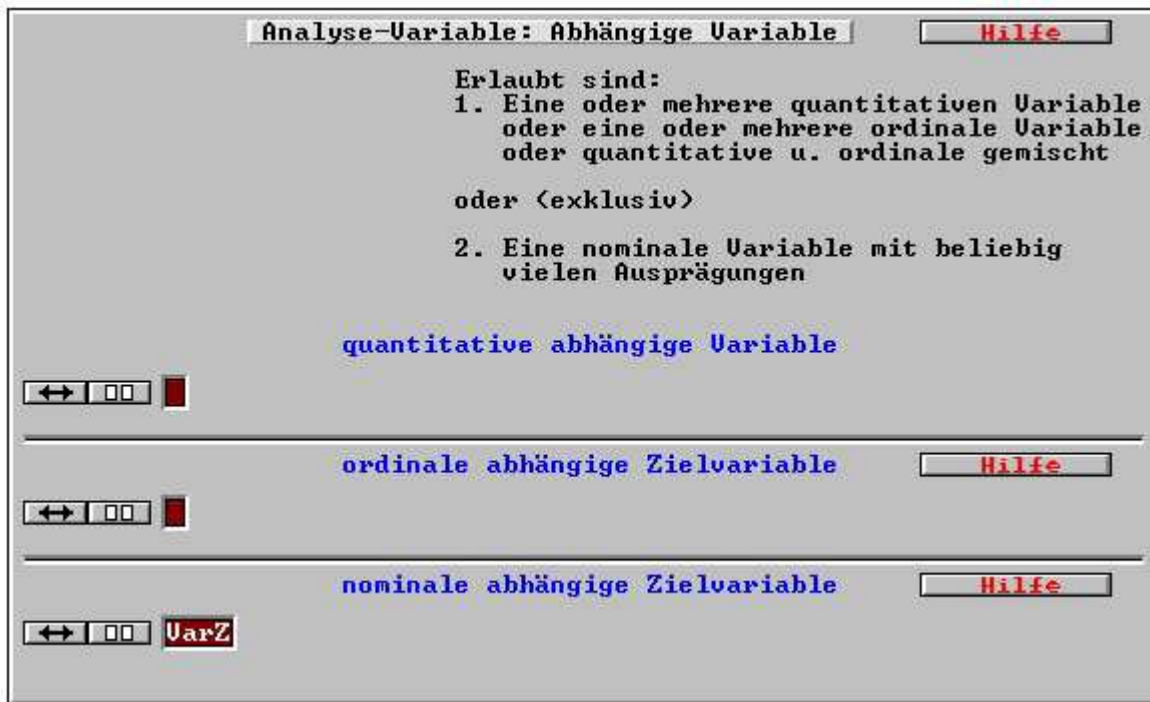


Die Variablen erhalten Namen. Die 3. Variable in der Datenmatrix enthält die Häufigkeit je Zelle der Tabelle. Wir nennen sie "Häufigkeit".

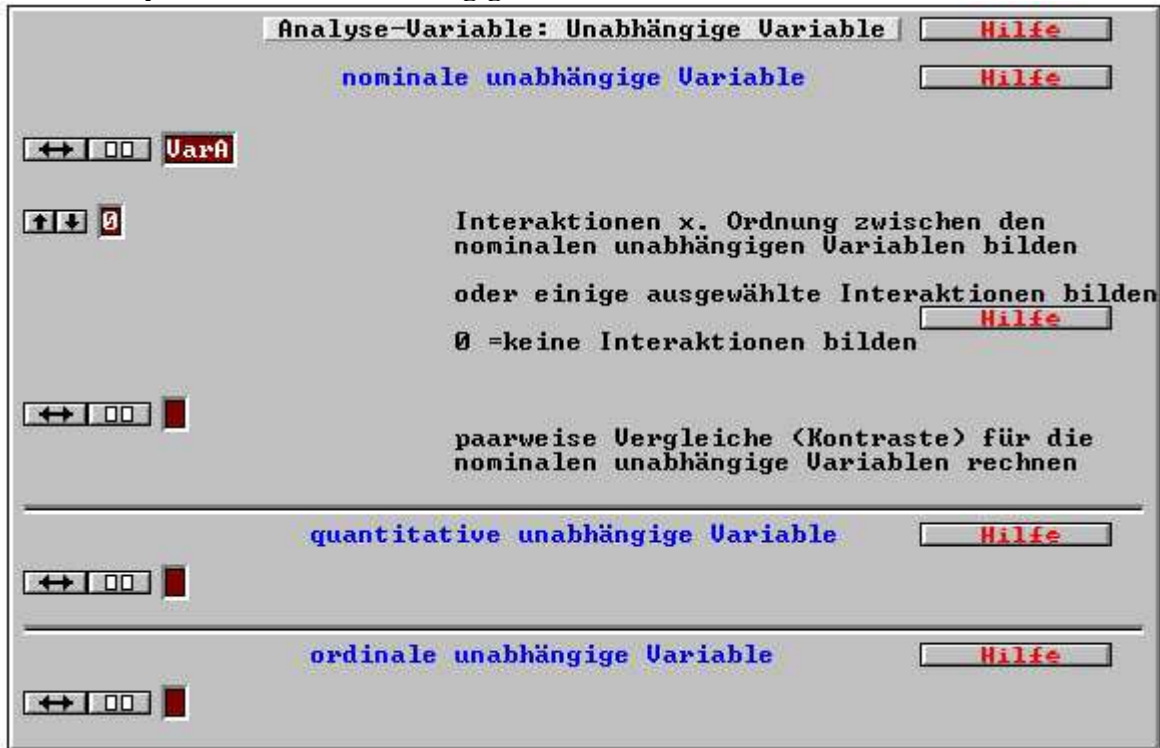
Box "Datei aus der gelesen wird"



Box "Analyse-Variable: Abhängige Variable"



Box "Analyse-Variable: Unabhängige Variable"



Box "Option: Untersuchungseinheiten gewichten"



Die Optionsbox wird geöffnet:

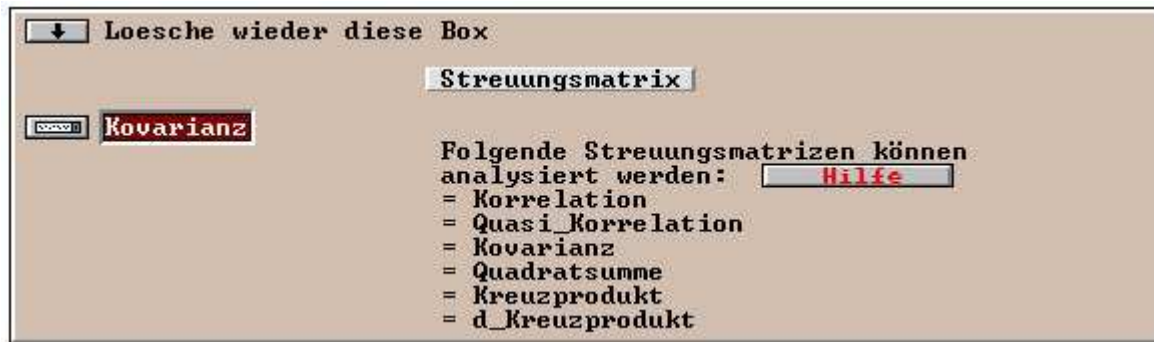


In (fast) allen Almo-Programmen wird die Almo-Variable "Gewicht1" für die Gewichtung der Datensätze verwendet. Die Variable "Häufigkeit" wird deswegen der Variablen "Gewicht1" zugewiesen. Beachte: Die Zuweisung muss mit einem Semikolon abgeschlossen werden.

Box "Option: Streungsmatrix"



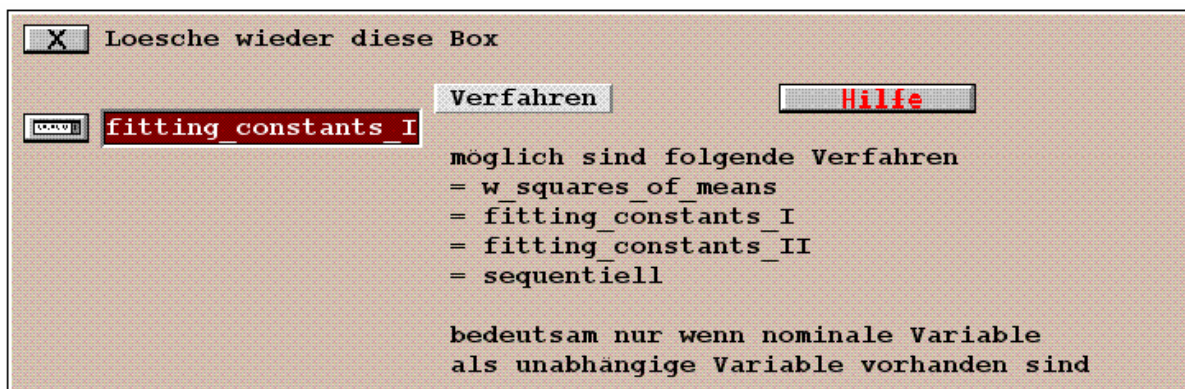
Die Optionsbox wird geöffnet:



Im Verlauf des Kalküls werden Determinanten aus Submatrizen der Streuungsmatrix errechnet. Dabei können sehr hohe Zahlenwerte entstehen - mit der Folge, dass die Rechengenauigkeit reduziert wird und dass die generalisierten erklärten Streuungen extrem hohe Zahlenwerte annehmen.

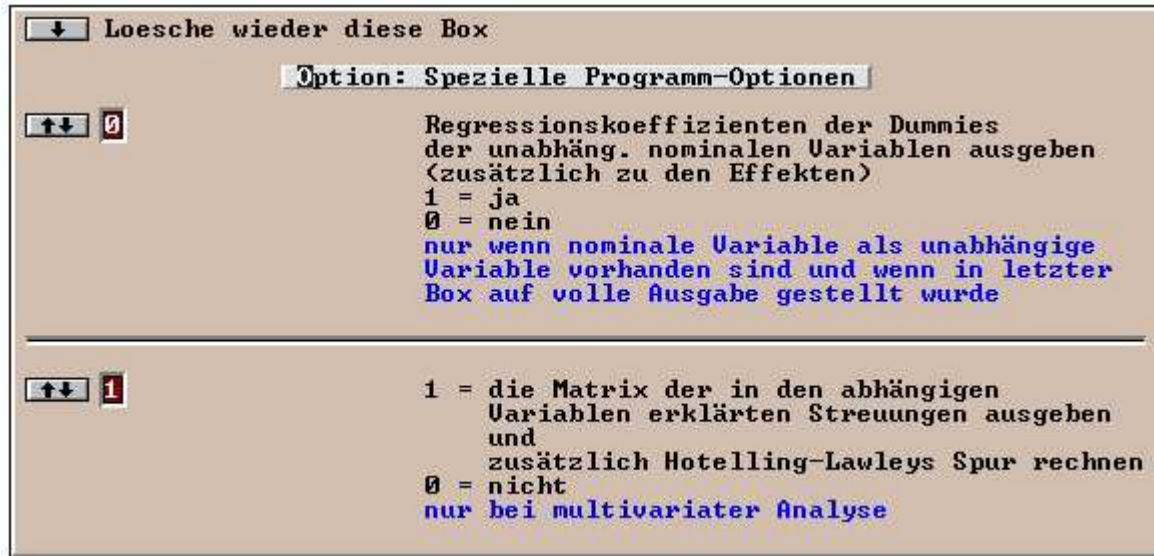
Es ist deswegen sinnvoll, nicht die Quadratsummen-Matrix zu verwenden, obwohl das im Prinzip möglich wäre, sondern die Kovarianz-Matrix.

Box "Verfahren"



Wir stellen das Verfahren auf "fitting constants I" ein. Das ist nicht notwendig. Es könnte problemlos auch mit dem voreingestellten Verfahren der "weighted squares of means" gerechnet werden. Die Ergebnisse wären - mit Ausnahme der Effekte - dieselben.

Box "Option: Spezielle Programm-Optionen"



Der Benutzer kann, wenn er will, sich noch die Hotelling-Lawley-Spur und die daraus gebildete Korrelation ausgeben lassen. Notwendig ist dies jedoch nicht.

Dasselbe Programm liegt auch als „selbst geschriebenes“ Almo-Syntaxprogramm vor. Man findet es im Menu Almo/Liste aller Almo-Programme.

Zuerst rechnen wir nun ein einfaches Tabellierungsprogramm mit denselben Daten. Man findet es als Beispielpogramm unter dem Namen "ZweiTab.Alm" durch offnen des Menus "Almo/Liste aller Almo-Programme". Das Ergebnis aus diesem Programm (stark gekurzt) ist folgendes:

```
Chi-Quadrat =          10.5440
N =                174
Signifikanz
(1-p)*100 =          96.8300
Cramers V =          0.1741
V2 =                0.0606
```

Bedeutsam fur die folgenden Erluterungen ist vor allem Cramers V.

Ausgabe aus Prog20mo.

Aus der Ausgabe, die ALMO fur oben abgebildetes Programm liefert, greifen wir die wichtigsten Ergebnisse heraus. Zur Interpretation der Ergebnisse siehe auch Abschnitt P20.6.8. Der besseren ubersichtlichkeit wegen bezeichnen wir V1 mit A und die Auspragungen (bzw. Dummies) mit A₁, A₂, A₃.

Die abhangige Variable V2 bezeichnen wir mit Z und ihre Auspragungen mit Z₁, Z₂, Z₃.

Hufigkeitstabelle		Tabelle der Anteilswerte (=Zellenmittelwerte der abhangigen Variablen)							
		<u>V2</u>							
		Z ₁	Z ₂	Z ₃					
<u>V1</u>	A ₁	7	6	19	32	0.22	0.19	0.59	1.0
	A ₂	31	15	46	92	0.34	0.16	0.50	1.0
	A ₃	27	8	15	50	0.54	0.16	0.30	1.0
			65	29	80				

Die Formel für Cramers V^2 ist

$$V^2 = \text{ChiQuadrat} / N$$

ALMO analysiert mit Prog20 folgende 3 Gleichungen:

$$(1) z'_i = \tau_i * t + \alpha_{1i} * a_1 + \alpha_{2i} * a_2 + \alpha_{3i} * a_3$$

für $i = 1, 2, 3$.

a_1, a_2, a_3 = dies sind die Dummies, die den Ausprägungen A1, A2, A3 entsprechen. Der Index i bezieht sich auf die 3 abhängigen Variablen z_1, z_2, z_3 . Wir erhalten für die 3 Gleichungen folgende Haupteffekte:

	z_1	z_2	z_3
τ	= 0.3740	0.1670	0.4600
α_1	= -0.1549	0.0209	0.1339
α_2	= -0.0366	-0.0036	0.0402
α_3	= 0.1665	-0.0067	-0.1597

Wir können nun sehr einfach die Anteilswerte (= Zellenmittelwerte) in obiger Tabelle reproduzieren. Betrachten wir beispielsweise die Zelle A_3Z_1 . Hier müssen wir den Nominaldummies in Gleichung 1 folgende Werte zuordnen:

$$a_1 = 0 \quad a_2 = 0 \quad a_3 = 1$$

Durch Einsetzen in die Regressionsgleichung erhalten wir den Anteilswert

$$\begin{aligned} p(a_3z_1) = z'_1 &= \tau_1 + \alpha_{31} * a_3 \\ &= 0.374 + 0.1665 * 1 \\ &= 0.54 \end{aligned}$$

Für Zelle A_2Z_2 erhalten wir

$$\begin{aligned} p(a_2z_2) = z'_2 &= \tau_2 + \alpha_{22} * a_2 \\ &= 0.167 - 0.0036 * 1 \\ &= 0.163 \end{aligned}$$

Aus der Kovarianzmatrix der Dummies entnehmen wir die Submatrix der 3 abhängigen Variablen z_1, z_2, z_3 .

	z_1	z_2	z_3
z_1	0.233	-0.062	-0.172
z_2	-0.062	0.139	-0.077
z_3	-0.172	-0.077	0.248

Die letzte Zeile/Spalte der Matrix wird gestrichen, um die in der Matrix enthaltene lineare Abhängigkeit zu beseitigen. Die Determinante der so verkleinerten Matrix liefert uns die **generalisierte Varianz**. Die Determinante einer 2x2-Matrix ergibt sich sehr einfach durch "Überkreuz"-Multiplizieren: $0,233 * 0,139 - (-0,062 * -0,062) = 0,02863$.

Im allgemeinen linearen Modell berechnen wir dann auch noch die Matrix der durch die unabhängigen Variablen erklärten Varianzen und Kovarianzen, sowie die verbleibende

Fehlermatrix. Die Determinante der Fehlermatrix ergibt dann die "generalisierte Fehlerstreuung". Die generalisierte erklärte Streuung erhalten wir residual.

ALMO liefert uns u.a. folgende Koeffizienten:

Alle im Folgenden angegebenen Streuungen und erklarte Streuungen sind Varianzen

generalisierte Gesamtstreuung	0.028626
=====	
Durch alle unabh. Variable	
erklarte generalisierte Streuung	0.001734
generalisierte Fehlerstreuung	0.026891

Wilks Lambda	0.939423
Korrelation (aus Wilks Lambda)	0.175389
quadriert	0.030761
Signifikanz: p	0.030238
Signifikanz: (1-p)*100	96.976235 %

Pillais Spur	0.060597
Signifikanz: p	0.031546
Signifikanz: (1-p)*100	96.845360 %

multiple Korrelation (aus Pillais Spur)	0.174065
quadriert	0.030299

Hotelling-Lawley Spur	0.064461
Korrelation (aus Hotelling-Lawley Spur)	0.176703
quadriert	0.031224
Signifikanz: p	0.029006
Signifikanz: (1-p)*100	97.099398 %

Wilks, Pillais und Hotellings Korrelation können interpretiert werden als Korrelationskoeffizient zwischen 2 polytomen Variablen. Da Korrelationskoeffizienten symmetrisch sind, muss derselbe Koeffizient entstehen, wenn wir umgekehrt V2 als unabhängige und V1 als abhängige Variable einsetzen. Dies ist auch tatsächlich der Fall.

Pillais Korrelation und Cramers V sind gleich. Es gilt:

$$\begin{aligned} \text{Pillais Korrelation} &= \sqrt{\text{PillaisSpur}/t} = \sqrt{0.0606/2} = 0.1741 \\ &= \text{Cramers V} = \sqrt{\text{ChiQuadrat}/N} = \sqrt{10.544/174} = 0.1741 \end{aligned}$$

Wobei t = die Zahl der Ausprägungen minus 1 (sind die Ausprägungszahlen verschieden, dann wird die kleinere genommen).

Siehe dazu Bortz, Lienert, Boehnke, 1990, S. 357 unten.

Der Unterschied zwischen dem über Wilks Lambda errechneten Korrelationskoeffizienten und dem über Pillais Spur berechnetem (bzw. Cramers V) wird ersichtlich, wenn wir mit Maskenprogramm Prog29m6 eine bivariate Korrespondenzanalyse (als kanonische Korrelation) rechnen. Almo liefert dann folgendes Ergebnis:

Faktor	Kanonische Korrelation R_K	Eigenwert (= Inertia) R^2_K	Wilks Lambda
1	0.24547	0.06025	0.93942
2	0.01854	0.00034	0.99966
Summe		0.06060	

Pillais Spur aus dem Allgemeinen Linearen Modell und Cramers V^2 aus der Tabellenanalyse sind identisch mit der Summe der Eigenwerte aller kanonischen Faktoren. Das t in obiger Gleichung für Pillais Korrelation lässt sich nun auch definieren als die Zahl der kanonischen Faktoren. Pillais Korrelation ist dann definiert als Wurzel aus dem Mittelwert aller quadrierten kanonischen Korrelationen.

Wilks Lambda, wie wir es aus dem Allgemeinen Linearen Modell erhalten, ist identisch mit Wilks Lambda aus dem 1. kanonischen Faktor. Wilks Korrelation berücksichtigt also nur den 1. kanonischen Faktor.

Ist eine der beiden nominalen Variablen dichotom, dann sind die über Wilks Lambda, über die Hotelling-Lawley-Spur und über Pillais Spur berechneten Korrelationen gleich. Es existiert dann nur ein kanonischer Faktor. Alle drei sind dann gleich zur kanonischen Korrelation und auch gleich zu Cramers V .

Wird die dichotome Variable als abhängige Variable (mit einer nicht-redundanten Dummy) und die polytome Variable (mit mehreren Dummies) als unabhängige Variable betrachtet, dann kann eine univariate Regressionsanalyse gerechnet werden, die eine multiple Korrelation erbringt, die identisch ist mit Pillais, Wilks und Hotellings Korrelation und Cramers V .

P20.9.5.2 Die Diskriminanzanalyse (lineares Wahrscheinlichkeitsmodell)

Bei der Diskriminanzanalyse befinden sich auf Seiten der unabhängigen Variablen quantitative Variable und auf Seiten der abhängigen Variablen eine nominale Variable. Siehe hierzu auch das Beispiel in P20.6.8.

Von "einfacher" Diskriminanzanalyse sprechen wir, wenn die abhängige nominale Variable eine Dichotomie ist, von "multivariater" Diskriminanzanalyse, wenn die abhängige nominale Variable polytomisch ist.

Für den Fall, dass die abhängige Variable nominal ist stehen 4 Verfahren zur Verfügung:

- (1) Die im Folgenden beschriebene multivariate Variante des allgemeinen linearen Modells,
- (2) das in P20.9.3.3 beschriebene nach dem kleinsten Quadrate-Kalkül gerechnete Logit- und Probit-Modell;
- (3) das im Handbuch Teil4: Fortgeschrittene Verfahren, Abschnitt P22 beschriebene nach dem Maximum-Likelihood-Kalkül gerechnete Logit- und Probit-Modell und
- (4) die "klassische" kanonische Diskriminanzanalyse, die wir im Handbuch, Teil4, Abschnitt P29.2 beschreiben.

Die Logit- und Probit-Modelle erbringen, da bei Ihnen die abhängige Variable nicht-linear transformiert wird, andere Ergebnisse als Verfahren 1 und 4.

Verfahren 1 und 4 liefern für den Fall, dass die abhängige Variable dichotom ist, dasselbe Ergebnis. Ist die abhängige Variable polytom, dann erbringen Verfahren 1 und 4 verschiedenen Ergebnisse, die jedoch in gewisser Weise äquivalent sind. Die klassische

kanonische Diskriminanzanalyse liefert $w-1$ orthogonale "Diskriminanzfunktionen" (w = Zahl der Ausprägungen der abhängigen nominalen Variablen). Das allgemeine lineare Modell hingegen liefert w nicht orthogonale Diskriminanzfunktionen. Siehe nachfolgende Gleichung (1). Der aus diesen Diskriminanzfunktionen ermittelbare Prognosewert für eine Untersuchungseinheit ist jedoch für beide Verfahren exakt gleich. Dies ist selbstverständlich kein Zufall. Der Beweis für die Äquivalenz von klassischer kanonischer Diskriminanzanalyse und der multivariaten Version des allgemeinen linearen Modells wird in Holm, 1979, S. 56 ff. ausgeführt.

Die "klassische" kanonische Diskriminanzanalyse wird in der Regel nur angewendet, wenn die unabhängigen Variablen quantitativ sind.

Betrachten wir ein Beispiel: Die abhängige Variable sei die Studienrichtung, die ein Student gewählt hat (mit den Ausprägungen Naturwissenschaft, Geisteswissenschaft, Wirtschaftswissenschaft). Als unabhängige Variable werden mehrere quantitativ gemessene psychologische Variable verwendet. Für die unabhängigen quantitativen Variablen werden hinsichtlich jeder Dummy-Variablen der abhängigen nominalen Variablen Regressionskoeffizienten ermittelt. Mit ihrer Hilfe kann dann die Wahrscheinlichkeit errechnet werden, mit der ein Student eine Studienrichtung wählt.

Wir wollen mit unseren Testdaten (".\Almo\Testdat\Testdat.fre") ein Beispiel rechnen.

Die quantitativen unabhängigen Variablen seien V6, V7, V8. Die abhängige Variable sei V3. Wir bezeichnen sie der besseren Übersichtlichkeit wegen mit C. Sie besitzt die 3 Ausprägungen C1, C2, C3.

Wir verwenden das in Abschnitt P20.8.1 ausführlich kommentierte Maskenprogramm Prog20mo.

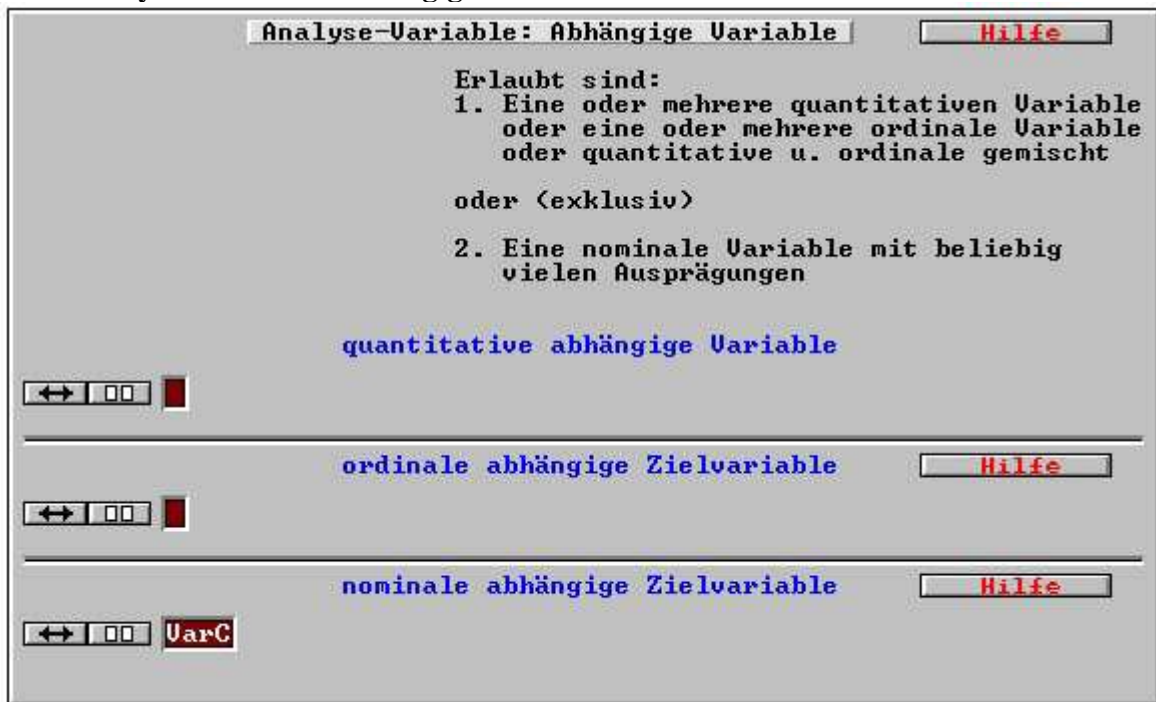
Die Box für die Variablennamen, die abhängige Variable und für die unabhängigen Variablen ist folgendermaßen auszufüllen.

Box "Freie Namensfelder"



Die Namen sind beliebig. Es wäre auch möglich, keine Namen zu geben. Also verwendet dann automatisch die Bezeichnung "Vx", wobei für x die Variablennummer eingesetzt wird.

Box "Analyse-Variable: Abhängige Variable"



Box "Analyse-Variable: Unabhängige Variable"

Analyse-Variable: Unabhängige Variable Hilfe

nominale unabhängige Variable Hilfe

Interaktionen x. Ordnung zwischen den nominalen unabhängigen Variablen bilden
 oder einige ausgewählte Interaktionen bilden Hilfe
 0 =keine Interaktionen bilden

paarweise Vergleiche (Kontraste) für die nominalen unabhängige Variablen rechnen

quantitative unabhängige Variable Hilfe

Quant1,Quant2,Quant3

ordinale unabhängige Variable Hilfe

Das allgemeine lineare Modell bestimmt folgendes Gleichungssystem:

$$c'_1 = {}_1\beta_0 + {}_1\beta_1V6 + {}_1\beta_2V7 + {}_1\beta_3V8$$

$$c'_2 = {}_2\beta_0 + {}_2\beta_1V6 + {}_2\beta_2V7 + {}_2\beta_3V8$$

$$c'_3 = {}_3\beta_0 + {}_3\beta_1V6 + {}_3\beta_2V7 + {}_3\beta_3V8$$

Der Index **vor** dem β -Koeffizienten bezieht sich auf die abhängige Variable, der **hinter** dem β -Koeffizienten auf die unabhängige Variable.

Wir erhalten folgende Regressionskoeffizienten (ohne β_0)

	C_1	C_2	C_3
β_1	-0.006	-0.035	0.040
β_2	-0.015	-0.032	-0.047
β_3	-0.003	-0.028	-0.030

Zu beachten ist wieder, dass sich die Regressionskoeffizienten einer Zeile zu .0 aufsummieren.

```

generalisierte Gesamtstreuung          121.704918
=====
Koeffizienten fuer Gesamt-Modell

Durch alle unabh. Variable
erklärte generalisierte Streuung        15.211663
generalisierte Fehlerstreuung          106.493255
-----
Wilks Lambda                            0.875012
F-Wert f. erklärte Streuung              1.288703
Freiheitsgrade Nenner =      6
                               Zaehler=  112
Signifikanz: p                          0.267382
Signifikanz: (1-p)*100                  73.261831 %
Teststaerke von F                        0.488179
    
```

```

-----
Pillais Spur                                0.125232
F-Wert f. erklarte Streuung                1.269177
Freiheitsgrade Nenner =      6
      Zaehler=    114
Signifikanz: p                             0.276352
Signifikanz: (1-p)*100                   72.364831 %
Teststaerke von F                         0.483604
-----
multiple Korrelation (aus Pillais Spur) 0.250232
quadriert                                0.062616

```

Die einzelnen Variablen erklären von der generalisierten Gesamtstreuung der nominalen Variablen C folgende Anteile - und es ergeben sich folgende partielle Korrelationskoeffizienten:

	part. Korrel	erkl.gen Streuung	Wilks Lambda	F-Wert	df1	df2	Signifikanz p	(1-p)100	Test- staerke
V6	0.2143	5.1261	0.9541	1.3478	2	56	0.267	73.27	0.2826
V7	0.2712	8.4515	0.9265	2.2221	2	56	0.116	88.41	0.4397
V8	0.1890	3.9435	0.9643	1.0369	2	56	0.363	63.74	0.2254

Sobald eine nominale Variable als abhängige Variable verwendet wird treten 2 Probleme auf:

1. Es besteht notwendigerweise Varianzheterogenität
2. Es treten Regressionskoeffizienten und Effekte auf, die Wahrscheinlichkeiten außerhalb des Bereichs von 0 bis 1.0 prognostizieren können.

Das 1. Problem haben wir ausführlich im Handbuch zum Almo-Data-Mining, Abschnitt P45.15.1.0 diskutiert.

Die Lösung für dieses Problem ist die "gewichtete Kleinste-Quadrate-Schätzung. Siehe dazu Abschnitt P20.9.3.2 und im Almo-Data-Mining, Abschnitt P45.15.1.8 und P45.15.2.2.

Mit dem 2. Problem wollen wir uns im folgenden Abschnitt beschäftigen.

P20.9.5.3 Prognostizierte Wahrscheinlichkeiten in der linearen Wahrscheinlichkeitsanalyse

Betrachten wir ein sehr einfaches Beispiel: Es soll untersucht werden, wie das Einkommen den Kauf einer Ware x bestimmt. Dieses Beispiel ist deswegen ein Einfaches, weil nur eine unabhängige Variable verwendet wird.

Die unabhängige Variable (das Einkommen) ist quantitativ. Wir verwenden dafür in unseren Testdaten ("C:\Almo\Testdat\Testdaten.fre") die Variable 5.

Die abhängige Variable (der Kauf) ist nominal. Kauf besitzt 2 Ausprägungen: ja und nein. Wir verwenden dafür in unseren Testdaten die Variable 10. Siehe auch Handbuch, Teil 4, P22 (Logitanalyse).

Wir rechnen mit Maskenprogramm Prog20mo.Msk ein allgemeines lineares Modell, genauer: eine Diskriminanzanalyse bzw. eine lineare Wahrscheinlichkeitsanalyse.

Das Programm ist unter dem Namen "Diskrim2.Alm" als Beispiel-Programm in Almo enthalten. Der Benutzer findet es durch Klick auf das Menü "Almo / Liste aller Almo-

Programme".

Die Box für die Variablennamen, die abhängige Variable und für die unabhängigen Variablen ist folgendermaßen auszufüllen.

Box "Freie Namensfelder"

Freie Namensfelder Hilfe

↔ Name 5 =Einkommen;
↔ Name 10=Kauf:Kauf Ja,KaufNein;

⋮ erzeuge zusätzliche Namensfelder

Box "Analyse-Variable: Abhängige Variable"

Analyse-Variable: Abhängige Variable Hilfe

Erlaubt sind:

1. Eine oder mehrere quantitativen Variable oder eine oder mehrere ordinale Variable oder quantitative u. ordinale gemischt oder <exklusiv>
2. Eine nominale Variable mit beliebig vielen Ausprägungen

quantitative abhängige Variable

↔

ordinale abhängige Zielvariable Hilfe

↔

nominale abhängige Zielvariable Hilfe

↔

Box "Analyse-Variable: Unabhängige Variable"

Analyse-Variable: Unabhängige Variable Hilfe

nominale unabhängige Variable Hilfe

Interaktionen x. Ordnung zwischen den nominalen unabhängigen Variablen bilden oder einige ausgewählte Interaktionen bilden Hilfe
 0 = keine Interaktionen bilden

paarweise Vergleiche (Kontraste) für die nominalen unabhängigen Variablen rechnen

quantitative unabhängige Variable Hilfe

Einkommen

ordinale unabhängige Variable Hilfe

Es wäre auch noch möglich - um das oben angesprochene Problem der Varianzheterogenität zu lösen - eine "gewichtete Kleinste-Quadrate-Schätzung" zu rechnen. Zu diesem Zweck müsste die Optionsbox "Option: Gewichtete Kleinste-Quadrate-Schätzung" geöffnet werden. Dadurch wird allerdings nicht das Problem gelöst, dass Wahrscheinlichkeiten prognostiziert werden können, die außerhalb 0-1 liegen.

Almo liefert folgende Ergebnisse (gekürzt):

```

Koeffizienten fuer quantitative Variable aus univariater Analyse
hinsichtlich der abhaengigen Variablen   V10=0 Kauf: ja

          95%
          Konfidenz-
          bereich
          nach
Variable      Regr. Standard  oben  erklarte  part. F-Wert  Signifikanz  df1  df2  Test-
          koef.  fehler  u.unten  Streuung  Korrel.      p      (1-p)100  stärke
-----
V5 Einkommen  0.0640  0.0330  0.0659   0.9092  0.245   3.767  0.057  94.31  1   59  0.4806

Koeffizienten fuer Konstante:   0.210491
    
```

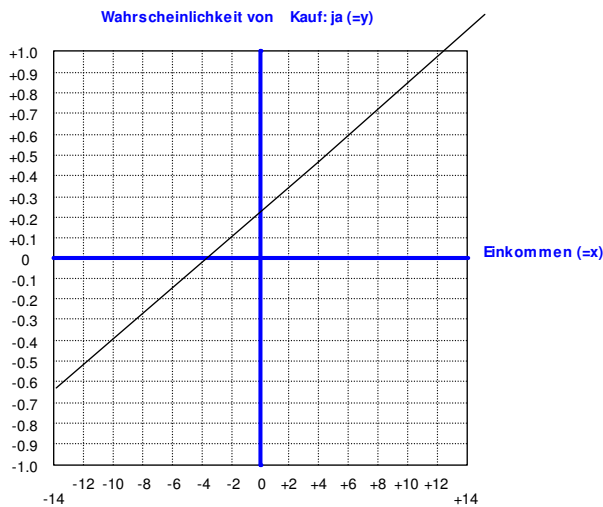
Die wesentlichen Ergebnisse sind also. Der Regressionskoeffizient beträgt 0.064. Er ist mit $(1-p)100 = 94.31\%$ signifikant. Die Konstante hat einen Wert von 0.2105

Wir können also die lineare Gleichung schreiben:

$$p = 0.064 * \text{Einkommen} + 0.2105$$

p = das ist die Wahrscheinlichkeit für "Kauf:ja"

Wir wollen die Gleichung als Gerade zeichnen:



Die Wahrscheinlichkeit eines Kaufs ist z.B. bei einem Einkommen

$$\begin{aligned} \text{von 4 Einheiten: } p &= 0.064 * 4 + 0.2105 = 0.4665 \\ \text{von 8 Einheiten: } p &= 0.064 * 8 + 0.2105 = 0.7225 \\ \text{von 10 Einheiten: } p &= 0.064 * 10 + 0.2105 = 0.8505 \end{aligned}$$

Der höchste Einkommenswert in unseren Daten ist 10.

Nun wollen wir wissen, wie die Kaufwahrscheinlichkeit bei einem Einkommen von 14 ist

$$14 \text{ Einheiten: } p = 0.064 * 14 + 0.2105 = 1.1950$$

Es entsteht eine Wahrscheinlichkeit größer als 1.0. Das gibt es nicht. Das ist die Schwäche der linearen Wahrscheinlichkeitsanalyse. Es können Wahrscheinlichkeiten prognostiziert werden, die über 1.0 oder unter 0 liegen.

Wenn der Zusammenhang zwischen Einkommen und Kaufwahrscheinlichkeit enger ist, die Gerade also steiler steigt, dann können auch bei einem Einkommen von 10 Einheiten Wahrscheinlichkeiten außerhalb 0-1 auftreten. Erfahrungsgemäß wird der zulässige Wahrscheinlichkeitsbereich 0-1 auch eher verlassen je mehr unabhängige Variable in das Modell eingeführt werden.

Mit Prog20mo aus Abschnitt P20.8.0 können die auftretenden Wahrscheinlichkeiten ermittelt werden. In der Box „Prognosewerte und Residuen“ dieses Programms muss das 1. Eingabefeld auf 1 gesetzt werden. Es werden dann die prognostizierten Wahrscheinlichkeiten errechnet und ausgegeben. Siehe dazu auch Abschnitt P20.9.3.1 (Prognosewerte und Residuen).

Für unsere Beispieldaten treten keine Wahrscheinlichkeiten außerhalb 0 - 1 auf. Wir zeigen deswegen die Ausgabe für eine andere Datenmatrix.

Datensatz	prognostizierte Wahrscheinlichkeit der Zugehörigkeit zu Gruppe	
	1	2
	KaufJa	KaufNein
1 (2)	0.168	0.832*
2 (2)	0.353	0.647*
3 (2)	0.123	0.877*
4 (2)	0.644*	0.356
5 (2)	0.311	0.689*
6 (2)	0.523*	0.477
7 (2)	0.377	0.623*
8 (1)	0.841*	0.159
9 (2)	0.652*	0.348
10 (2)	-0.058	1.058*
11 (2)	0.595*	0.405
12 (1)	0.759*	0.241
13 (2)	0.539*	0.461
14 (2)	-0.112	1.112*
15 (2)	0.102	0.898*
16 (2)	-0.013	1.013*
17 (2)	0.223	0.777*

Die Gruppe mit maximaler Wahrscheinlichkeit ist mit * markiert. Die tatsächliche Gruppenzugehörigkeit wird hinter der Datensatznummer in Klammern angegeben.

Wir erkennen, dass in den Datensätzen 10, 14, 16 Wahrscheinlichkeiten auftreten, die kleiner 0 bzw. größer 1 sind.

Streng genommen müssten wir es also aufgeben, die Prognosewerte als Wahrscheinlichkeiten zu interpretieren

Allerdings ist folgendes zur Verteidigung dieses Verfahrens zu sagen:

1. Wahrscheinlichkeiten größer 1.0 treten nur auf, wenn man den „Anwendungsbereich“ des Modells verlässt, d.h. wenn man für die unabhängige Variable Werte einsetzt, die sehr niedrig oder sehr hoch sind.
2. Man setzt Wahrscheinlichkeiten größer 1.0 einfach auf 1.0 und solche unter 0 auf 0 oder man normiert die Wahrscheinlichkeiten auf den Bereich 0-1. So wird in Almo verfahren.
3. Die Koeffizienten, die die lineare Wahrscheinlichkeitsanalyse liefert, sind einfach und klar zu interpretieren – so wie man es bei der Regressionsanalyse gewohnt ist. Dies gilt für die Koeffizienten der Logit- und Probitanalyse keinesfalls, wie wir noch im Handbuch, Teil 4, in Kapitel P22 sehen werden.
4. Der Prognoseerfolg ist beim linearen Wahrscheinlichkeitsmodell erfahrungsgemäß so gut wie beim Logit-Modell. Der Prognoseerfolg wird in folgender Weise festgestellt: Wir ermitteln für jeden Datensatz die wahrscheinlichste Gruppenzugehörigkeit. Das ist die, die oben mit * markiert wurde. Dann wird ausgezählt, wie oft die tatsächliche Gruppenzugehörigkeit prognostiziert wurde. Almo liefert folgenden Output:

Kauf:ja	28 Fälle.	Davon richtig prognostiziert	20	(=71,4%)
Kauf:nein	33		24	(=72,7%)

P20.9.5.4 Diskriminanzanalyse mit unabhängigen quantitativen und nominalen Variablen (lineares Wahrscheinlichkeitsmodell)

Im Rahmen des linearen Wahrscheinlichkeits-Modells ist es selbstverständlich auch möglich zu den unabhängigen quantitativen Variablen auch noch unabhängige nominale Variable hinzuzufügen.

Wir rechnen ein Beispiel mit unseren Testdaten. Siehe hierzu auch P20.6.8.

Es soll die Frage geklärt werden: Was sind die Bestimmungsgründe der Wahl eine Studienrichtung von jungen Menschen.

Es kann sowohl das Maskenprogramm Prog20mx (Abschnitt P20.8.0) als auch das Maskenprogramm mit Optionen Prog20mo (Abschnitt P20.8.1) verwendet werden.

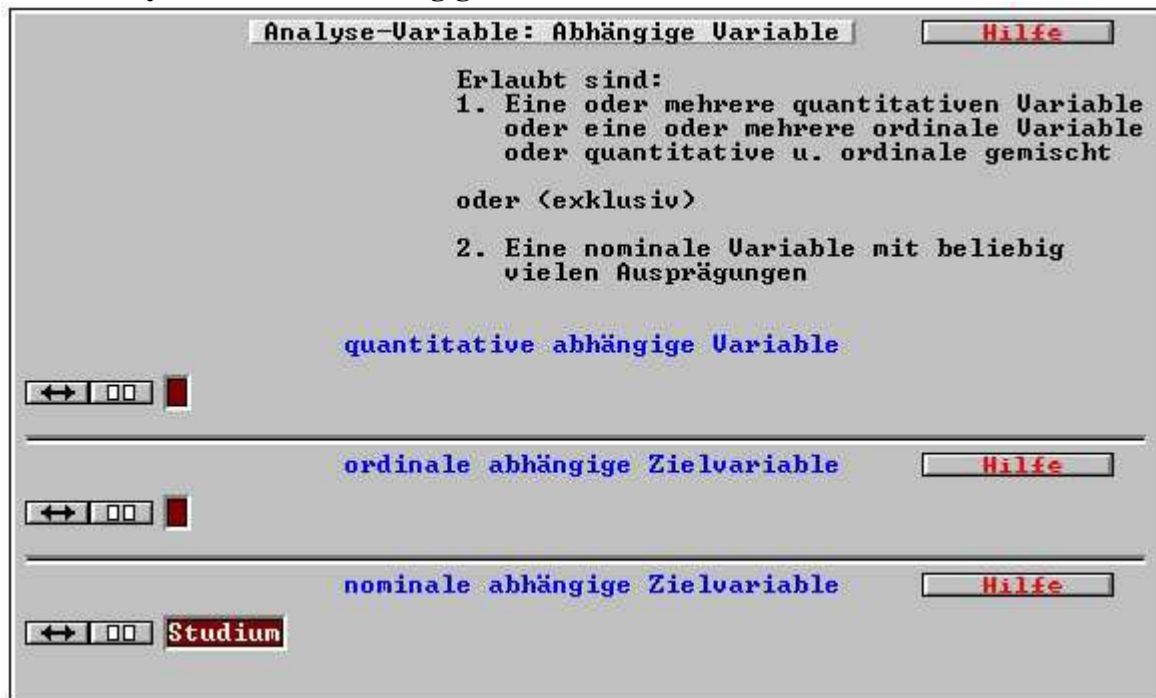
Wir zeigen nur die Boxen, die spezifisch für dies Analyse sind.

Box "Freie Namensfelder"



Als abhängige Variable verwenden wir V5, das den Namen "Studium" erhält und trichotomisiert wird. Wohnort und Geschlecht sind unabhängige nominale Variable und Alter und VaterEinkommen sind unabhängige quantitative Variable.

Box "Analyse-Variable: Abhängige Variable"



Box "Analyse-Variable: Unabhängige Variable"

Analyse-Variable: Unabhängige Variable Hilfe

nominale unabhängige Variable Hilfe

Wohnort, Geschlecht

2 Interaktionen x. Ordnung zwischen den nominalen unabhängigen Variablen bilden oder einige ausgewählte Interaktionen bilden Hilfe
 0= keine Interaktionen bilden

Wohnort, Geschlecht paarweise Vergleiche (Kontraste) für die nominalen unabhängigen Variablen rechnen

quantitative unabhängige Variable Hilfe

Alter, VaterEinkommen

ordinale unabhängige Variable Hilfe

Box "Option: Umkodierungen und Kein-Wert-Angaben"

Die Variable V5 muss trichotomisiert werden. Dazu wird die Optionsbox für das Umkodieren geöffnet:

Loesche wieder diese Box

Umkodierungen und Kein-Wert-Angaben

Umkodierungen Hilfe
 Kein_Wert-Angabe Hilfe

Studium(1:3=1; 4:5=2; 7:9=3)

erzeuge zusätzliche Felder für Umkodierungen / Kein_Wert-Angaben

Kontrollieren, ob Umkodierung so erfolgt wie gewünscht Hilfe

diese Variablen ...

... aus diesen Datensätzen vor und nach der Umkodierung zur Kontrolle anzeigen

Die Ergebnisse dieses Programms sind (gekürzt) folgende:

generalisierte Gesamtstreuung 91.803279
 =====

Koeffizienten fuer Gesamt-Modell

Durch alle unabh. Variable
 erklarte generalisierte Streuung 28.442333
 generalisierte Fehlerstreuung 63.360946
 multipler Korrelat.koeff. 0.556613

Wilks Lambda 0.690182
 F-Wert f. erklarte Streuung 1.513200
 Freiheitsgrade Nenner = 14
 Zaehler= 104

Signifikanz: p 0.118965
 Signifikanz: (1-p)*100 88.103506 %
 Teststaerke von F 0.817369
 =====

Koeffizienten fuer quantitat./ordinale Variable aus univariater Analyse

hinsichtlich der abhaeng. Var. V5-1 Studium: Geisteswissenschaft

	Regr. koeff.	Standard fehler	95% Konfidenz- bereich nach oben u.unten	erklaerte Streuung	part. Korrel.	F-Wert	Signifikanz p	(1-p)100	df2	Test- staerke
V6	0.0093	0.0294	0.0589	0.0234	-0.044	0.101	0.753	24.70	53	0.0612
V7	0.0117	0.0264	0.0530	0.0454	0.061	0.196	0.659	34.07	53	0.0719

hinsichtlich der abhaeng. Var. V5-2 Studium: Wirtschaftswissenschaft

V6	0.0061	0.0286	0.0574	0.0099	0.029	0.045	0.834	16.65	53	0.0550
V7	-0.0465	0.0257	0.0516	0.7167	-0.241	3.266	0.076	92.36	53	0.4272

hinsichtlich der abhaeng. Var. V5-3 Studium: Naturwissenschaft

V6	0.0033	0.0207	0.0415	0.0028	0.022	0.025	0.875	12.45	53	0.0527
V7	0.0348	0.0186	0.0373	0.4012	0.249	3.495	0.067	93.30	53	0.4513

general. Streuung in den abhaengigen Variablen,
 die durch die unabhaeng. quantitat. Variablen erklart wird

	part. Korrel	erkl.gen Streuung	Wilks Lambda	F-Wert	df1	df2	Signifikanz p	(1-p)100	Test- staerke
V6	0.0439	0.1221	0.9981	0.0501	2	52	0.941	5.91	0.0573
V7	0.2966	6.1121	0.9120	2.5081	2	52	0.089	91.06	0.4861

Koeffizienten fuer Konstante

hinsichtlich der abh.Variablen V5-1 Studium Geistesw
 Effekt (Regressionskoeffizient) 0.434324

hinsichtlich der abh.Variablen V5-2 Studium Wirtschaftsw
 Effekt (Regressionskoeffizient) 0.597527

hinsichtlich der abh.Variablen V5-3 Studium Naturwiss
 Effekt (Regressionskoeffizient) -0.031851

```

=====
Koeffizienten fuer Variable   V3 Wohnort

Korrelat.koeff.                0.268281
erklarte generalisierte Streuung 4.914074

Wilks Lambda                    0.928025
F-Wert f. erklarte Streuung    0.989413
Freiheitsgrade Nenner =      4
                          Zaehler= 104
Signifikanz: p                  0.417846
Signifikanz: (1-p)*100        58.215395 %
Teststaerke von F              0.303473
=====

```

```

Koeffizienten fuer Variable   V4 Geschlecht

Korrelat.koeff.                0.363416
erklarte generalisierte Streuung 9.641514

Wilks Lambda                    0.867929
F-Wert f. erklarte Streuung    3.956370
Freiheitsgrade Nenner =      2
                          Zaehler= 52
Signifikanz: p                  0.024481
Signifikanz: (1-p)*100        97.551904 %
Teststaerke von F              0.685578
=====

```

```

Koeffizienten fuer Variable :   Interaktion V3*V4

Korrelat.koeff.                0.275601
erklarte generalisierte Streuung 5.208236

Wilks Lambda                    0.924044
F-Wert f. erklarte Streuung    1.047492
Freiheitsgrade Nenner =      4
                          Zaehler= 104
Signifikanz: p                  0.387032
Signifikanz: (1-p)*100        61.296783 %
Teststaerke von F              0.320440
=====

```

Koeffizienten der Dummies
hinsichtlich der abh. Var. V5-1 Studium: Geisteswissenschaft

Effekte von A Wohnort

	Effekte	Standard- fehler	erklarte Streuung	partielle Korrelat.	t-Wert	Signifikanz p	Signifikanz (1-p)100	Test- staerke
A1	-0.1252	0.0968	0.3878	-0.1750	1.2937	0.2015	79.85%	0.2460
A2	-0.0539	0.0886	0.0858	-0.0833	0.6086	0.5449	45.51%	0.0918
A3	0.1791	0.0996	0.7498	0.2399	1.7989	0.0778	92.22%	0.4239

Effekte von B Geschlecht

	Effekte	Standard- fehler	erklarte Streuung	partielle Korrelat.	t-Wert	Signifikanz p	Signifikanz (1-p)100	Test- staerke
B1	-0.1842	0.0671	1.7480	-0.3530	2.7467	0.0082	99.18%	0.7699
B2	0.1842	0.0671	1.7480	0.3530	2.7467	0.0082	99.18%	0.7699

Effekte von AB

	Effekte	Standard- fehler	erklarte Streuung	partielle Korrelat.	t-Wert	Signifikanz p	(1-p)100	Test- staerke
A1 B1	0.0095	0.0963	0.0023	0.0136	0.0992	0.9197	8.03%	0.0511
A1 B2	-0.0095	0.0963	0.0023	-0.0136	0.0992	0.9197	8.03%	0.0511
A2 B1	0.0457	0.0891	0.0610	0.0703	0.5130	0.6097	39.03%	0.0795
A2 B2	-0.0457	0.0891	0.0610	-0.0703	0.5130	0.6097	39.03%	0.0795
A3 B1	-0.0553	0.0968	0.0756	-0.0782	0.5713	0.5697	43.03%	0.0867
A3 B2	0.0553	0.0968	0.0756	0.0782	0.5713	0.5697	43.03%	0.0867

***** entsprechend werden auch die Effekte hinsichtlich
der Wirtschafts- und der Naturwissenschaft ausgegeben

"Multivariate Effekte"

Generalisierte Streuung in den abhaengigen Variablen
die erklart wird durch die Dummies von A Wohnort

	part. Korrel	erkl.gen Streuung	Wilks Lambda	F-Wert	df1	df2	Signifikanz p	(1-p)100	Test- staerke
A1	0.2099	2.9207	0.9559	1.1985	2	52	0.310	69.01	0.2541
A2	0.0992	0.6303	0.9901	0.2587	2	52	0.776	22.43	0.0894
A3	0.2484	4.1676	0.9383	1.7102	2	52	0.189	81.08	0.3477

"Multivariate Effekte"

Generalisierte Streuung in den abhaengigen Variablen
die erklart wird durch die Dummies von B Geschlecht

	part. Korrel	erkl.gen Streuung	Wilks Lambda	F-Wert	df1	df2	Signifikanz p	(1-p)100	Test- staerke
B1	0.3634	9.6415	0.8679	3.9564	2	52	0.025	97.55	0.6903
B2	0.3634	9.6415	0.8679	3.9564	2	52	0.025	97.55	0.6903

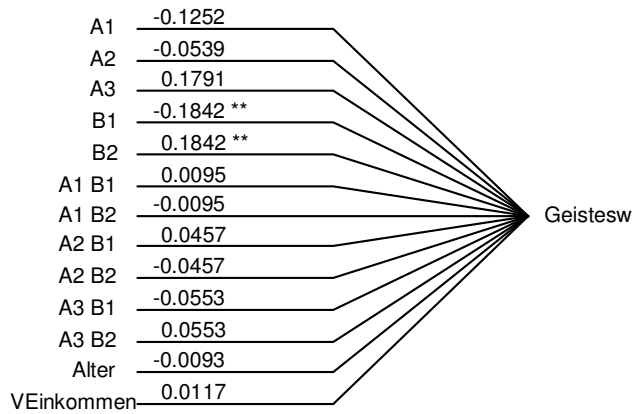
"Multivariate Effekte"

Generalisierte Streuung in den abhaengigen Variablen
die erklart wird durch die Dummies von AB

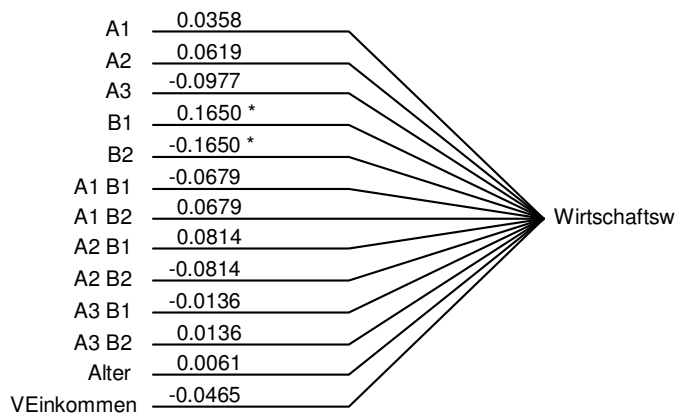
	part. Korrel	erkl.gen Streuung	Wilks Lambda	F-Wert	df1	df2	Signifikanz p	(1-p)100	Test- staerke
A1 B1	0.1336	1.1511	0.9822	0.4723	2	52	0.632	36.81	0.1246
A1 B2	0.1336	1.1511	0.9822	0.4723	2	52	0.632	36.81	0.1246
A2 B1	0.2709	5.0187	0.9266	2.0594	2	52	0.136	86.41	0.4099
A2 B2	0.2709	5.0187	0.9266	2.0594	2	52	0.136	86.41	0.4099
A3 B1	0.1400	1.2659	0.9804	0.5194	2	52	0.603	39.66	0.1326
A3 B2	0.1400	1.2659	0.9804	0.5194	2	52	0.603	39.66	0.1326

Almo zeichnet noch folgende 3 Flussdiagramme der Effekte und Regressionskoeffizienten hinsichtlich der 3 Studienrichtungen.

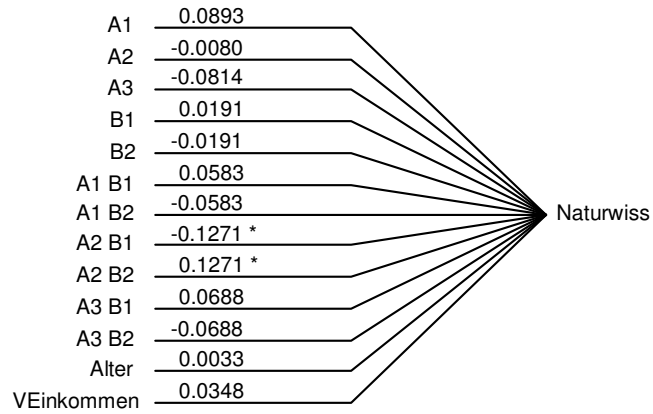
Effekte und Regressionskoeffizienten
 A Wohnort: A1=Stadt A2=Kleinstadt A3=Dorf
 B Geschlecht: B1=männlich B2=weiblich



Effekte und Regressionskoeffizienten
 A Wohnort: A1=Stadt A2=Kleinstadt A3=Dorf
 B Geschlecht: B1=männlich B2=weiblich



Effekte und Regressionskoeffizienten
A Wohnort: A1=Stadt A2=Kleinstadt A3=Dorf
B Geschlecht: B1=männlich B2=weiblich



P20.9.6 RESERVIERT

P20.9.7 Allgemeines Lineares Modell mit Rangvariablen

In Almo sind die beiden Maskenprogramme Prog20m8 und Prog20am enthalten, die es erlauben, "Rangvariable" in die Analyse mit einzuschließen. Der Benutzer findet diese Programme durch Klick auf "Verfahren / Allgemeines Lineares Modell" oder durch Klick auf das Menü "Almo / Liste aller Almo-Programme".

Das Maskenprogramm Prog20m8 rechnet ein Allgemeines Lineares Modell mit einer Rangvariablen als abhängiger Variablen. Dabei wird diese Rangvariable behandelt wie eine quantitative Variable.

Wird für eine unabhängige nominal-dichotome Variable und eine abhängige Rangvariable ein Prog20m8 gerechnet, dann entspricht dies dem U-Test nach Mann-Whitney. Die Ergebnisse stimmen überein, sind aber nicht exakt gleich. Siehe dazu Handbuch, Teil 3, Abschnitt P8.2.8.

Das zweite Maskenprogramm Prog20am ist als "Experimentalprogramm" für den Statistiker gedacht. Es rechnet wie Prog20m8 ein Allgemeines Lineares Modell mit einer Rangvariablen als abhängiger Variablen. Zusätzlich ist es aber nun auch möglich, Rangvariable als unabhängige Variable einzuführen. Die Rangvariablen werden dabei wie quantitative Variable behandelt.

Beide Programme, Prog20m8 und Prog20am stimmen in ihren Boxen und Eingabefeldern weitgehend überein mit dem in Abschnitt P20.8.1 dargestellten und erläuterten Programm Prog20mo - so dass wir es uns ersparen können, die einzelnen Boxen dieser beiden Programme zu kommentieren.

P20.9.7.1 Zum Begriff der "Rangvariablen"

Die Werte einer Rangvariablen (auch "Rangwertvariable" genannt) sind die Rangplätze, die Untersuchungsobjekte in einer Messdimension hintereinander einnehmen.

Beispiel: Bei einem Marathon-Lauf trifft ein Läufer nach dem anderen im Ziel ein (dabei können auch 2 oder mehr gleichzeitig eintreffen). Ihre Rangplätze bilden die Rangvariable.

Eine Rangvariable kann auch durch eine einfache Transformation aus einer ordinalen oder quantitativen Variablen hervorgehen.

Betrachten wir folgende ordinale Variable

Schulbildung	Codeziffer
-----	-----
Volksschule	1
Hauptschule	2
Gymnasium	3
Fachschule	3
Universität	4

Die Codeziffern 1 bis 4 drücken eine Rangordnung im Bildungsniveau aus. 4 ist mehr als 3 und 3 ist mehr als 2 etc. Um wie viel mehr ist allerdings unbekannt. Die Differenzen zwischen den Bildungsstufen sind nicht bekannt. Die Ziffern drücken die Relation "mehr" oder "weniger" oder "gleich" aus.

Beachte: Gymnasium und Fachschule wurden gleichrangig betrachtet und deswegen beide mit 3 kodiert.

Unterschied zwischen ordinaler und Rang-Variabler

Bei der ordinalen Variablen werden den Ausprägungen der Variablen Rangplätze zugewiesen. Bei der Rang-Variablen werden den Untersuchungseinheiten Rangplätze zugewiesen.

Aus der ordinalen Variablen "Schulbildung" kann nun eine Rangvariable gebildet werden.

Von 7 Personen kennen wir die Schulbildung in Form ordinaler Codeziffern.

Person	Schulbildung	Wert in der ordinalen Variablen	Wert in der Rangvariablen
1	Volksschule	1	1
2	Hauptschule	2	2.5
3	Hauptschule	2	2.5
4	Gymnasium	3	5
5	Gymnasium	3	5
6	Fachschule	3	5
7	Universität	4	7

Der Wert der Rangvariablen ist sehr einfach der Rangplatz der Person, wenn alle Personen ihrer Schulbildung nach hintereinander gestellt werden. Da manche Personen dieselbe Schulbildung besitzen wie z.B. Person 2 und 3 bzw. eine als gleichrangig erachtete Schulbildung besitzen, wie z.B. Person 6 mit 4 und 5, wird eine "Rangteilung" vorgenommen. Person 2 und 3 teilen sich die Plätze 2 und 3. Der mittlere Wert ist 2.5. Person 4,5,6 teilen sich die Plätze 4,5,6. Der Wert in der Mitte ist 5.

Von "Bindung" wird gesprochen, wenn 2 oder mehr Personen denselben Rangplatz einnehmen. Wie am Beispiel gezeigt, wird dann üblicherweise eine "Rangteilung" vorgenommen.

In entsprechender Weise kann natürlich auch eine quantitative Variable in eine Rangvariable überführt werden (wobei allerdings ein Informationsverlust eintritt). Die Untersuchungsobjekte werden entsprechend ihren Werten in der quantitativen Variablen hintereinander gestellt. Bei gleichem Wert wird eine Rangteilung vorgenommen.

Werden 2 Rangvariable nach dem Kalkül des Produkt-Moment-Korrelationskoeffizienten korreliert (d.h. werden sie wie quantitative Variable behandelt), dann entsteht der Spearman'sche Rangkorrelationskoeffizient Rho.

Wird für eine unabhängige nominal-dichotome Variable und eine abhängige Rangvariable ein Allgemeines Lineares Modell gerechnet, dann entspricht dies dem U-Test nach Mann-Whitney. Bei dieser Analyse wird die Rangvariable also so behandelt, wie wenn sie quantitativ wäre.

P20.9.8 Analysen mit vielen unabhängigen nominalen Variablen

Um eine Analyse mit vielen unabhängigen nominalen Variablen rechnen zu können, darf der Benutzer

(1)keine Interaktionen verlangen

(2)und er muss (in der Programm-Maske Prog20mo in der letzten Optionsbox) durch "VERZICHTE=ZELLEN;" auf die Zellenhäufigkeiten verzichten.

Außerdem kann es auch notwendig werden, "VERZICHTE=EFFEKTE;" zu schreiben, um Speicherplatz einzusparen.

P20.10 Hierarchische Regression

Eine ausführliche Darstellung der hierarchischen Regression ist enthalten in Gaensslen/ Schubö (1973, Kapitel 10.2) und in Holm, 1979, Abschnitt 18.2.2 (wir sprechen dort von "schrittweiser" Regression).

Betrachten wir ein Beispiel mit einer abhängigen Variablen y und 3 unabhängigen Variablen V_1, V_2, V_3 . Die nur aus 1.0 bestehende Konstante bezeichnen wir mit K .

Der Vorgang ist folgender:

1. Aus der Variablen V_3 werden die Konstante K und die Variablen V_1, V_2 "herausgenommen". Es entsteht die Partialvariable $V_{3,12}$. Mit V_3 als abhängiger Variabler und V_1, V_2 als unabhängiger Variabler wird eine Regressionsanalyse gerechnet.

$$V_3^* = \beta_0 + \beta_{13,2} V_1 + \beta_{23,1} V_2$$

Die Partialvariable $V_{3,12}$ ergibt sich dann als die Differenz

$$V_{3,12} = V_3 - V_3^*$$

Das Herausnehmen von K hat keine spezifische Wirkung.

2. Aus der Variablen V_2 werden die Konstante K und V_1 "herausgenommen". Es entsteht die Partialvariable $V_{2,1}$.

3. Aus V_1 wird die Konstante K herausgenommen. Es entsteht die Partialvariable $V_{1,k}$. Dabei gilt

$$V_{1,k} = V_1 - M_1$$

(wobei M_1 = Mittelwert von V_1).

Die Gleichung der hierarchischen Regression für dieses Beispiel lautet nun:

$$y = M_y + \beta_{1y}(V_1 - M_1) + \beta_{2y,1} V_{2,1} + \beta_{3y,12} V_{3,12} + e$$

M_y ist die Regressionskonstante. Sie ist im Fall der hierarchischen Regression gleich dem Mittelwert von y .

Im Programm Prog20mo (Abschnitt P20.8.1) wird die hierarchische Regression sehr einfach dadurch veranlasst, dass in der Optionsbox „Verfahren“ der Eintrag „sequentiell“ ausgewählt wird. Siehe dazu die Erläuterungen zu dieser Box in Abschnitt P20.8.1.1.

Die im Folgenden dargestellte Methode, eine hierarchische Regression herzustellen, ist etwas komplizierter. Sie ist jedoch flexibler und eröffnet zusätzliche Möglichkeiten.

Die Eingabe in ALMO für diese Methode lautet im Maskenprogramm Prog20mx (Abschnitt P20.8.0) oder Prog20mo (Abschnitt P20.8.1).



Die Variablen-Hierarchie wird also durch Schrägstriche / ausgedrückt.

Nach dem Schrägstrich braucht „V“ nicht wiederholt zu werden. Sie können aber auch V1 / V2 / V3 schreiben. Selbstverständlich können anstelle der Vx auch Variablennamen geschrieben werden, z.B. Alter / Größe / Körpergewicht.

Die Variablen V1, V2, V3 könnten auch ordinal sein. Dann wäre im Maskenprogramm im letzten Eingabefeld der abgebildeten Box die durch Schrägstriche getrennten Variablen zu schreiben

Die Variablen könnten auch gemischt quantitativ und ordinal sein.

Im Maskenprogramm Prog20mo werden in der Box für die unabhängigen Variablen die quantitativen und ordinalen Variablen in ihre jeweiligen Eingabefelder eingetragen. Schrägstriche werden dabei nicht verwendet.

Als Trennzeichen (wenn mehrere Variable in ein Eingabefeld geschrieben werden) dient das Komma. Dann muss die Optionsbox "Option: Programm-Optionen lt. Handbuch" geöffnet werden.



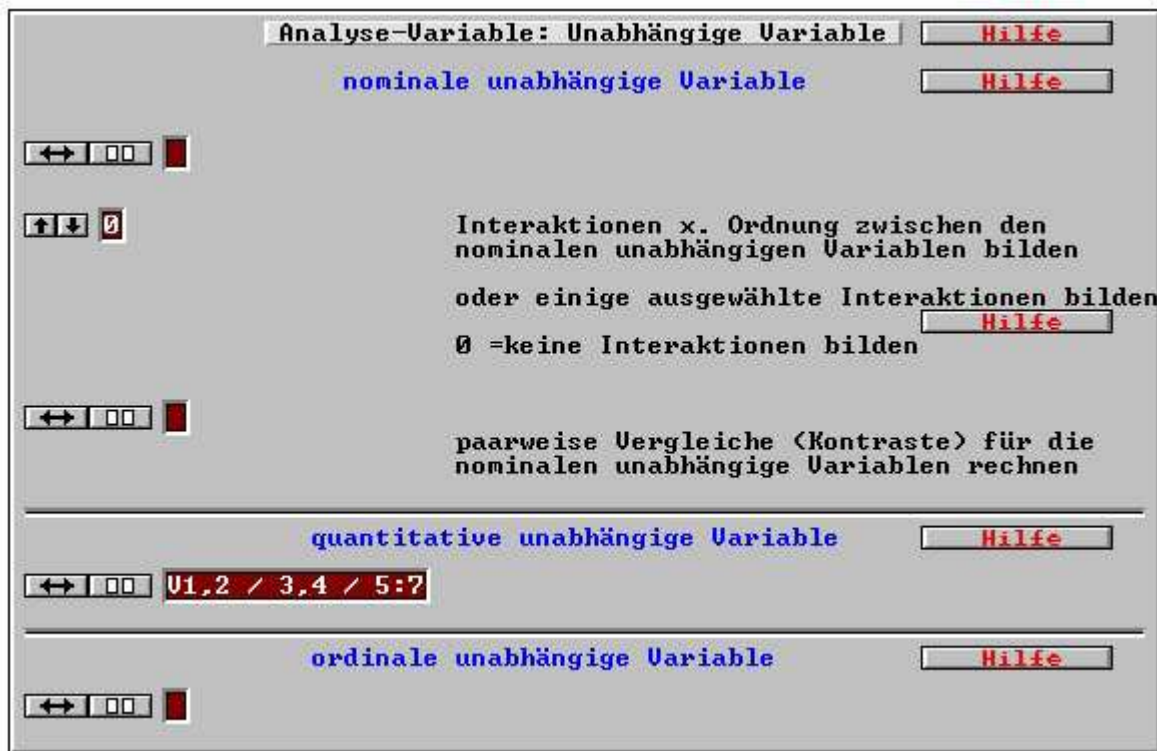
In ihr wird dann eingetragen:



Beachte: Sobald der Benutzer die unabhängigen quantitativen Variable durch einen Schrägstrich trennt, gibt ALMO als Regressionskonstante (= als Effekt der Konstanten) den Mittelwert der abhängigen Variablen aus.

P20.10.1 Gruppenweise hierarchische Regression

Im Maskenprogramm Prog20mx (Abschnitt P20.8.0) oder Prog20mo (Abschnitt P20.8.1) wird die Box für die unabhängigen Variablen so ausgefüllt:



Durch die Schrägstriche werden die hierarchischen Gruppen voneinander getrennt.

V1,2	/	3,4	/	5:7
1. hierarchische Gruppe		2. hierarchische Gruppe		3. hierarchische Gruppe

Es gilt das Prinzip: Aus den Variablen einer hierarchischen Gruppe werden die Variablen, der vor ihr stehenden hierarchischen Gruppen auspartiielliert. Außerdem werden die Variablen innerhalb einer hierarchischen Gruppe gegeneinander auspartiielliert.

V5:7 bilden die 3. hierarchische Gruppe. Aus ihnen werden die Variablen, der vor ihnen stehenden hierarchischen Gruppen auspartiielliert. Dann werden V5 bis 7 gegeneinander auspartiielliert. Die 3. hierarchische Gruppe hat jedoch keinen Einfluss auf die vor ihr stehenden hierarchischen Gruppen. V3,4 bilden die 2. hierarchische Gruppe. Aus ihnen werden V1,2 auspartiielliert. Dann werden V3 und 4 gegeneinander auspartiielliert.

Die Gleichung, die dieser ALMO-Eingabe entspricht, lautet also:

$$\begin{aligned}
 y' = & M_y + \\
 & \beta_{1y,2} (V_1 - M_1) + \beta_{2y,1} (V_2 - M_2) + \\
 & \beta_{3y,124} V_{3,12} + \beta_{4y,123} V_{4,12} + \\
 & \beta_{5y,123467} V_{5,1234} + \beta_{6y,123457} V_{6,1234} + \beta_{7y,123456} V_{7,1234}
 \end{aligned}$$

Beachte: Sobald bei ordinalen bzw. quantitativen Variablen ein Schrägstrich verwendet und in Prog20mo in der Optionsbox "Streuungsmatrix" QUADRATSUMME gesetzt wird, behandelt ALMO die Variablen der 1. Gruppe (in unserem Beispiel V1, V2) als Partialvariable, aus denen die Wirkung der Konstanten "herausgenommen" wurde. Das bedeutet, dass ALMO als Regressionskonstante den y-Mittelwert ausgibt.

ALMO liefert folgendes Ergebnis (gekürzt):

Almo gibt die Ergebnisse in umgekehrter Reihenfolge aus, beginnt also mit der hinteren hierarchischen Gruppe.

Alle im Folgenden angegebenen Streuungen und erklarte Streuungen sind Varianzen standardisierter Variabler

```
=====
Gesamtstreuung                                1.000000
=====
```

Koeffizienten fuer Gesamt-Modell

```
Durch alle unabh. Variable
erklarte Streuung                                0.116536
Fehlerstreuung                                0.883464
multipler Korrelat.koeff.                      0.341374
F-Wert f. erklarte Streuung                    0.998732
Freiheitsgrade Nenner = 7
                Zaehler= 53
Signifikanz: p                                0.443311
Signifikanz: (1-p)*100                        55.668894 %
Teststaerke von F                             0.387681
=====
```

Koeffizienten fuer quantitat./ordinale Variable aus univariater Analyse hinsichtlich der abhaeng. Var. V20 Y

	Regr. koeff.	Standard fehler	95% Konfidenz- bereich nach oben u.unten	erklarte Streuung	part. Korrel.	F-Wert	Signifikanz p	(1-p)100	df1	df2	Test- staerke
V5	-0.2043	0.1444	0.2896	0.0334	-0.191	2.001	0.163	83.68	1	53	0.2847
V6	-0.0770	0.1375	0.2757	0.0052	-0.077	0.314	0.577	42.27	1	53	0.0853
V7	-0.0967	0.1379	0.2765	0.0082	-0.096	0.492	0.486	51.41	1	53	0.1058
hierarchische Gruppe V5,6				0.0436	0.217	0.872	0.464	53.62	3	53	0.2295
V3	0.1799	0.1291	0.2590	0.0324	0.188	1.942	0.169	83.06	1	53	0.2777
V4	-0.1954	0.1300	0.2608	0.0376	-0.202	2.258	0.139	86.10	1	53	0.3147
hierarchische Gruppe V3,4				0.0700	0.271	2.100	0.131	86.93	2	53	0.4174
V1	-0.0055	0.1318	0.2644	0.0000	-0.006	0.002	0.965	3.51	1	53	0.0502
V2	0.0525	0.1318	0.2644	0.0026	0.055	0.159	0.692	30.84	1	53	0.0677
hierarchische Gruppe V1,2				0.0029	0.057	0.087	0.909	9.07	2	53	0.0629

Zur Ermittlung des F-Wertes und der partiellen Korrelation wird die Fehlerstreuung des Gesamtmodells verwendet.

Anwendungsbeispiel

Bei einer Untersuchung der Leistung von Probanden (in irgend einer Eigenschaft) werden als Ursachen mehrere physische und mehrere psychologische Variable verwendet. Dabei könnte es sich erweisen, dass die Gruppe der psychologischen Variablen zusammen erheblich mehr an Streuung erklären als die der physischen Variablen. Wie im nächsten Abschnitt dargestellt, müssen dabei die Variablen-Gruppen nicht notwendigerweise hierarchisch hintereinander stehen. Sie können auch gleichrangig sein.

P20.10.2 Regression von Polynomen als hierarchische Regression

Betrachten wir folgende Gleichung:

$$y = \beta_0 + \beta_1x + \beta_2z + \beta_3x^2 + \beta_4z^2 + \beta_5x^3 + \beta_6z^3$$

Die Variable y wird nicht durch x und z , sondern auch durch x^2 , x^3 , und z^2 , z^3 erklärt. In obiger Gleichung wird eine Gleichrangigkeit aller Variablen angenommen.

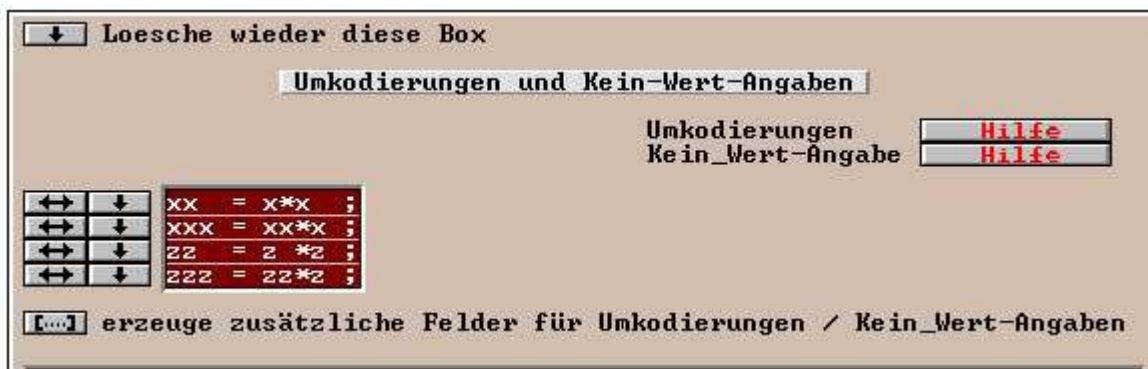
Hier ist es nun aber sinnvoll, eine Hierarchie der Variablen einzuführen, und zwar in der Weise, dass den kubischen Variablen x^3 , z^3 nicht erlaubt wird, (auspartiellierenden) Einfluss auf die vor ihnen stehenden Variablen, x , z , x^2 , z^2 zu nehmen. Weiterhin wird den quadratischen Variablen x^2 , z^2 nicht erlaubt, (auspartiellierenden) Einfluss auf x und z zu nehmen. Siehe auch bei Winer, 1971, S. 507 bis 510.

Beim Maskenprogramm Prog20mx oder Prog20mo wird man zuerst in der Box „Freie Namensfelder“ den einfachen, den quadratischen und den kubischen Variablen einen Namen geben – etwa so:



Aus einer Datei, die 3 Variable umfasst, werden x, z, y eingelesen. Die quadratischen und kubischen Variablen erhalten die Nummern 4 bis 7.

Die quadratischen und kubischen Variablen müssen dann noch aus den einfachen (aus der Datei eingelesenen) Variablen x und z gebildet werden. Das geschieht in der Umkodierungs-Box.



In der Box "Analyse-Variable: Unabhängige Variable wird dann eingetragen



P20.10.3 Gründe für die Hierarchisierung

Folgende 3 Gründe sind möglich:

1. Die Regressionsgleichung ist, wie oben beschrieben, ein Polynom.
2. Es ist dann sinnvoll, hierarchische Gruppen zu bilden, wenn eine kausale Ordnung zwischen den unabhängigen quantitativen Variablen besteht (bzw. wenn man eine solche unterstellt). In unserem Beispiel in P20.10.1 unterstellen wir, dass die Variablen der 1. Gruppe, also V1 und V2 nicht nur die abhängige Variable y determinieren, sondern auch kausal die Variablen der 2. und 3. Gruppe bestimmen. Die Variablen der 2. Gruppe stehen kausal vor denen der 3. Gruppe.
3. Die durch die hierarchischen Gruppen erklärten Streuungen addieren sich zur gesamten erklärten Streuung. Wenn man nach jeder Variablen einen Schrägstrich zieht, dann erhält man für die unabhängigen Variablen erklärte Streuungen, die orthogonal sind, d.h. sich zur gesamten erklärten Streuung addieren.

P20.11 Gleichrangige Gruppen bei den Kovarianten: Partielle multiple Korrelation

Durch einen senkrechten Strich werden gleichrangige Gruppen gebildet. Betrachten wir ein Beispiel.

Unabh.QUANTITATIVE Var.	V1,2		3:6		7	;
	1. gleichrangige Gruppe		2. gleichrangige Gruppe		3. gleichrangige Gruppe	

Beim Maskenprogramm Prog20mx oder Prog20mo in Abschnitt P20.8.0 bzw. P20.8.1, muss in der Box „Analyse-Variable: Unabhängige Variable“ geschrieben werden.



Hier werden, wie in einer normalen Analyse, **alle** Variable gegeneinander auspartiiert. Der senkrechte Strich veranlasst ALMO dazu, die durch jede einzelne Gruppe erklärte Streuung, sowie den partiellen multiplen Korrelationskoeffizienten für jede Gruppe und dessen Signifikanz zusätzlich zu ermitteln.

ALMO liefert für obige Anweisung folgendes Ergebnis (gekürzt):

Alle im Folgenden angegebenen Streuungen und erklarte Streuungen sind Varianzen standardisierter Variabler

```

=====
Gesamtstreuung                1.000000
=====

Koeffizienten fuer Gesamt-Modell

Durch alle unabh. Variable
erklarte Streuung              0.198070
Fehlerstreuung                0.801930
multipler Korrelat.koeff.     0.445051
F-Wert f. erklarte Streuung   1.870080
Freiheitsgrade Nenner =      7
                          Zaehler= 53
Signifikanz: p                 0.092730
Signifikanz: (1-p)*100       90.726963 %
Teststaerke von F            0.688309
=====

```

Koeffizienten fuer quantitat./ordinale Variable aus univariater Analyse hinsichtlich der abhaeng. Var. V8

	Regr. koeff.	Standard fehler	95% Konfidenz- bereich nach oben u. unten	erklarte Streuung	part. Korrel.	F-Wert	Signifikanz p	(1-p)100	df1	df2	Test- staerke
V1	0.0400	0.1365	0.2737	0.0013	0.040	0.086	0.772	22.81	1	53	0.0595

V2	0.0971	0.1295	0.2596	0.0085	0.103	0.563	0.456	54.38	1	53	0.1141
zusammen				0.0087	0.104	0.289	0.754	24.63	2	53	0.0944
V3	0.0692	0.1307	0.2621	0.0042	0.073	0.281	0.598	40.19	1	53	0.0815
V4	-0.4093	0.1299	0.2605	0.1502	-0.397	9.928	0.003	99.73	1	53	0.8718
V5	-0.2530	0.1376	0.2759	0.0512	-0.245	3.382	0.072	92.85	1	53	0.4394
V6	-0.0731	0.1310	0.2627	0.0047	-0.076	0.311	0.579	42.12	1	53	0.0850
zusammen				0.1844	0.432	3.047	0.024	97.56	4	53	0.7708
V7	-0.1235	0.1314	0.2635	0.0134	-0.128	0.884	0.351	64.88	1	53	0.1519
zusammen				0.0134	0.128	0.884	0.351	64.88	1	53	0.1519

Erläuterung:

Die 1. Gruppe erklärt eine Streuung von 0.0087 und „partiellen multiplen Korrelationskoeffizienten“ von

$$R_{G_1} = R_{8.12.34567} = 0.104$$

Zur Notation: Vor dem 1. Punkt steht die abhängige Variable V8. Hinter dem 1. Punkt stehen die unabhängigen Variablen der 1. Gruppe V1 und V2. Hinter dem 2. Punkt stehen die „auspartiierten übrigen Variablen, also V3, 4, 5, 6, 7.

P20.12 Hierarchische und gleichrangige Gruppen bei quantitativen und ordinalen Kovarianten

Auch innerhalb einer hierarchischen Gruppe können gleichrangige Subgruppen gebildet werden. Das geschieht durch Schrägstriche und senkrechte Striche, z.B. so

V1,2 / 5,6 | 7,8 | 9 / 10 | 11,12 ;

Beim Maskenprogramm Prog20mx oder Prog20mo in Abschnitt P20.8.0 bzw. P20.8.1, muss in der Box „Analyse-Variable: Unabhängige Variable“ geschrieben werden.

Analyse-Variable: Unabhängige Variable Hilfe

nominale unabhängige Variable Hilfe

Interaktionen x. Ordnung zwischen den nominalen unabhängigen Variablen bilden
 oder einige ausgewählte Interaktionen bilden
 Ø =keine Interaktionen bilden Hilfe

paarweise Vergleiche (Kontraste) für die nominalen unabhängigen Variablen rechnen

quantitative unabhängige Variable Hilfe

U1,2 / 5,6 | 7,8 | 9 / 10 | 11,12

ordinale unabhängige Variable Hilfe

Beachte: Die gleichrangigen Gruppen bestehen **innerhalb** der hierarchischen Gruppe. Wir haben also zunächst 3 hierarchische Gruppen:

V1,2 / 5,6,7,8,9 / 10,11,12

Betrachten wir die 2. hierarchische Gruppe. Aus ihr werden die Variablen der 1. hierarchischen Gruppe auspartiiert. Dann werden die Variablen 5,6,7,8,9 gegeneinander auspartiiert.

Innerhalb der 2. hierarchischen Gruppe bestehen 3 gleichrangige Subgruppen:

V5,6 bilden die 1. Subgruppe

V7,8 bilden die 2. Subgruppe

V9 bildet die 3. Subgruppe.

Für jede der 3 Subgruppen werden nun noch die erklärte Streuung und der partielle multiple Korrelationskoeffizient ermittelt.

ALMO liefert für obige Anweisung folgendes Ergebnis (gekürzt):

Almo gibt die Ergebnisse in umgekehrter Reihenfolge aus, beginnt also mit der hinteren hierarchischen Gruppe.

Das Wort "zusammen" kennzeichnet die gleichrangige Gruppe.

Alle im Folgenden angegebenen Streuungen und erkläerte Streuungen sind Varianzen standardisierter Variabler

```
=====
Gesamtstreuung                                1.000000
=====
```

Koeffizienten fuer Gesamt-Modell

```
Durch alle unabh. Variable
erkläerte Streuung                            0.686781
Fehlerstreuung                               0.313219
multipler Korrelat.koeff.                    0.828723
F-Wert f. erkläerte Streuung                 10.963286
Freiheitsgrade Nenner = 10
                Zaehler= 50
Signifikanz: p                               0.000006
Signifikanz: (1-p)*100                       99.999433 %
Teststaerke von F                            1.000000
=====
```

Koeffizienten fuer quantitat./ordinale Variable aus univariater Analyse hinsichtlich der abhaeng. Var. V20 Y

	Regr. koeff.	Standard fehler	95% Konfidenz-bereich nach oben u. unten	erkläerte Streuung	part. Korrel.	F-Wert	Signifikanz p	(1-p)100	df1	df2	Test-staerke
V10	0.0884	0.1152	0.2313	0.0037	0.108	0.589	0.446	55.37	1	50	0.1170
zusammen				0.0037	0.108	0.589	0.446	55.36	1	50	0.1170
V11	-0.0500	0.1223	0.2456	0.0010	-0.058	0.167	0.684	31.60	1	50	0.0686
V12	0.0062	0.1177	0.2363	0.0000	0.007	0.003	0.956	4.37	1	50	0.0503
zusammen				0.0011	0.059	0.088	0.909	9.13	2	50	0.0629
hierarchische Gruppe V10,11,12				0.0038	0.109	0.202	0.893	10.70	3	50	0.0857

V5	-0.0452	0.0849	0.1705	0.0018	-0.075	0.284	0.596	40.41	1	50	0.0818
V6	-0.0251	0.0820	0.1648	0.0006	-0.043	0.094	0.762	23.79	1	50	0.0604
zusammen				0.0024	0.086	0.188	0.828	17.15	2	50	0.0783
V7	0.0241	0.0833	0.1673	0.0005	0.041	0.084	0.774	22.55	1	50	0.0593
V8	0.8190	0.0809	0.1624	0.6427	0.820	102.603	0.000	100.00	1	50	1.0000
zusammen				0.6469	0.821	51.633	0.000	99.99	2	50	1.0000
V9	-0.0800	0.0814	0.1636	0.0060	-0.138	0.964	0.331	66.93	1	50	0.1612
zusammen				0.0060	0.138	0.964	0.331	66.92	1	50	0.1612
hierarchische Gruppe				0.6801	0.827	21.713	0.000	99.99	5	50	1.0000
V5,6,7,8,9											
V1	-0.0055	0.0808	0.1623	0.0000	-0.010	0.005	0.943	5.66	1	50	0.0505
V2	0.0525	0.0808	0.1623	0.0026	0.092	0.422	0.518	48.19	1	50	0.0977
hierarchische Gruppe				0.0029	0.096	0.232	0.795	20.48	2	50	0.0851
V1,2											

Die Bildung von hierarchischen und gleichrangigen Gruppen ist auch bei den unabhängigen ordinalen Variablen möglich.

Beim Maskenprogramm erfolgt die Eingabe dann im Eingabefeld für die ordinalen Variablen.

Unabhängige ordinale und quantitative Variable können auch gemischt werden.

Dann muss im Maskenprogramm Prog20mo folgendermaßen verfahren werden:

In der Box für die unabhängigen Variablen werden die quantitativen und ordinalen Variablen in ihre jeweiligen Eingabefelder eingetragen. Schrägstriche und senkrechte Striche werden dabei nicht verwendet. als Trennzeichen (wenn mehrere Variable in ein Eingabefeld geschrieben werden) dient das Komma.

The screenshot shows two input sections. The top section is titled "quantitative unabhängige Variable" and contains a text box with the input "v1,5,6,7,8" and a "Hilfe" button. The bottom section is titled "ordinale unabhängige Variable" and contains a text box with the input "v2,10" and a "Hilfe" button.

Dann muss die Optionsbox "Option: Programm-Optionen lt. Handbuch" geöffnet werden.

The screenshot shows a dialog box titled "Option: Programm-Optionen lt. Handbuch" with a "Hilfe" button in the top right corner and a downward-pointing arrow button on the left side.

In ihr wird dann eingetragen:



P20.13 Frei gewählte Hierarchie bei den nominalen Variablen

Auch die unabhängigen nominalen Variablen können in hierarchische Gruppen eingeteilt werden.

Beim „selbst geschriebenen“ Syntaxprogramm:

V 7,9	/	10,11	/	13;
1.hierarchische Gruppe		2.hierarchische Gruppe		3.hierarchische Gruppe

Bei den Maskenprogrammen, z.B. bei Prog20mx oder Prog20mo in Abschnitt P20.8.0 bzw. P20.8.1, muss in die Box "Analysevariable: Unabhängige Variable" geschrieben werden.



Die hierarchische Ordnung ist sinnvoll, wenn (1) eine kausale Ordnung zwischen den unabhängigen nominalen Variablen besteht, (2) wenn orthogonale, d.h. addierbare erklärte Streuungen gewünscht sind. Siehe dazu unsere ausführlichen Darstellungen bei der Hierarchisierung der Kovarianten in P20.10.3.

Die Bildung von **gleichrangigen** Gruppen bei den nominalen Variablen ist (vorläufig noch) nicht möglich.

Auch eine Kovarianzanalyse ist möglich, bei der die nominalen hierarchisch geordnet und die Kovarianten hierarchisch und gleichrangig geordnet sind.

Beispiel:

Im Maskenprogramm Prog20mo (Abschnitt P20.8.1) ist die Box für die unabhängigen Variablen in folgender Weise auszufüllen:

Analyse-Variablen: Unabhängige Variable Hilfe

nominale unabhängige Variable Hilfe

U20 / 21,24

0

Interaktionen x. Ordnung zwischen den nominalen unabhängigen Variablen bilden oder einige ausgewählte Interaktionen bilden Hilfe

0 =keine Interaktionen bilden

paarweise Vergleiche (Kontraste) für die nominalen unabhängige Variablen rechnen

quantitative unabhängige Variable Hilfe

U1, 5, 7, 8, 10

ordinale unabhängige Variable Hilfe

U2, 6

Die nominalen Variablen werden durch Schrägstriche getrennt. Wären zusätzlich nur quantitative oder nur ordinale Variable vorhanden, dann würden diese ebenfalls durch Schrägstriche getrennt in ihr jeweiliges Eingabefeld geschrieben werden.

In unserem Beispiel sind jedoch quantitative **und** ordinale Variable vorhanden.

In diesem Fall müssen die quantitativen und die ordinalen Variablen in ihre jeweiligen Eingabefelder eingetragen werden. Schrägstriche und senkrechte Striche werden dabei nicht verwendet. Als Trennzeichen (wenn mehrere Variable in ein Eingabefeld geschrieben werden) dient das Komma.

Dann muss die Optionsbox "Option: Programm-Optionen lt. Handbuch" geöffnet werden.

Hilfe

Option: Programm-Optionen lt. Handbuch

In ihr wird dann eingetragen:



Die unabhängigen quantitativen und ordinalen Variablen bilden die Menge der Kovarianten. Sie werden, wie unter REIHUNG angegeben, hierarchisch und gleichrangig geordnet.

Die Klasse der unabhängigen nominalen Variablen und die Klasse der Kovarianten werden - wie bei der normalen Kovarianzanalyse - zuerst **gegeneinander** auspartielliert. Erst dann findet die hierarchische bzw. gleichrangige Gruppierung innerhalb der beiden Klassen statt.

Beachte:

1. Nach dem Wort REIHUNG dürfen nur quantitative und ordinale, nicht jedoch nominale Variable stehen. Eine Mischung der nominalen mit unabhängigen quantitativen und ordinalen Variablen ist also nicht möglich.
2. Durch Verwendung der PARTIAL-Anweisung können weitere komplizierte hierarchische Gruppierungen erzeugt werden. Siehe dazu im Handbuch Teil3b, Abschnitt P19.3 und hier nachfolgend Abschnitt P20.15.

Hinweis zu 1. Die Mischung von nominalen mit quantitativen und ordinalen Variablen ist jedoch möglich, wenn die Partial-Anweisung verwendet wird. Beispiel: Man möchte schreiben

$$\text{Reihung} = V1 / 2, 3, 4;$$

wobei V1 nominal und V2,3,4 quantitativ/ordinal sind. Diese Anweisung ist jedoch unzulässig. Die gewünschte Hierarchisierung kann allerdings durch folgende Partial-Anweisung erreicht werden:

$$\text{Partial} = V1 \text{ aus } V2, 3, 4;$$

Siehe dazu Abschnitt P 19.3 und P 20.15.

P20.13.1 Kovarianzanalyse und multivariate Analyse mit frei gewählter Hierarchie

Selbstverständlich sind alle Möglichkeiten und Optionen des Allgemeinen linearen Modells, sofern sie inhaltlich sinnvoll sind, möglich. So kann auch eine Kovarianzanalyse gerechnet werden. Bei ihr werden die nominalen Variablen (Dummies) und die Kovarianten zuerst gegeneinander auspartielliert. Erst danach werden hierarchische und gleichrangige Gruppen bei den Kovarianten bzw. die hierarchischen Gruppen bei den nominalen Variablen gebildet.

Durch Verwendung der PARTIAL-Anweisung kann die gegenseitige Auspartiellierung von Kovarianten und nominalen Variablen ganz oder zum Teil zu einer einseitigen Auspartiellierung verändert werden. Siehe dazu Abschnitt P 20.15.1.

Auch die multivariate Analyse ist möglich. Die abhängigen Variablen können mehrere ordinale bzw. quantitative oder eine nominale Variable sein.

P20.14 Bildung von Interaktionsvariablen

Die Bildung der Interaktionsvariablen wollen wir an einem Beispiel betrachten:

A1			A2			A3		
B1	B2	B3	B1	B2	B3	B1	B2	B3
1	2	-	3	4	-	-	-	-

Die Interaktionsvariable AB entsteht aus der Kombination von A und B - wie in obiger Tabelle dargestellt. Diese beiden besitzen 3 Ausprägungen. ALMO betrachtet nur die notwendigen (nicht-redundanten) Dummies A1, A2 sowie B1, B2.

Die Interaktionsvariable AB besitzt also die 4 notwendigen Dummies A1B1 A1B2 A2B1 A2B2, denen die Werte 1,2,3,4 zugewiesen werden. Die anderen Kombinationen, die wir in obiger Tabelle durch einen Strich markiert haben, erhalten den Wert 5. Im Verlauf des Kalküls werden sie ausgeschlossen. ALMO ermittelt also für die Interaktionsvariable AB die Obergrenze von 5.

Bildung von Interaktionsvariablen

Bei den Maskenprogramm Prog20mx und Prog20mo (Abschnitt P20.8.0 und P20.8.1) wird die Interaktion in der Box "Analyse-Variable: Unabhängige Variable" festgelegt.

a. Der Normalfall

Betrachten wir ein Beispiel:



Die unabhängigen nominalen Variablen sind A, B, C.

Wird als Interaktion in das Eingabefeld 3 geschrieben, dann werden alle Interaktionen bis zur 3. Ordnung gebildet, also

Interaktionen 2. Ordnung: AB, AC, BC

Interaktionen 3. Ordnung: ABC

b. Spezifische Interaktionsvariable

Sollen Interaktionen explizit und spezifisch gebildet werden, dann ist die Vorgehensweise beim Maskenprogramm (etwas kompliziert) folgende.

1. In der Box "Freie Namensfelder" werden den Interaktionsvariablen Namen gegeben

Die Namen können Sie wählen, wie Sie wollen.

Verwenden Sie für die Interaktionen hinter "Name xx" Nummern die frei sind; am besten Nummern, die höher sind als die Nummer der letzten eingelesenen Variablen. Die Zahl der vereinbarten Variablen in der obersten Eingabebox der Programm-Maske muss dann eventuell erhöht werden.

2. In der Box "Analysevariable: unabhängige Variable" schreiben Sie folgende Eingabe:

Sätze zu Interaktionsvariablen

1. Der Benutzer muss angeben, aus welchen nominalen Variablen die Interaktionsvariablen gebildet werden sollen. Anstelle des Wortes "mal" kann auch das Multiplikationszeichen, der Stern *, geschrieben werden, also

2. Die Interaktionsvariable erhalten Variablennummern, die frei sind. Am besten ist es man verwendet Variablennummern, die höher sind als die letzte eingelesene Variable. Beispiel:

Der Datensatz umfasst die Variablen V1 bis V20

Die Vereinbare-Anweisung lautet: VereinbareVariable = 30

Dann kann man den Interaktionsvariablen die Nummern 21 bis 30 zuweisen.

- Die Variablen, aus denen die Interaktionsvariablen gebildet werden, müssen im Eingabefeld vor der Interaktionsvariablen stehen.
- Interaktionen der Ordnung 3 und höher können ihrerseits aus Interaktionen gebildet werden.

IntABCDE ist IntABC mal IntDE

Voraussetzung dafür ist, dass die Elemente (in unserem Beispiel) IntABC und IntDE schon definiert sind. Die Bildung von Interaktionsvariablen in dieser Weise verringert die Schreiarbeit und etwas die Rechenzeit.

- Rechts von "ist" können maximal nur 4 Variable stehen. Es können also nur maximal 4er-Interaktionsvariable aus Hauptvariablen gebildet werden. Sollen Interaktionen höherer Ordnung gebildet werden, z.B. 5er-Interaktionen, dann müssen diese ihrerseits aus Interaktionsvariablen gebildet werden.

Wird bei dieser Art der Eingabe die Optionsbox "Verfahren" nicht geöffnet, dann entsteht ein Ergebnis nach dem Verfahren der weighted squares of means (das empfehlenswerte Verfahren)

Hierarchisierung durch Schrägstriche

Wird die Box "Analyse-Variable: Unabhängige Variable" in folgender Form ausgefüllt dann wird nach dem Verfahren der fitting_constants_I gerechnet.



Die Optionsbox "Verfahren" darf nicht geöffnet werden. Almo bringt dann die irreführende Mitteilung, es würde nach dem Verfahren der weighted squares of means rechnen. Tatsächlich jedoch wird durch die Schrägstriche ein gruppenweise hierarchisches Modell erzeugt - eben ein fitting_constants I-Modell.

Die Schrägstriche in A,B,C / IntAB,IntAC,IntBC / IntABC bewirken eine Hierarchisierung. Es wird also ein der Methode der "fitting constants I" äquivalente Lösung ermittelt. Werden die Schrägstriche weggelassen (d.h. durch Beistriche ersetzt), dann entsteht eine "weighted-squares-of-means"-Lösung, die in der Regel vorzuziehen ist. Die Optionsbox "Verfahren" darf nicht geöffnet werden.

P20.14.1 Einbeziehung von nur einigen Interaktionsvariablen

Im Maskenprogramm Prog20mx oder Prog20m0 schreiben wir:



Der Schrägstrich in "V1:10 / V12,23 kann weggelassen werden. Dann entsteht eine "weighted-squares-of-means"-Lösung. Siehe oben "Hierarchisierung durch Schrägstriche".

Beachte: Die Variablennummern V12 und V23 können für diese beiden Interaktionsvariablen nur dann verwendet werden, wenn sie frei sind. Das setzt voraus, dass der eingelesene Datensatz nur die Variablen V1 bis V11 umfasst und die Zahl der vereinbarten Variablen mindestens 23 ist. Werden Interaktionsvariable auf diese Art gebildet, dann kommt man ohne explizite Namensgebung aus. Die beispielsweise aus V1 und V2 gebildete Interaktionsvariable könnte aber auch V33 sein. In der 2. Zeile der Eingabebox würde man dann definieren

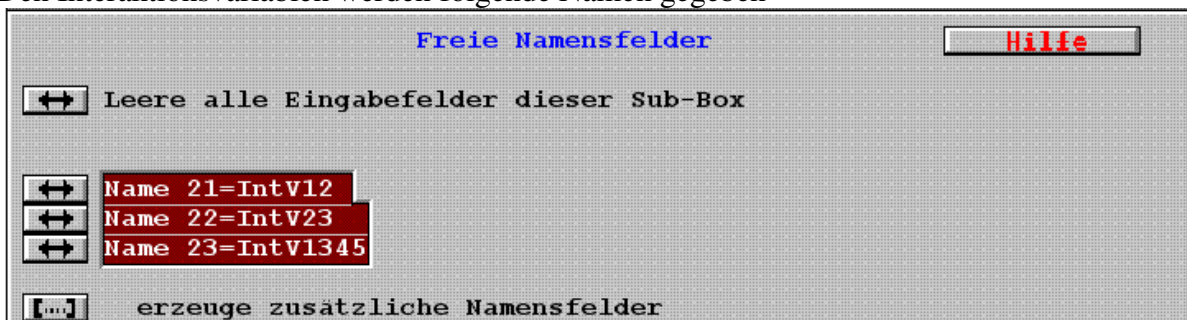
V33 ist V1 mal V2

Damit würde ausgedrückt, dass die Interaktion von V1 und V2 in die (freie) Variable V33 eingeschrieben wird. Wir raten ab, so vorzugehen. Es ist übersichtlicher und damit weniger fehleranfällig, die oben beschriebene Namensgebung zu verwenden.

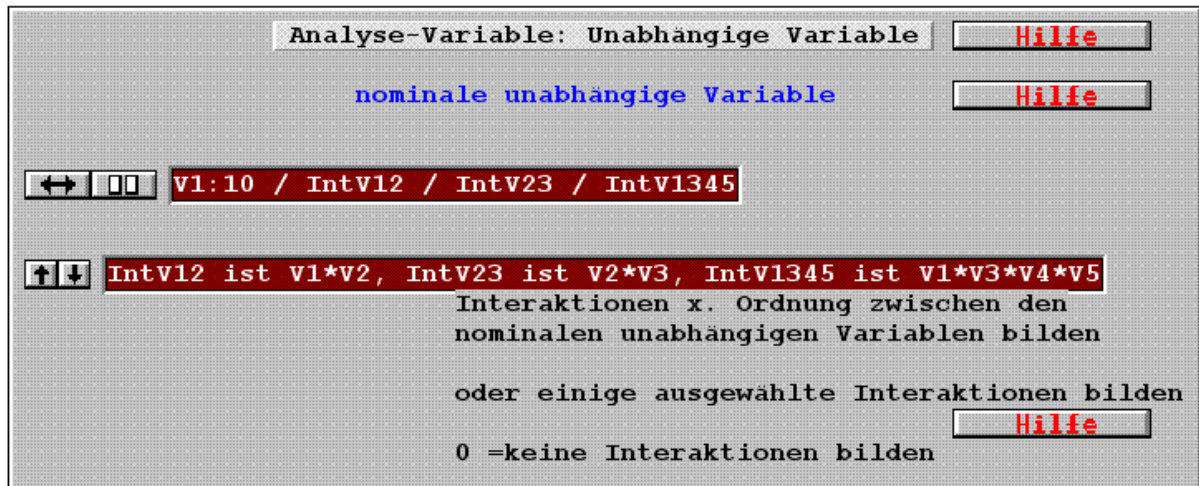
Bei gleichen Zellenhäufigkeiten sind die erklärten Streuungen, Effekte und Kontraste dieses "unvollständigen" Modells gleich denen des vollständigen - aber nur dann. Die Fehlerstreuung (die in ALMO als Reststreuung ermittelt wird), die PRE-Korrelationen, die Freiheitsgrade und damit der F-Wert und das Signifikanzniveau sind aber verschieden. Das "unvollständige" Modell ist nicht illegal. Es wird auf Informationen verzichtet, die an und für sich in den Daten enthalten sind.

Werden Schrägstriche verwendet, also eine Hierarchisierung durchgeführt, dann wäre folgende Anweisung problematisch:

Den Interaktionsvariablen werden folgende Namen gegeben



In der Box für die unabhängigen nominalen Variablen wird geschrieben



Wird eine Interaktion höherer Ordnung eingeführt, in unserem Beispiel IntV1345, dann sollten in das Modell noch die jeweiligen Interaktionen niedrigerer Ordnung aufgenommen werden, die in der betreffenden Interaktion höherer Ordnung enthalten sind. In unserem Beispiel wären das die Interaktionen

1*3, 1*4, 1*5, 3*4, 3*5, 4*5,
1*3*4, 1*3*5, 3*4*5.

(Wir schreiben die Interaktionen als Zahlenverknüpfungen der Hauptvariablen V1,3,4,5, siehe dazu auch Satz 1 in P20.14.1).

Werden keine Schrägstriche verwendet, dann ist eine solche Konstruktion eher akzeptabel. Der Benutzer muss sich aber dann im Klaren darüber sein, dass die gegenseitige Auspartiellierung der Variablen unvollständig ist.

Bei ungleichen Zellenhäufigkeiten ist das unvollständige Modell außerordentlich problematisch, da die 2er-Interaktionen nicht unabhängig voneinander sind. In unserem Beispiel erhalten wir für die Interaktionen V12, V23 des unvollständigen Modells andere Werte, als wir sie in einem Modell mit allen 2er-Interaktionen erhalten würden.

P20.14.2 Leere Zellen und Interaktionen

Im SPSS wird ein Beispiel von Milliken/Johnson vorgetragen, bei dem das unvollständige Modell – trotz ungleicher Zellenhäufigkeiten – sinnvoll ist. Die Daten sind folgende:

V1 FAT	V2 SURF	V3 FLOUR			
		1	2	3	4
1	1	6.7	4.3	5.7	-
1	2	7.1	-	5.9	5.6
1	3	-	5.5	6.4	5.8
2	1	-	5.9	7.4	7.1
2	2	-	5.6	-	6.8
2	3	6.4	5.1	6.2	6.3
3	1	7.1	5.9	-	-
3	2	7.3	6.6	8.1	6.8
3	3	-	7.5	9.1	-

ALMO liefert folgende Ergebnisse (gekürzt):

Streuungsquelle	Streuung	F-Wert	df	Signifikanz p	(1-p)100	Korrel Koeff.	Test- staerke
Gesamtstreuung	24.8354						
Fehlerstreuung	2.3159		14				
alle unabh. Var. zusammen	22.5195	12.3761	11	0.0001	99.9905	0.9522	1.0000
V1 FAT	10.1178	30.5826	2	0.0001	99.9950	0.9021	0.9999
V2 SURF	0.9972	3.0142	2	0.0804	91.9622	0.5486	0.4913
V3 FLOUR	8.6908	17.5128	3	0.0002	99.9850	0.8886	0.9999
V5 FatSurf	5.6388	8.5220	4	0.0014	99.8641	0.8419	0.9864

Betrachten wir ein weiteres Beispiel. Wir analysieren die Datei "C:\Almo\Testdat\Leertzell.fre" mit dem Maskenprogramm Prog20mo. Das Programm ist als Beispielprogramm unter dem Namen "Leertzell.Alm" in Almo enthalten. Der Benutzer findet es durch Klick auf das Menü "Almo / Liste aller Almo-Programme". Wir zeigen im Folgenden nur einige Boxen dieses Programms.

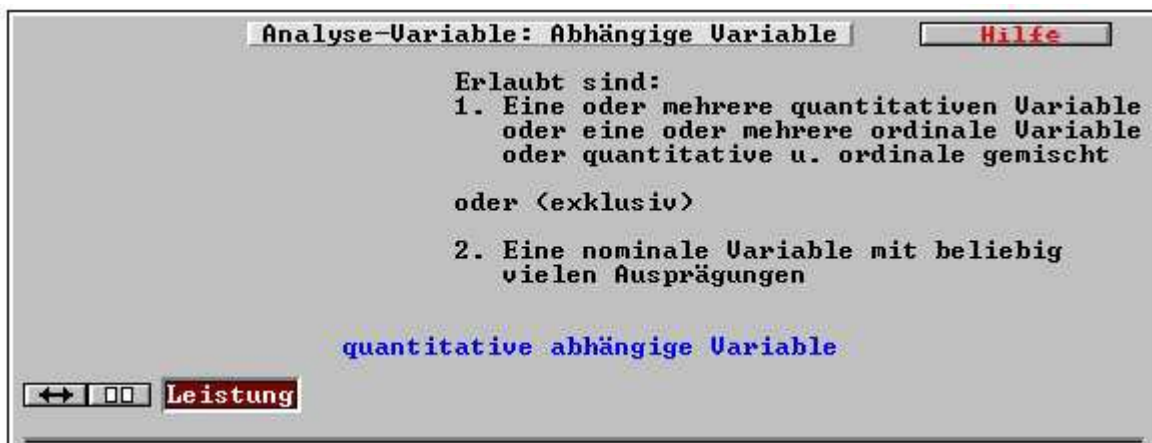


Die Datei enthält die 3 unabhängigen nominalen Variablen.

Geschlecht
Beruf
Familienstand

und die abhängige quantitative Variable

Leistung





Für die 3 nominalen Variablen werden alle Interaktionen angefordert.

Im Ergebnis wird folgende Warnung ausgegeben:

```

***** WARNUNG
Leere Zellen aufgetreten

3. Zelle ist leer
9. Zelle ist leer

Siehe weiter unten die Tabelle der Zellenhäufigkeiten

***** WARNUNG
Da leere Zellen vorhanden sind
sollten Sie mit "Verfahren=fitting_constants" rechnen

```

Die Tabelle der Zellenhäufigkeiten ist folgende:

Geschlec	Beruf	Familien	Leistung
männlich	Arbeiter	ledig	3
		verheira	5
		geschied	0
	Angestel	ledig	8
		verheira	8
		geschied	2
	Beamter	ledig	7
		verheira	1
		geschied	0
weiblich	Arbeiter	ledig	2
		verheira	3
		geschied	3
	Angestel	ledig	5
		verheira	4
		geschied	2
	Beamter	ledig	3
		verheira	4
		geschied	1
			-

<----- 3. Zelle leer

<----- 9. Zelle leer

Almo überprüft die Quadratsummenmatrix auf lineare Abhängigkeiten.

Diagonalglieder der Choleskymatrix
zur Ermittlung und zum Ausschluss
linearer Abhängigkeiten

Unabhängige Variable

```

A1          60.196721
B1          32.000000
B2          33.725490
C1          26.265988
C2          27.673711
A1 B1       30.156411
A1 B2       31.980819
A1 C1       22.770465
A1 C2       23.390180
B1 C1       10.467020
B1 C2       11.844843
B2 C1       11.318523
B2 C2       11.153186
A1 B1 C1    5.210161
A1 B1 C2    0.000000
A1 B2 C1    6.109091
A1 B2 C2   -0.000000
    
```

Almo eliminiert eine Variable i, wenn ihr Diagonalglied
aus der Choleskymatrix kleiner ist als $0.0001 * SS(i)$
 0.0001 kann ueber Option 48 veraendert werden.
Almo gibt eine Warnung aus, wenn das Cholesky-Glied
kleiner ist als $0.09 * SS(i)$ - einstellbar ueber Option 49
 $SS(i)$ ist das Diagonalglied ii der Var.i aus Streuungsmatrix

Folgende Variable werden eliminiert
A1 B1 C2
A1 B2 C2

Um die linearen Abhängigkeiten zu beseitigen eliminiert Almo 2 Dummy-Variable. Als
Effekte für die 3-er Interaktion gibt Almo aus (gekürzt):

Effekte von ABC

	Effekte	Standard- fehler	erklarte Streuung	
A1 B1 C1	1.8871	1.2681	16.2427	
A1 B1 C2	0.0000	0.0000	0.0000	<-----
A1 B1 C3	-1.8871	1.2681	16.2427	
A1 B2 C1	0.1163	1.0957	0.0826	
A1 B2 C2	0.0000	0.0000	0.0000	<-----
A1 B2 C3	-0.1163	1.0957	0.0826	
A1 B3 C1	-2.0034	1.3519	16.1051	
A1 B3 C2	0.0000	0.0000	0.0000	<-----
A1 B3 C3	2.0034	1.3519	16.1051	
A2 B1 C1	-1.8871	1.2681	16.2427	
A2 B1 C2	0.0000	0.0000	0.0000	<-----
A2 B1 C3	1.8871	1.2681	16.2427	
A2 B2 C1	-0.1163	1.0957	0.0826	
A2 B2 C2	0.0000	0.0000	0.0000	<-----
A2 B2 C3	0.1163	1.0957	0.0826	
A2 B3 C1	2.0034	1.3519	16.1051	
A2 B3 C2	0.0000	0.0000	0.0000	<-----
A2 B3 C3	-2.0034	1.3519	16.1051	

Bei den mit einem Pfeil gekennzeichneten Effekten ist die erklärte Streuung gleich 0.0000.

Das von Almo gelieferte Ergebnis ist problematisch. Almo bringt folgende Fehlermeldung:

Folgende Variable werden eliminiert

A1 B1 C2

A1 B2 C2

***** FEHLER

Muessen Dummies (insbesondere Interaktions-Dummies) eliminiert werden, dann koennen falsche Ergebnisse entstehen

Rechnen Sie eine Analyse ohne Interaktionen

Almo bricht den Kalkuel nicht ab

was tun? ---->

Hilfe

Werden nominale Dummies oder Interaktions-Dummies wegen linearer Abhängigkeiten eliminiert, dann kann die "Struktur der Datenmatrix", die beim Verfahren der weighted squares of means gefordert wird, so gestört sein, dass ein Teil der Ergebnisse falsch ist.

Für den Benutzer gibt es nun 4 Möglichkeiten zu reagieren.

1. Er akzeptiert das Ergebnis so wie es ausgegeben wurde. Nicht zu empfehlen.
2. Er rechnet ein Haupteffekte-Modell, verzichtet also auf Interaktionen.
3. Er rechnet, wie oben von Almo empfohlen das vollständige Modell, jedoch mit dem Verfahren der fitting-constants I. Dabei treten allerdings "wechselnde Werte" bei den Interaktionseffekten 3. und höherer Ordnung auf. Diese Lösung des Problems ist also problematisch.
4. Der Benutzer rechnet mit dem Verfahren der weighted squares of means - verzichtet dabei allerdings auf die 3-er Interaktionen. Es werden nur 2-er Interaktionen angefordert. Almo muss dann keine Variablen mehr eliminieren. In die Box "Analyse-Variable: Unabhängige Variable" wird eingetragen:



Nun rechnen wir dieselbe Analyse mit der Datei "C:\Almo\Testdat\Leerzel2.fre". Das Programm ist als Beispielprogramm unter dem Namen "Leerzel2.Alm" in Almo enthalten. Der Benutzer findet es durch Klick auf das Menü "Almo / Liste aller Almo-Programme".

Die Tabelle der Zellenhäufigkeiten ist nun folgende:

Geschlec	Beruf	Familien	Leistung
männlich	Arbeiter	ledig	3
		verheira	5
		geschied	0
	Angestel	ledig	8
		verheira	8
		geschied	2

<----- 3. Zelle leer

	Beamter	ledig	0	<----- 7. Zelle leer
		verheira	0	<----- 8. Zelle leer
		geschied	0	<----- 9. Zelle leer
weiblich	Arbeiter	ledig	2	
		verheira	3	
		geschied	3	
	Angestel	ledig	5	
		verheira	4	
		geschied	2	
	Beamter	ledig	3	
		verheira	4	
		geschied	1	
			-	

Um lineare Abhängigkeiten zu vermeiden eliminiert Almo die Dummies:

A1 B2
A1 B1 C2
A1 B2 C1
A1 B2 C2

Die Tabelle der Effekte weist für die 2-er Interaktion AB und die 3-er Interaktionen ABC für einige der Effekte eine erklärte Streuung von 0.0000 auf.

Der Benutzer sollte auf die Interaktionen AB und ABC verzichten und nur die verbleibenden Interaktionen AC und BC einbeziehen. In die Box "Freie Namensfelder" wird dann eingetragen:

In die Box "Analyse-Variable: Unabhängige Variable" ist einzutragen:

Hier wäre es auch möglich eine Hierarchie zwischen den nominalen Variablen und den Interaktionsvariablen einzuführen. Dies geschieht indem zwischen die beiden Variablengruppen ein Schrägstrich geschrieben wird. Natürlich werden dadurch die Ergebnisse verändert.

Unsere Empfehlung ist folgende:

Wenn Almo leere Zellen meldet, dann sollte man die Interaktionen, bei denen einzelne Dummies eliminiert werden (bzw für die Effekte ausgegeben werden, deren erklärte Streuung 0.0000 ist), aus der Analyse ausschließen. Das scheint uns eine klare Lösung des Problems der leeren Zellen zu sein.

In unserem letzten Beispiel meldet Almo, dass es die Dummy-Variable A1B2 eliminieren muss. Also nehmen wir die Interaktion AB zur Gänze heraus. Weiter meldet Almo, das die Dummy-Variable A1B1C2, A1B2C1, A1B2C2 eliminiert werden müssen. Also nehmen wir die Interaktion ABC zur Gänze heraus.

Vielleicht ist es noch sinnvoller, nur ein Haupteffekte-Modell zu rechnen, d.h. eine Analyse ohne Interaktionen zu rechnen.

Ist bei einer nominalen Variablen eine Ausprägung leer, dann wird man diese sehr einfach mit einer anderen Ausprägung zusammenfassen. Beispiel: Die Variable V1 besitzt 3 Ausprägungen. Die Ausprägung 1 ist leer. Dann schreibt man folgende Umkodierungs-Anweisung:

V1 (1,2=1; 3=2)

P20.14.3 Zelleneffekte

Betrachten wir ein Beispiel mit 3 nominalen Variablen A,B,C. Wir verzichten auf die 2er- und 3er-Interaktionen, die sich aus A,B,C kombinieren lassen und bilden statt dessen eine "Zellenvariable" ABC, die die gemeinsame Wirkung aller 2er- und 3er-Interaktionen auf die abhängige Variable in sich vereinigt.

Die ALMO-Syntax-Anweisungen für ein solches Modell lauten:

```
Vereinbare
Variable = 99;
Anfang
Name1=A; N2=B; N3=C; N4=Y;
N99=ABC;

Programm=20;
UnominaleV = A, B, C / ABC;
Untergrenze A,B,C,ABC = 4*1;
Obergrenze A,B,C,ABC = 2,2,3,12;
Matrix      = Quadratsumme;
Verzichte  = Zellen, Effekte,
            Vektorausgabe,
            Matrixausgabe;
Ende_Programmparameter;

Lese A:Y aus Eingabe
      Format fix Leer_zu Ende;
      Feld 4*1;

ABC = A mit B mit C;

GP
Gehe_zu Lese
```

Die Obergrenze für die Zellenvariable ergibt sich aus $2*2*3 = 12$

Die Zellenvariable ABC wird durch die MIT-Operation aus A,B,C gebildet.

Ende	Daten
1117	
...	

Mit den Maskenprogrammen kann dieses Programm nur sehr umständlich nachgebildet werden. Dieses Modell liefert dieselbe gesamte erklärte Streuung, bzw. Fehlerstreuung, wie das vollständige Modell, das alle Interaktionen mit einschließt.

Der Schrägstrich in "A,B,C / ABC" kann weg gelassen werden. Siehe hierzu die Anmerkung "Schrägstrich" in P20.14.

Beachte: Die Zellenvariable ABC wird durch die MIT-Operation (und nicht die MAL-Operation) in der Leseschleife gebildet. Die Anweisung INTERAKTIONEN=ABC IST A MAL B MAL C ist nicht verwendbar. Demzufolge muss im Programmparameter-Block die Unter- und Obergrenze für ABC angegeben werden.

P20.15 Die Verwendung der PARTIAL-Anweisung

In Handbuch, Teil 3, Abschnitt P19.3 haben wir gezeigt, wie mit der PARTIAL-Anweisung eine Matrix partieller Streuungen gebildet werden kann. Diese PARTIAL-Anweisung kann auch im Rahmen von Programm 20 verwendet werden.

Betrachten wir ein Beispiel:

Die abhängige quantitative Variable ist V5.

Variablen V1 bis V3 und V9 sind die unabhängigen quantitativen Variablen

Beim Maskenprogramm Prog20mo wird die Partial-Anweisung in die Box "Option: Programm-Optionen lt. Handbuch" geschrieben.



Optionsbox geöffnet:



Hier soll eine Regressionsanalyse gerechnet werden. Der ALMO-interne Ablauf des Kalküls ist folgender:

1. Die Korrelationsmatrix der Variablen V1,2,3,9,5 wird gebildet.
2. Die Variablen V1 bis 3 werden aus V9 auspartielliert. Es entsteht die entsprechende Matrix partieller Korrelationen. V1,2,3 korrelieren mit V9 nunmehr mit 0.
3. Auf diese Matrix wird der übliche Regressionskalkül angewendet.

Die Regressionsgleichung, deren Koeffizienten errechnet werden, lautet:

$$y = k + \beta_{1y.23} V1 + \beta_{2y.13} V2 + \beta_{3y.12} V3 + \beta_{9y.123} V9.123$$

Beachte, dass V9 als Partialvariable $V_{9.123}$ in die Gleichung eingeht und dass V9 keinen Einfluss auf die anderen Variablen V1:3 besitzt.

Prinzip: Durch die PARTIAL-Anweisung wird also die gegenseitige Auspartiellierung, die in Programm 20 die Voreinstellung ist, aufgehoben. Die PARTIAL-Anweisung bewirkt eine einseitige (hierarchische) Auspartiellierung. Die Variablen vor dem Schlüsselwort AUS werden aus den Variablen nach AUS auspartielliert - aber nicht umgekehrt.

Dieselbe Wirkung - ohne PARTIAL-Anweisung - würden wir durch einen Schrägstrich, also eine Hierarchisierung erreichen: Beim Maskenprogramm Prog20mx oder Prog20mo wird in der Box "Analyse-Variable: Unabhängige Variable" im Eingabefeld für die quantitativen Variablen geschrieben:



Betrachten wir ein Beispiel mit nominalen Variablen:

```
Anfang
Name1=A; N2=B; N3=C; N4=AB; N5=AC;
N6=BC; N7=ABC; N8=Y;

Programm=20;
U_Nominale_V = A,B, C, AB, AC, BC, ABC;
UG A,B,C      = 3*1;
OG A,B,C      = 2,2,3;

Interaktionen = AB ist A mal B,
                AC ist A mal C,
                BC ist B mal C,
                ABC ist A mal B mal C;

Partial =      A,B,C aus AB /
                A,B,C aus AC /
                A,B,C aus BC /
                A,B,C,AB,AC,BC aus ABC;

A_aquantitative_V = y;

Ende_Programmparameter;
```

Beim Maskenprogramm Prog20mo müssen zuerst die Namen der Variablen und der Interaktionsvariablen angegeben werden. Das geschieht in der Box "Freie Namensfelder".



Dann müssen die Interaktionen in der Box "Analyse-Variable: Unabhängige Variable" definiert werden:



Dann muss die Partial-Anweisung in die Box "Option: Programm-Optionen lt. Handbuch" geschrieben werden:



Optionsbox geöffnet:



Sehr lange Partial-Anweisungen können auch über beide vorhandenen Eingabefelder geschrieben werden.

Dieselbe Wirkung - ohne PARTIAL-Anweisung - würden wir durch Hierarchisierung erreichen:

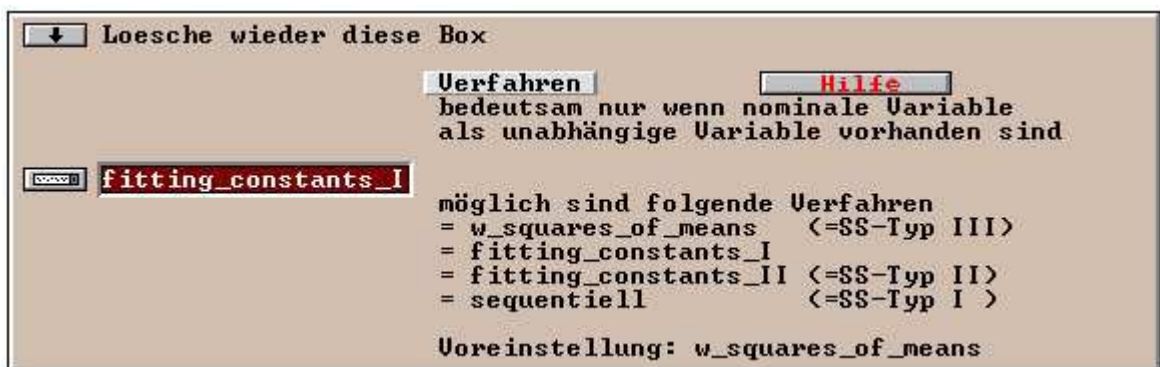
Beim Maskenprogramm Prog20mx oder Prog20mo wird in der Box "Analyse-Variable: Unabhängige Variable" im Eingabefeld für die nominalen Variablen geschrieben:



Dieselbe Wirkung würde erreicht, wenn die Box so ausgefüllt würde:



In der Optionsbox "Verfahren" muss dann noch auf "fitting_constants I" gesetzt werden.



BEACHTTE: Wenn 3 Variable und ihre Interaktionen nach dem Verfahren der "fitting_constants I" analysiert werden, dann treten "wechselnde Werte" bei den Interaktionseffekten auf. Die Analyse ist in Bezug auf die Interaktionseffekte nicht korrekt.

P20.15.1 Die einseitige Anpassung von Kovarianten und nominalen Variablen in der Kovarianzanalyse

Siehe hierzu auch Teil I, Abschnitt P20.7.1.1. Dort haben wir diese Konstellation "das unvollständige Modell" genannt.

Betrachten wir ein weiteres Beispiel.

Beim in Almo-Syntax "selbst geschriebenen" Programm wird folgender Programmparameter-Block geschrieben:

```

Programm=20
Name1=A; Name2=B; Name3=AB; Name10=X1;
Name11=X2; Name12=Y;

U_Nominale_V      = A,B/AB;
Untergrenze A,B  = 2*1;
Obergrenze A,B   = 2,3;
Interaktionen    = AB ist A mal B;

U_Quantitative_V = X1,X2;

Partial          = A,B,AB aus X1,X2;

A_Quantitative_V = Y
Ende_Programmparameter;

```

Beim Maskenprogramm **Prog20mo** müssen zwei Eingabeboxen in folgender Weise ausgefüllt werden

Analyse-Variable: Unabhängige Variable Hilfe

nominale unabhängige Variable Hilfe

A,B / AB

AB ist A mal B

Interaktionen x. Ordnung zwischen den nominalen unabhängigen Variablen bilden
oder einige ausgewählte Interaktionen bilden
0 =keine Interaktionen bilden Hilfe

paarweise Vergleiche (Kontraste) für die nominalen unabhängige Variablen rechnen

quantitative unabhängige Variable Hilfe

X1,X2

ordinale unabhängige Variable Hilfe

Die Optionsbox "Programm-Optionen lt. Handbuch" muss geöffnet werden und so ausgefüllt werden:

Hilfe

Loesche wieder diese Box (dann Voreinstellungen wieder gueltig)

Option: Programm-Optionen lt. Handbuch Hilfe

Partial = A,B,AB aus X1,X2;

Anmerkung:

Wird bei den nominalen Variablen A, B/AB der Schrägstrich durch einen Beistrich ersetzt, dann entsteht ein Lösung nach dem Verfahren der "weighted squares of means" – das in der Regel vorziehen ist. Mit Schrägstrich entsteht eine Lösung nach dem Verfahren der "fitting constants I".

Durch die PARTIAL-Anweisung werden zuerst A,B,AB aus den Kovarianten X1,X2 auspartiiert. Danach sind die beiden Variablenmengen orthogonal (unkorreliert). Wir führen also letzten Endes eine Kovarianzanalyse durch, bei der die Kovarianten an die

nominalen Variablen und ihre Interaktion angepasst sind, aber nicht umgekehrt. Die nominalen Variablen sind also nicht "kovarianzadjustiert". Sie erbringen Koeffizienten (Effekte, erklärte Streuungen) wie in einer Varianzanalyse ohne Kovariate. Wenn wir schreiben:

PARTIAL=X1, X2 AUS A, B, AB;

dann werden zuerst X1, X2 aus A,B,AB auspartielliert. Letzten Endes führen wir also eine Kovarianzanalyse durch, bei der die nominalen Variablen an die Kovarianten angepasst werden. Die Kovarianten erbringen jedoch Koeffizienten (Regressionskoeffizienten, erklärte Streuungen) wie in einer Regressionsanalyse ohne nominale Variable.

P20.16 Unvollständige, geschachtelte und hierarchische Versuchspläne

P20.16.1 Unvollständige Versuchspläne

In der experimentellen Forschung werden die unterschiedlichsten Formen unvollständiger Versuchspläne verwendet.

In der Literatur werden u.a. folgende unvollständige Versuchspläne unterschieden.

1. Unvollständige, ausgewogene und teilweise ausgewogene Block-Pläne
2. Lateinisches Quadrat (siehe Winer 1971, S. 685)
3. Griechisch-lateinisches Quadrat (siehe Winer, 1970, S. 709)
4. "confounding designs" (siehe Winer, 1971, S. 604).

Alle diese Versuchspläne werden von ALMO begriffen als Versuchspläne mit einigen leeren Zellen.

ALMO eliminiert jene Dummy-Variable, die lineare Abhängigkeiten verursachen. In der Regel sind das die Interaktionsvariablen. Bei "ausgewogenen" unvollständigen Versuchsplänen, wie etwa dem lateinischen Quadrat, kommt es zu keiner Eliminierung von Variablen, wenn der Benutzer auf Interaktionen verzichtet.

Betrachten wir ein Beispiel für einen unvollständigen, ausgewogenen Versuchsplan:

		Faktor A			
		A1	A2	A3	A4
Faktor B	B ₁				0
	B ₂			0	
	B ₃		0		
	B ₄	0			

Die mit 0 gekennzeichneten Zellen sind leer. Trotzdem entstehen keine linearen Abhängigkeiten bei den Dummy Variablen, sofern auf Interaktionen verzichtet wird. Man sagt, der Versuchsplan sei "ausgewogen". ALMO eliminiert deswegen auch keine Variablen.

Wird jedoch vom Benutzer "INTERAKTIONEN=2;" angegeben, dann eliminiert ALMO automatisch diejenigen Interaktionsdummies, die eine lineare Abhängigkeit verursachen.

P20.16.2 Geschachtelte Variable

Betrachten wir ein Beispiel:

			D1	D2
A1	B1	C1		
		C2		
	B2	C3		

		C4		
A2	B1	C5		
		C6		
	B2	C7		
C8				
C9				

Wir haben eine 4-faktorielle Versuchsanordnung - mit der Besonderheit, dass C mit A und B nicht voll "gekreuzt" ist. Die Ausprägungen C1, C2 beispielsweise stehen unter A1B1 und die Ausprägungen C7, C8, C9 stehen unter A2B2. C ist unter A und B "geschachtelt". C ist eine "geschachtelte" Variable. In ALMO können wir derartige Variable durch die PARTIAL-Anweisung behandeln. Der Programmparameter-Block des Syntax-Programms lautet

```
P=20
N1=A; N2=B; N3=C; N4=D; N12=AB;

U_Nominale_V          = A,B,C,D /AB;
Untergrenze A,B,C,D  = 1,1,1,1;
Obergrenze  A,B,C,D  = 2,2,9,2;
Interaktionen        = AB ist A mal B,
Partial              = A,B,AB aus C;
.
.
.
Ende_Programmparameter;
```

Anmerkung:

Wird in U_nominale_V der Schrägstrich durch einen Beistrich ersetzt, dann entsteht eine Lösung nach dem Verfahren der "weighted squares of means" – das in der Regel vorziehen ist. Mit Schrägstrich entsteht eine Lösung nach dem Verfahren der "fitting constants I".

Regel: Aus der geschachtelten Variablen werden die vor ihr stehenden Variable und deren Interaktionen auspartielliert. Steht nur eine Variable vor der geschachtelten Variablen, so wird nur diese auspartielliert.

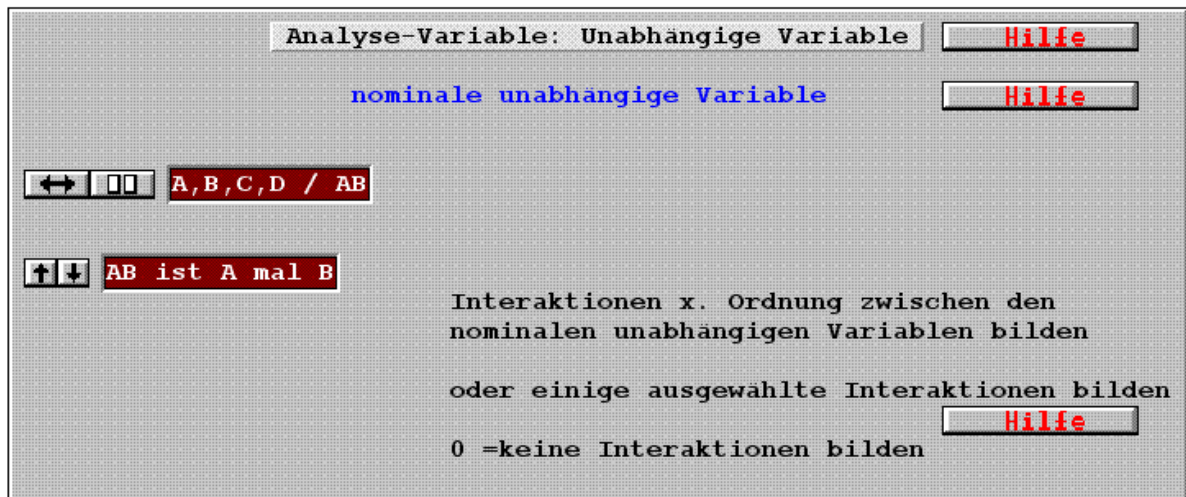
Bei gleichen Zellenhäufigkeiten könnten wir - ohne Verwendung der PARTIAL-Anweisung - auch schreiben:

$$U_Nominale_V = A,B / AB / C,D; \quad (\text{Lösung nach "fitting constants I"})$$

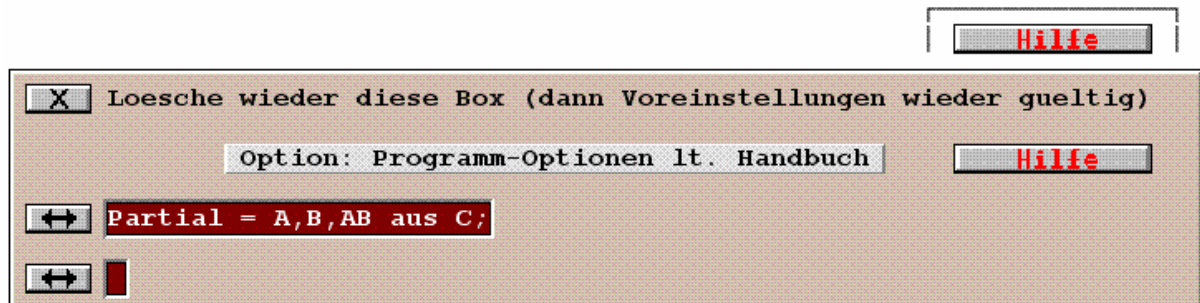
oder

$$U_Nominale_V = A,B,AB / C,D; \quad (\text{Lösung nach "weighted squares of means"})$$

Im Maskenprogramm Prog20mo wird geschrieben



In der Box "Option: Programm-Optionen lt. Handbuch" wird geschrieben



Entscheidend dafür, wie wir "geschachtelte Versuchspläne" analysieren, ist, ob wir die geschachtelte Variable als fixen Faktor oder als Zufallsvariable behandeln. Siehe dazu Winer,1971,S. 313,361.

In den meisten Fällen handelt es sich bei den geschachtelten Variablen um die Untersuchungseinheiten (die explizit als Variable in den Versuchsplan aufgenommen werden). Diese Untersuchungseinheiten werden als Zufallsauswahl aus einer größeren Grundgesamtheit betrachtet.

Die geschachtelte Variable der Untersuchungseinheiten wird deswegen als Zufallsvariable begriffen. Derartige Versuchspläne werden in der Literatur üblicherweise als "Versuchspläne mit Messwiederholungen" bezeichnet. Wir beschreiben diesen Typus in Abschnitt P20.17.2 ff.

Unsere oben beschriebene Methode gilt nur für Versuchspläne, bei denen die geschachtelte Variable ein fixer Faktor ist.

P20.16.3 Hierarchische Versuchspläne

Betrachten wir ein Beispiel:

Y

A1	B1	C1	
		C2	
	B2	C3	
		C4	
	B3	C5	
		C6	
		C7	
A2	B4	C8	
		C9	
		C10	
	B5	C11	
		C12	

Die 2. Variable B ist unter A geschachtelt. Die 3. Variable C ist unter A und B geschachtelt. Allgemein: Jede Variable außer der ersten ist unter den ihr vorausgehenden Variablen geschachtelt.

Diese Versuchs-Anordnung analysieren wir dadurch, dass wir als unabhängige nominale Variable angeben

A/B/C

Alle Variable werden also durch Schrägstrich voneinander getrennt.

Oder ohne Schrägstriche durch

A, B, C

In der Box "Option: Verfahren" wird auf "sequentiell" eingestellt.

Unsere Aussage gilt nur für den Fall, dass die geschachtelten Variablen "fixe" Faktoren sind. Siehe Abschnitt P20.16.2.

P20.17 Analysen mit wiederholten Messungen

Im folgenden wird nur ein einfaches Messwiederholungs-Design dargestellt. Komplexe Designs werden im Handbuch "P20: Allgemeines Lineares Modell" ausgeführt.

Betrachten wir ein Beispiel, das wir aus Winer (1971, S.268) entnehmen:

Die Reaktionszeit von 5 Autofahrern wird nach der Einnahme von 4 verschiedenen Drogen gemessen. Die Frage lautet: Haben die Drogen einen Einfluss auf die Reaktionszeit? Und wenn ja, unterscheiden sie sich in ihrer Einflusstärke?

Wir besitzen in diesem Beispiel also 4 wiederholte Messungen an derselben Variablen, der Reaktionszeit. Siehe nachfolgende Tabelle. Wir führen in diesem Falle eine Varianzanalyse durch, bei der wir die Reaktionszeit als abhängige, quantitative Variable verwenden. Die unabhängige nominale Variable ist die "Drogeneinnahme" (mit den Ausprägungen: 1. Droge, 2. Droge, 4. Droge). Da wir unterstellen müssen, dass auch die Versuchspersonen unterschiedlich auf die Drogen reagieren, müssen wir auch die "Versuchsperson" (mit den Ausprägungen: 1. Versuchsperson, ... 5. Versuchsperson) als unabhängige nominale Variable einführen.

Dies ist das Charakteristikum der Analysen mit wiederholten Messungen: Wir können die Streuung, die durch die unterschiedlichen Messungen (nach Einnahme unterschiedlicher Drogen) erklärt wird, trennen von der Streuung, die durch die Unterschiede zwischen den Versuchspersonen erklärt wird.

P20.17.0 Uni- und multivariater Ansatz

Beim univariaten Ansatz, den wir in Abschnitt P20.17.1 ff. darstellen, werden die Untersuchungseinheiten, an denen die Messwiederholungen vorgenommen wurden (in unserem Beispiel: die Autofahrer) explizit als Faktor in die Varianzanalyse eingeführt. Beim multivariaten Ansatz ist dies implizit der Fall. Hier werden die Messwiederholungen, genauer: die Differenzen zwischen ihnen als abhängige Variable einer multivariaten Varianzanalyse betrachtet.

Wenn alle Verteilungsannahmen erfüllt sind, dann ist der univariate Ansatz der statistisch mächtigere. Diese Verteilungsannahmen werden im Begriff der "Sphärizität" zusammengefasst. Einige dieser Annahmen der Sphärizität sind: (1) Die Varianzen der Messwiederholungen sind gleich, (2) die Korrelationen zwischen den Paaren, die sich aus je zwei Messwiederholungen bilden lassen, sind gleich.

Es darf unterstellt werden, dass diese Verteilungsannahmen nur in Ausnahmefällen erfüllt sind. Für den multivariaten Ansatz sind diese Annahmen nicht notwendig.

Beim univariaten Ansatz werden die Untersuchungseinheiten als Ausprägungen einer nominalen Variablen betrachtet, so dass sehr schnell große Matrizen entstehen, die auch einen Rechner mit sehr viel Speicher überfordern. Beim multivariaten Ansatz ist dies nicht der Fall. Uni- und multivariater Ansatz erbringen unterschiedliche Ergebnisse, d.h. Almo liefert als Signifikanz für die jeweiligen unabhängigen Variablen verschiedene Werte. Nur wenn 2 Messwiederholungen vorliegen, wie in unserem unmittelbar nachfolgendem Beispiel, dann ist das Ergebnis dasselbe.

P20.17.0.1 Der multivariate Ansatz

Beim multivariaten Ansatz werden die Differenzen zwischen den wiederholten Messungen als abhängige quantitative Variable betrachtet. Siehe dazu Brien/Kaiser (1985).

Betrachten wir folgendes einfache Beispiel:

Person	Droge 1	Droge 2
1	50	40
2	47	53
3	39	48
.		
.		

Person 1 hat nach Einnahme von Droge 1 eine Reaktionszeit von 50 hundertstel Sekunden, nach Droge 2 von 40 hundertstel Sekunden. Die Differenz zwischen den beiden Drogen ist 10 hundertstel Sekunden. Wir könnten aus den Differenzen eine neue Datenmatrix bilden.

Person	Reaktionsdifferenz Droge1 - Droge 2
1	10
2	6
.	
.	

Mit diesen Daten rechnen wir eine normale einfache Varianzanalyse mit der Droge als unabhängiger nominalen Variablen und der Reaktionsdifferenz als abhängiger quantitativen Variablen. Dadurch testen wir die Signifikanz des Unterschieds zwischen den beiden Drogen.

Da wir hier nur eine abhängige Variable haben, liegt im Prinzip keine multivariate Analyse vor. Dies ist erst dann der Fall, wenn wir die Differenzen zwischen mindestens 3 Messwiederholungen als abhängige Variable einsetzen. Trotzdem kann diese Datenkonstellation mit nachfolgender Programm-Maske gerechnet werden, die den multivariaten Ansatz verwendet.

Wir wollen unser Beispiel nun dadurch etwas komplizieren, dass wir 4 Messwiederholungen vornehmen. Das zu analysierende Design hat also folgende Gestalt (wir übernehmen unser Beispiel aus Winer, 1971, S.268).

Subjekt	Droge 1	Droge 2	Droge 3	Droge 4
1	30	28	16	34
-	-	-	-	-
-	-	-	-	-

Das Maskenprogramm

Das nachfolgende Maskenprogramm verwendet ein Beispiel aus Winer (1971) mit 4 Messwiederholungen. Es ist aber selbstverständlich für beliebig viele Messwiederholungen ausgelegt.

Prog20M2
Messwiederholungsdesign Kurzprogramm
 als multivariate Varianzanalyse
 Analysiert wird ein Design mit 4 Messwiederholungen. Die Reaktionszeit von Autofahrern nach der Einnahme von 4 verschiedenen Drogen wird gemessen.
 Beispiel aus Winer: Statistical Principles in experimental research, 1971, S. 268

Das Design hat folgende Gestalt:

	Droge1	Droge2	Droge3	Droge4
Subjekt	Y1	Y2	Y3	Y4
1	30	28	16	34
2	14	18	10	22
3	24	20	18	30
4	38	34	20	44
5	26	28	14	30

Y1 ist die Reaktionszeit des Autofahrers nach Einnahme von Droge 1 (in 1/100 Sekunden)
 Y2 ist die Reaktionszeit des Autofahrers nach Einnahme von Droge 2 etc.

Die Zahl der Subjekte und der Auspraegungen der Variablen ist beliebig

Die Datenmatrix ist folgende:

	U1	U2	U3	U4	U5
1	30	28	16	34	
2	14	18	10	22	
3	24	20	18	30	
4	38	34	20	44	
5	26	28	14	30	

Was ist ein Kurzprogramm ? -->
 Bedienung -->

- 1
 Vereinbare Variable= 20 ;
- 2 Option: Weitere Vereinbarungen - nur wenn Also dazu auffordert
- 3
 Dateiname
zeige=Namensdatei in Output zeigen
leer =nicht
- 4
 Freie Namensfelder

5 bei Datei-Problemen

 Format der Daten
 der Datensatz enthält diese Variablen
Bei Format DIREKT schreiben Sie: alle_U

höchste Variablen-Nummer des Datensatzes

6 Wenn Dateiformat FIX oder Nicht-Standard-FREI

7 zu analysierende Variable
Variablen-Nummer der 1.Messung
Variablen-Nummer der letzten.Messung

8 Option: Ein- und Ausschliessen von Untersuchungseinheiten

9 Option: Umkodierungen und Kein-Wert-Angaben

10

Erläuterung zu den Boxen

Box „Datei aus der gelesen wird“

Siehe P0.3.

The dialog box is titled "Datei aus der gelesen wird". It contains the following elements:
- A file path field: "C:\Almo7\Testdat\Win268.fre"
- A format field: "frei"
- A variable range field: "U1:5"
- A "Hilfe" button next to the file path field with the text "bei Datei-Problemen" below it.
- A "Hilfe" button next to the format field.
- A "Hilfe" button next to the variable range field.
- A "Hilfe" button at the bottom right.
- Text: "Format der Daten" and "der Datensatz enthält diese Variablen Bei Format DIREKT schreiben Sie: alle_U".

Eingabefeld 4: Geben Sie hier die höchste verwendete Variablennummer an. In unserem Beispiel wird ein Datensatz mit den Variablen V1 bis V5 eingelesen. Die höchste verwendete Variablennummer ist also 5.

Box „Zu analysierende Variable“

The dialog box is titled "zu analysierende Variable". It contains the following elements:
- Two input fields for variable numbers: "2" and "5".
- Text: "Variablen-Nummer der 1.Messung" and "Variablen-Nummer der letzten.Messung".

Eingabefeld 1 und 2: Almo unterstellt, dass die Messwiederholungsvariablen fortlaufend hintereinander stehen. In unserem Beispiel sind sie V2, V3, V4, V5. Die Variablennummer der 1. Messung ist also 2, die der letzten ist 5.

Ist dies in den vorhandenen Daten nicht der Fall, dann ist folgendermaßen vorzugehen.

Beispiel: Die Messwiederholungen sind in den Variablen 2,4,6,8 enthalten. In der Umkodierungsbox schreiben wir

```
V20 = V2;  
V21 = V4;  
V22 = V6;  
V23 = V8;
```

Dabei müssen die Variablen V20,21,22,23 frei sein. In obiger Box wird dann angegeben

```
Variablen_Nummer der 1. Messung: 20  
Variablen_Nummer der letzten Messung: 23
```

Programm-intern bildet Almo nun die Differenzen zwischen den Reaktionszeiten in folgender Weise:

```
Diff1 = Reaktion1 - Reaktion4  
Diff2 = Reaktion2 - Reaktion4  
Diff3 = Reaktion3 - Reaktion4
```

Das Prinzip ist offenkundig folgendes: Die Differenzvariablen entstehen dadurch, dass die letzte abhängige Variable von den vor ihr stehenden subtrahiert wird. Jede andere Differenz ist redundant und darf deswegen (zur Vermeidung linearer Abhängigkeiten) nicht in das

Modell aufgenommen werden. Man könnte auch die erste oder eine andere beliebige abhängige Variable von den jeweils anderen subtrahieren.

Die im Rahmen des Kalküls entstehenden Streuungen können sehr große Zahlenwerte annehmen. Um diese um einen konstanten Faktor zu verringern, empfiehlt es sich, die Differenzvariable Diff1 bis Diff3 durch einen konstanten Wert zu dividieren – bewährt hat sich

$$\sqrt{2}$$

als Divisor.

Almo liefert folgendes Ergebnis (gekürzt):

```

generalisierte Gesamtstreuung      27026.000000
=====
Koeffizienten fuer Gesamt-Modell

Durch alle unabh. Variable
erklarte generalisierte Streuung    26406.400000
generalisierte Fehlerstreuung      619.600000
multipler Korrelat.koeff.          0.988470

Wilks Lambda                        0.022926
F-Wert f. erklarte Streuung         28.412309
Freiheitsgrade Nenner =            3
                                   Zaehler=  2
Signifikanz: p                      0.032859
Signifikanz: (1-p)*100              96.714091 %
Teststaerke von F                   0.773277
=====

```

Die Drogen haben mit 96.714% einen signifikant verschiedenen Einfluss auf die Reaktionszeit.

Da wir in unserer Analyse nur eine unabhängige Variable haben, den Drogeneffekt, ist die "durch alle unabh. Variable erklärte generalisierte Streuung" gleich der des Drogeneffekts. Die wesentlichen Koeffizienten sind das Wilks'sche Lambda, dessen F-Transformation und die Signifikanz.

Interessant ist der Vergleich mit den Ergebnissen aus dem univariaten Ansatz. Hier entsteht ein F-Wert von 24.76 (df1=3, df2=12) mit einer (höheren) Signifikanz von 99.99 %

P20.17.1 Univariater Ansatz

Wir wollen uns nun in den folgenden Abschnitten mit dem univariaten Ansatz der Analyse von Messwiederholungsdesigns beschäftigen.

Wir wollen im Folgenden das Beispiel von Winer mit ALMO durchrechnen: Die Untersuchungsdaten sind bei ihm in folgender Tabelle enthalten:

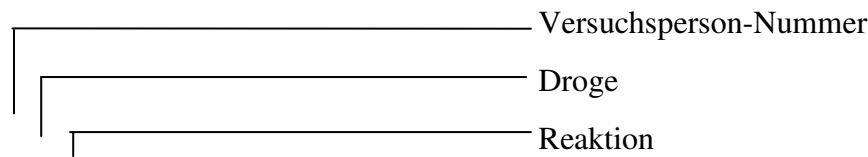
Versuchsperson	Droge 1	Droge 2	Droge 3	Droge 4
1	30	28	16	34
2	14	18	10	22
3	24	20	18	30
4	38	34	20	44
5	26	28	14	30

Betrachten wir die Zelle oben links: Nach Einnahme von Droge 1 hat Person 1 eine Reaktionsgeschwindigkeit von 30 Hundertstelsekunden. Entsprechend sind die anderen Zellen zu interpretieren.

Um diese Daten für den univariaten Ansatz in ALMO einzugeben, müssen sie die Form einer Datenmatrix (d.h. aufeinander folgender Datenvektoren) haben. Hier gibt es 2 Möglichkeiten:

1. Möglichkeit: Die Untersuchungseinheit ist die einzelne Messung. Der Datenvektor besteht aus den 3 Variablenwerten:

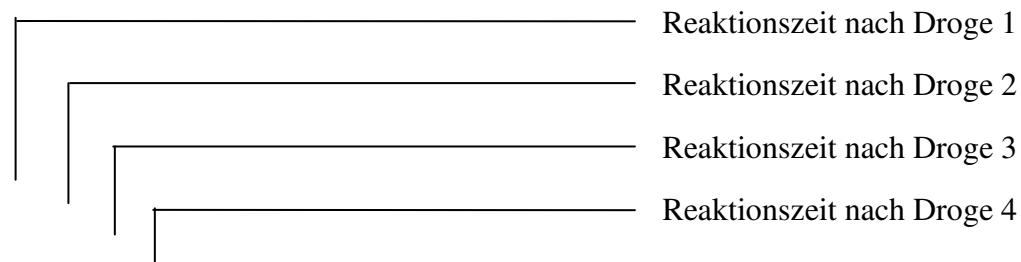
(1) Versuchsperson-Nummer, (2) Droge, (3) Reaktionszeit.



1 1 30
 1 2 28
 ...
 5 3 14
 5 4 30

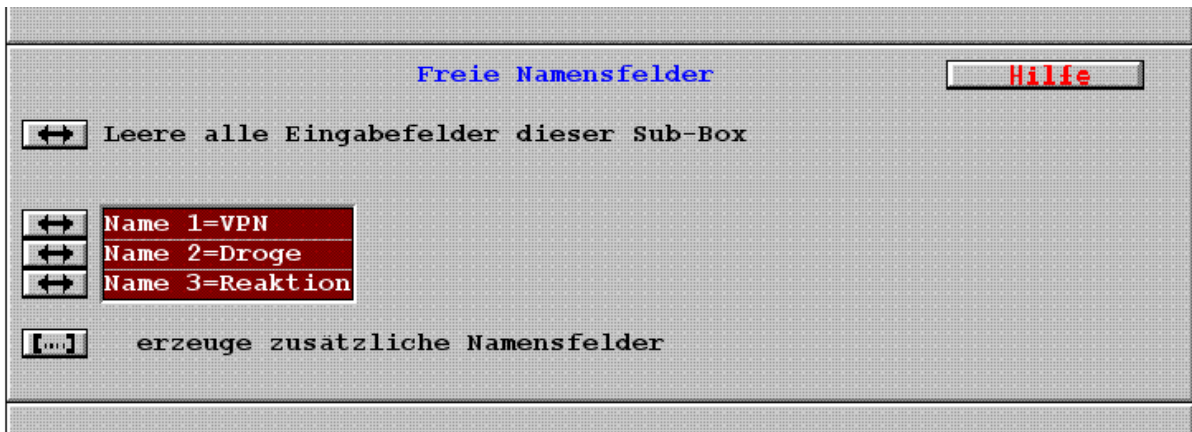
Für jede der 5 Versuchsperson gibt es 4 Daten-Zeilen (für die 4 Drogen). Die Datenmatrix umfasst also dann $5 \cdot 4 = 20$ Datenvektoren.

2. Möglichkeit: Die Untersuchungseinheit ist die Versuchsperson. Der Datenvektor besteht aus den 4 Reaktionszeiten einer Versuchsperson. Die Datenmatrix unserer 5 Probanden umfasst dann 5 Datenvektoren.

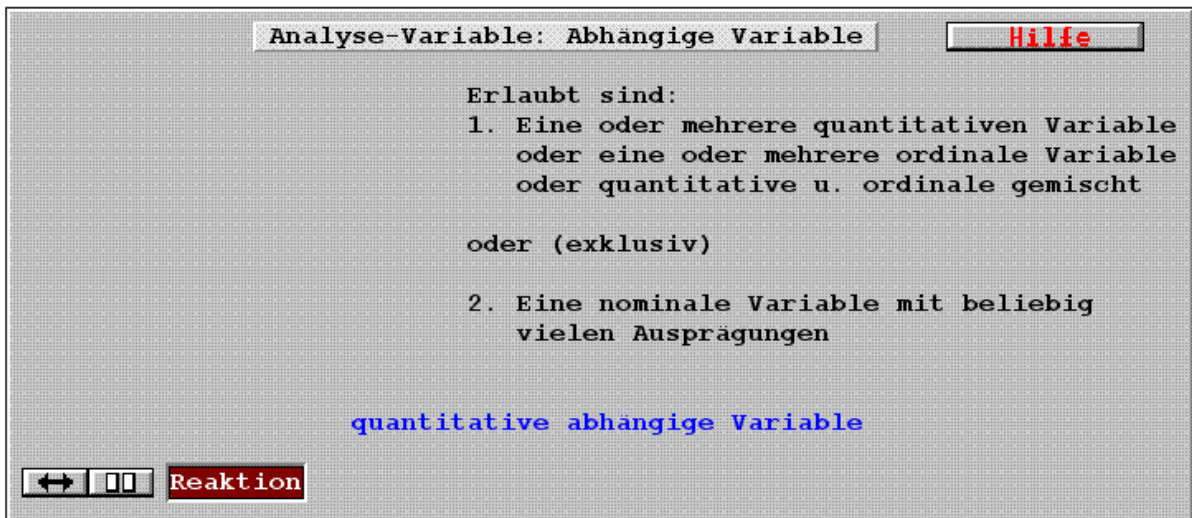


30 28 16 34
 14 18 10 22
 ...
 26 28 14 30

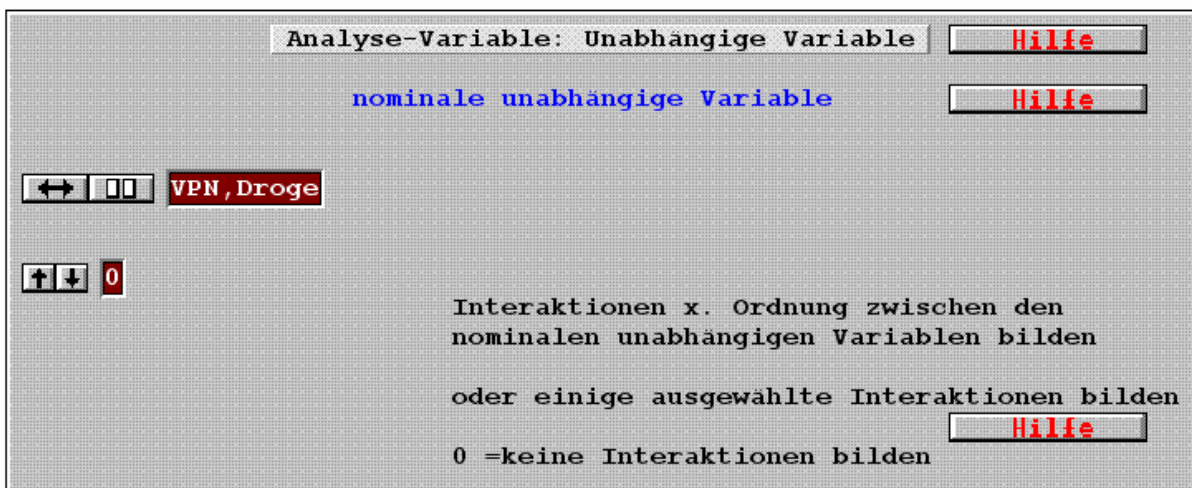
Liegen die **Daten in der 1. Form** vor, dann kann die Programmmaske Prog20mo (oder Prog20mx) verwendet werden. Folgende Variablennamen werden vergeben



Die Box für die abhängige quantitative Variable ist:



Und für die unabhängigen nominalen Variablen ist:



Normalerweise werden sich die **Daten in der 2. Form** befinden. D.h. der Datenvektor bezieht sich auf eine Versuchsperson und enthält die Messungen, die an der Versuchsperson durchgeführt wurden. Wir wollen nun überlegen, wie wir ein ALMO-Programm für diesen

Normalfall zu schreiben haben. Die Programmmasken Prog20mo und Prog20mx sind nicht verwendbar. Es muss ein etwas komplexes Almo-Syntax-Programm geschrieben werden

Wir schaffen 3 neue Variable, denen wir je einen Namen zuweisen.

	<u>Variable</u>	<u>Namen</u>
für die Person-Nummer	V10	Subjekt
für die Droge	V11	DROGE
für die Reaktionszeit	V12	REAKTION

Die Variablennummern 10,11,12 müssen selbstverständlich frei sein.

Das nachfolgende Programm ist als Beispielprogramm „Win268_U.Alm“ in Almo enthalten. Sie finden es durch Klick auf das Menü „Almo/Liste aller Almo-Programme“.

#

Messwiederholungsdesign
als univariate Varianzanalyse
Analysiert wird ein Design mit 4 Messwiederholungen. Die Reaktionszeit
von Autofahrern nach der Einnahme von 4 verschiedenen Drogen wird ge-
messen.
Beispiel aus Winer: Statistical Principles in experimental research,
1971, S. 268
Das Design hat folgende Gestalt:

Subjekt	Y1	Y2	Y3	Y4
1	30	28	16	34
2	14	18	10	22
3	24	20	18	30
4	38	34	20	44
5	26	28	14	30

```
VEREINBARE  
  VARIABLE=20;
```

```
ANFANG
```

```
Name10=Subjekt;  
Name11=Droge;  
Name12=Reaktion;
```

```
Programm=20;
```

```
  U_nominale_V      = Subjekt, Droge;
```

```
  Untergrenze Subjekt,Droge = 1,1;  
  Obergrenze  Subjekt,Droge = 5,4;
```

```
  A_quantitative_V  = Reaktion;
```

```
  Matrix            = Quadratsumme;
```

```
  Verzichte        = Matrixausgabe,  
                   Vektorausgabe;
```

```
ENDE_PROGRAMMPARAMETER
```

```
Lese V1:4  
  aus EINGABE          # Eingabe = die Daten stehen hinter dem Programm #  
  Format frei          # und werden gemeindam mit dem Prog eingelesen  #  
  leerzu ENDE;
```

```
Subjekt = DE;
```

```
Droge   = 1;  
Reaktion= V1;  
Gehe_in_Programm
```

```
Droge   = 2;  
Reaktion= V2;  
Gehe_in_Programm
```

```
Droge   = 3;  
Reaktion= V3;  
Gehe_in_Programm
```

```
Droge   = 4;  
Reaktion= V4;  
Gehe_in_Programm
```

```
Gehezu LESE
```

```

#-----#
#       Datenmatrix       #
#       #                 #
# Drog1 Drog2 Drog3 Drog4 #
#
ENDE

    30   28   16   34
    14   18   10   22
    24   20   18   30
    38   34   20   44
    26   28   14   30

```

Die Besonderheit an diesem Programm ist, dass die Person-Nummer als Variable verwendet wird und dass wir mit **einem** eingelesenen Datenvektor 4 mal "in das Programm gehen", da wir 4 Messungen für eine Versuchsperson besitzen.

Beachte: Da die Person-Nummer "Subjekt" eine unabhängige nominale Variable ist, können selbstverständlich nur Analysen gerechnet werden, bei denen die Zahl der Versuchspersonen nicht zu groß ist.

Für den Programmierspezialisten sei angemerkt, dass sich die Leseschleife unter Verwendung der SCHLEIFE-Anweisung sehr viel kürzer schreiben lässt:

```

Lese V1:4 aus Eingabe
  Format frei
  Leer_zu Ende;

Subjekt=DE;

Schleife von Droge = 1;
  Reaktion = V(Droge);
  Gehe_in_Programm
Bis Droge = 4;

Gehe_zu Lese

```

Innerhalb einer Schleife wird 4 mal "in das Programm" gegangen. Als Schleifenvariable wird die Drogennummer verwendet. Die Reaktionszeit erhält ihren Wert aus der indizierten Variablen V(DROGE). Siehe Abschnitt 38 im Handbuch "Teil 2: ALMO-Programmiersprache". Im 1. Schleifendurchlauf ist V(DROGE) gleich V(1). REAKTION ist also gleich V1. Im 4. Schleifendurchlauf ist V(DROGE) gleich V(4) gleich V4. REAKTION ist also gleich V4.

ALMO liefert für obiges Programm eine Ausgabe wie bei der Varianzanalyse (siehe Abschnitt P20.9.1).

Wir wollen nun einen etwas komplizierteren Fall betrachten. Die Daten seien folgende:

		Droge 1		Droge 2	
		Tag	Nacht	Tag	Nacht
Versuchsperson	1	30	36	16	20
	2	14	19	12	19

Zusätzlich zum Faktor "Droge" ist noch der Faktor des Einnahmezeitpunktes hinzugekommen. Die Datenmatrix muss, damit ALMO sie verarbeiten kann zu folgender Matrix transformiert werden.

VPN	Droge	Tag_Nacht	Reaktionszeit
1	1	1	30
1	1	2	36
1	2	1	16
1	2	2	20
.	.	.	.
.	.	.	.

So befinden sich die **Daten in der 1. Form.**

Die notwendige Daten-Transformation und die Analyse erreichen wir mit folgenden Almo-Programm:

```

Vereinbare
Variable = 10;
Anfang
Name5=Subjekt;
Name6=Droge;
Name7=TagNacht;
Name8=Reaktion;

Programm=20;
U_nominale_V      = Subjekt,Droge,TagNacht;
Untergrenze SUBJEKT,Droge,TagNacht = 1,1,1;
Obergrenze  SUBJEKT,Droge,TagNacht = 20,2,2;
A_quantitative_V = Reaktion;
Matrix       = Kovarianz;
Verzichte    = Matrixausgabe, Vektorausgabe,
              Zellen, Effekte;
Ende_Programmparameter;
Lese V1:4 aus Eingabe
  Format frei leerzu Ende;

SUBJEKT = DE;
H1=0;
Schleife von Droge = 1;
  Schleife von TagNacht = 1;
    H1 = H1 + 1;
    Reaktion = V(H1);
    Gehe_in_Programm
    Drucke SUBJEKT,Droge,TagNacht,Reaktion;
  Bis TagNacht = 2;
Bis Droge = 2;
Gehe_zu Lese
Ende
30 36 16 20
14 19 12 19
. . . .
. . . .

```

Daten stehen hinter Programm

neuer Datensatz zum Anschauen ausgeben

P20.17.1.1 Exkurs: Zuverlässigkeit und wiederholte Messungen:

Betrachten wir ein Beispiel:

Die Reaktionszeit von 5 Autofahrern wird durch ein bestimmtes Verfahren gemessen. Um die Zuverlässigkeit dieses Messverfahrens zu überprüfen, wird die Messung (nacheinander) 4 mal durchgeführt.

Wir führen in diesem Falle eine Varianzanalyse durch, bei der wir als abhängige quantitative Variable die Reaktionszeit verwenden. Als unabhängige nominale Variable verwenden wir den Messzeitpunkt (mit den Ausprägungen: 1. Messung, 2. Messung, ... 4. Messung) und die Versuchsperson (mit den Ausprägungen: 1. Versuchsperson, 2. Versuchsperson, ... 5. Versuchsperson).

Dieses Beispiel ist vollkommen strukturgleich zu dem Beispiel des vorausgegangenen Abschnitts P20.17.1, bei dem die Reaktionszeit von 5 Autofahrern 4 mal nach Einnahme einer jeweils anderen Droge gemessen wurde.

Wir schreiben also genau dasselbe ALMO-Programm wie in Abschnitt P20.17.1. Lediglich die Namensgebung ist zu ändern. Anstelle von "1. Droge, ... 4. Droge" heißt es jetzt 1. Messung, ... 4. Messung".

ALMO liefert die durch die Variable "Messzeitpunkt" (SSM) und die durch die Variable "Versuchsperson" (SSV) erklärte Streuung. Je größer SSM ist, umso stärker ist die Wirkung des Messzeitpunkts, umso weniger zuverlässig ist also die Messung.

Die Spearman-Brown-Formel der Zuverlässigkeit lautet (siehe Winer, 1971, S. 286f).

$$(1) p = \frac{\frac{SSV}{N-1} - \frac{SSM + SSE}{N * (k-1)}}{\frac{k * (SSM + SSE)}{N * (k-1)}} \quad (2) z = \frac{k * p}{1 + k * p}$$

- Z = Zuverlässigkeitskoeffizient
- SSV = die durch die Versuchspersonen erklärte Quadratsumme
- SSM = die durch die Variable "Messzeitpunkt" erklärte Quadratsumme
- SSE = Fehlerstreuung aus ALMO
- N = Zahl der Versuchspersonen
- k = Zahl der Messwiederholungen

P20.17.6 Aggregatdaten-Analyse (z.B. Wahlanalyse)

Betrachten wir ein Beispiel:

V1 Wahlbezirk	V2 abgegeben. Stimmen in 1000	V3 Gemeindetyp	V4 Geographische Lage	V5 Arbeiter Anteil
1	134	1 (städtisch)	1 (Nord)	0.70
2	8	2 (ländlich)	2 (Süd)	0.20
3	46	1 (städtisch)	3 (West)	0.40
4	96	1 (städtisch)	4(Ost)	0.30
.
.

V6 Anteil d. Jungwähler	V7 Anteil d. Pensionisten	V8 Partei A	V9 STIMMANTEIL Partei B	V10 Partei C
0.15	0.23	0.60	0.30	0.04
0.13	0.25	0.25	0.60	0.10
0.20	0.12	0.35	0.40	0.25
0.10	0.22	0.30	0.40	0.20
.
.
.

Die 1. Zeile dieser Tabelle ist in folgender Weise zu verstehen: Wahlbezirk 1 ist städtisch und liegt im Norden des Bundesgebietes. Der Arbeiteranteil an der wahlberechtigten Bevölkerung ist 70 %, der Anteil der Jungwähler 15 %.... etc. Partei A erhielt 60 % der abgegebenen Stimmen, Partei B 30 % und Partei C 4 %.

Die Frage lautet: In welcher Weise erklären Gemeindetyp, die geographische Lage, Arbeiteranteil, Jungwähler und Pensionistenanteil die Stimmverteilung auf die 3 Parteien A, B, C?

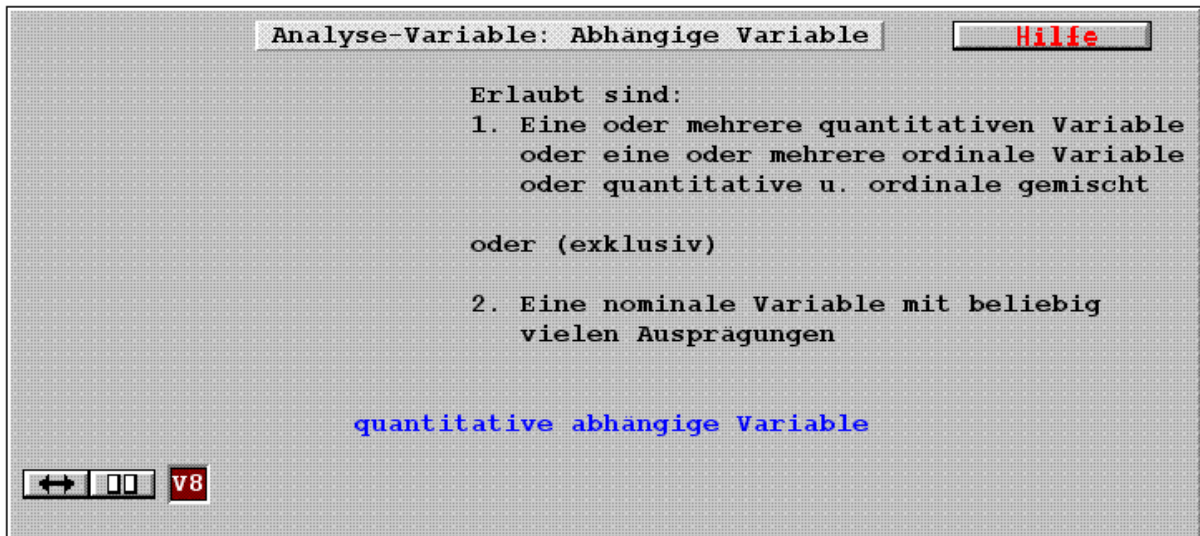
Wir haben also

- 1) 3 abhängige quantitative Variable:
Stimmanteil der Parteien A, B, C.
- 2) Zwei unabhängige nominale Variable;
 - a) Gemeindetyp: städtisch, ländlich, und
 - b) geographische Lage: Nord, Süd, West, Ost.
- 3) 3 Kovariate, also unabhängige quantitative Variable:
Arbeiteranteil, Jungwähleranteil, Pensionistenanteil.

Wir rechnen also 3 Kovarianzanalysen: für jede Partei eine.

Wir wollen ein ALMO-Programm für Partei A schreiben. Dazu kann die Programmaske Prog20mo oder Prog20mx verwendet werden.

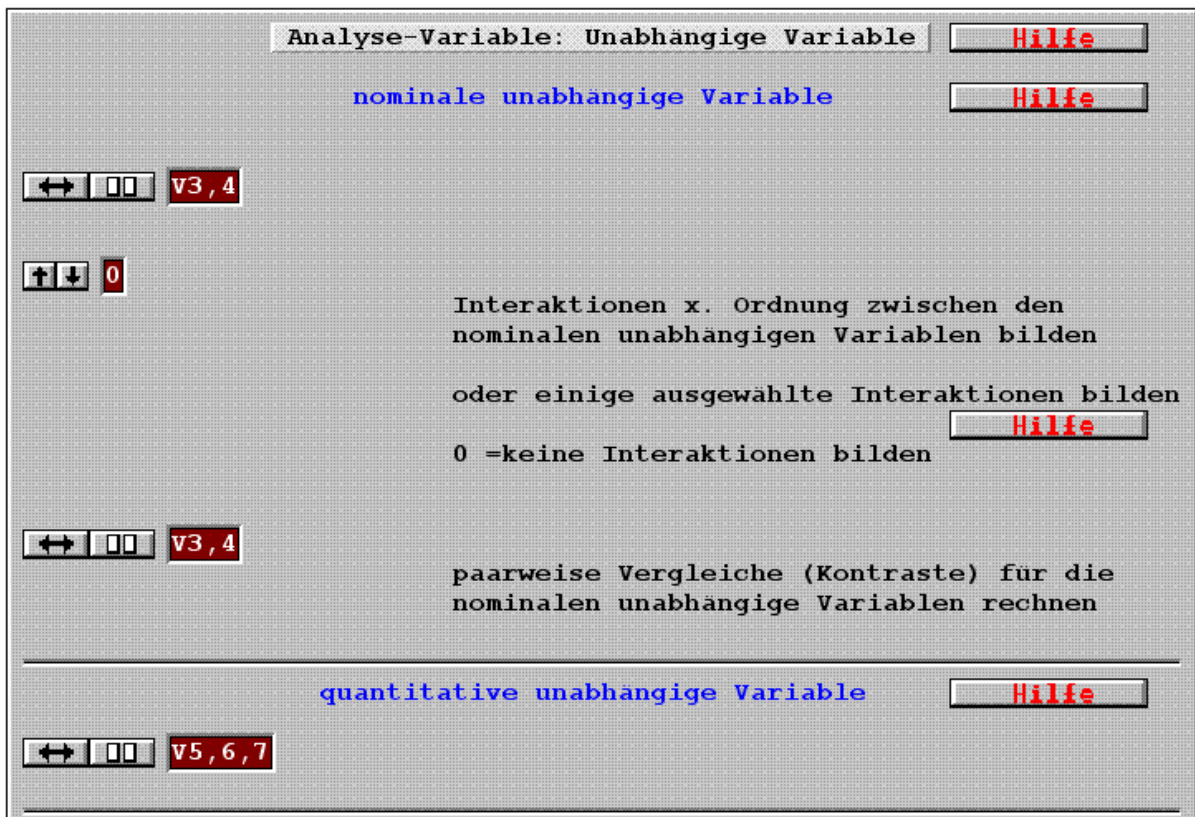
Die Box für die abhängige quantitative Variable V8 (Stimmanteil Partei A)



Box für die unabhängigen Variablen

Nominal: V3 (Gemeindetyp), V4 (Geografische Lage)

Quantitativ: V5 (Arbeiteranteil), V6 (Jungwähler), V7 (Pensionisten)



Es hat sich gezeigt, dass eine Gewichtung sinnvoll ist, da sonst jeder Bezirk unabhängig von seiner Größe mit demselben Gewicht in die Analyse eingeht. In der Literatur werden zwei verschiedene Gewichtungen vorgeschlagen. Die Dummies der nominalen Variablen, die Kovariaten und die abhängige Variable werden multipliziert.

1) mit $\sqrt{n_i}$ n_i = Anzahl der abgegebenen Stimmen im Wahlbezirk i

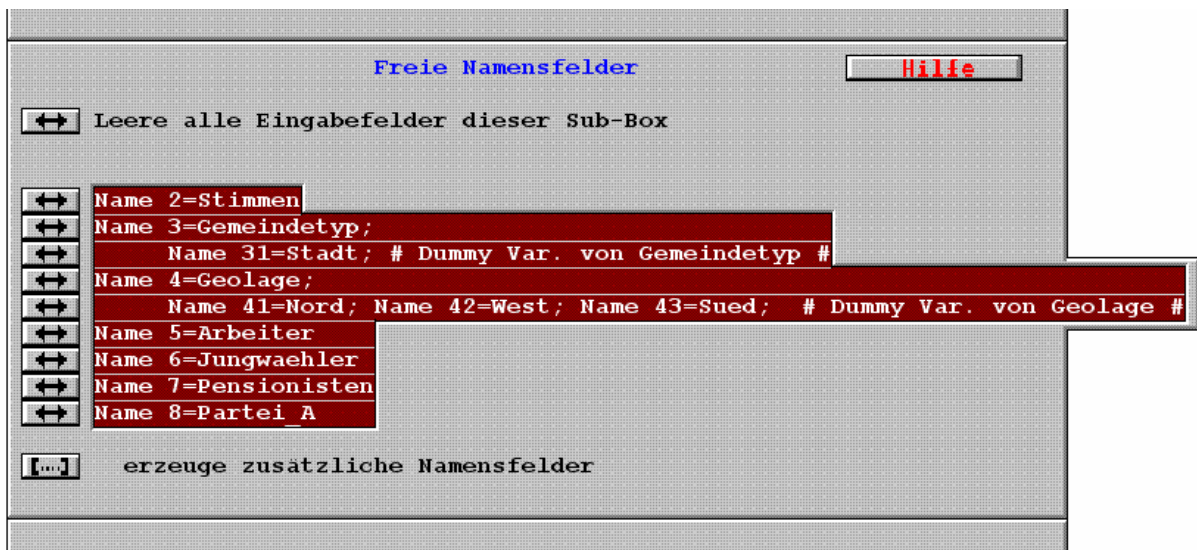
oder

2) mit $\sqrt{\frac{n_i}{p_i(1-p_i)}}$ p_i = Stimmenanteil der Partei im Wahlbezirk i

Siehe z.B. E.Neuwirth: Journal für Sozialforschung, 21, 1981, S.286

Wir wollen für Partei A mit der 2. Gewichtung eine Analyse mit Prog20mo bzw. Prog20mx rechnen.

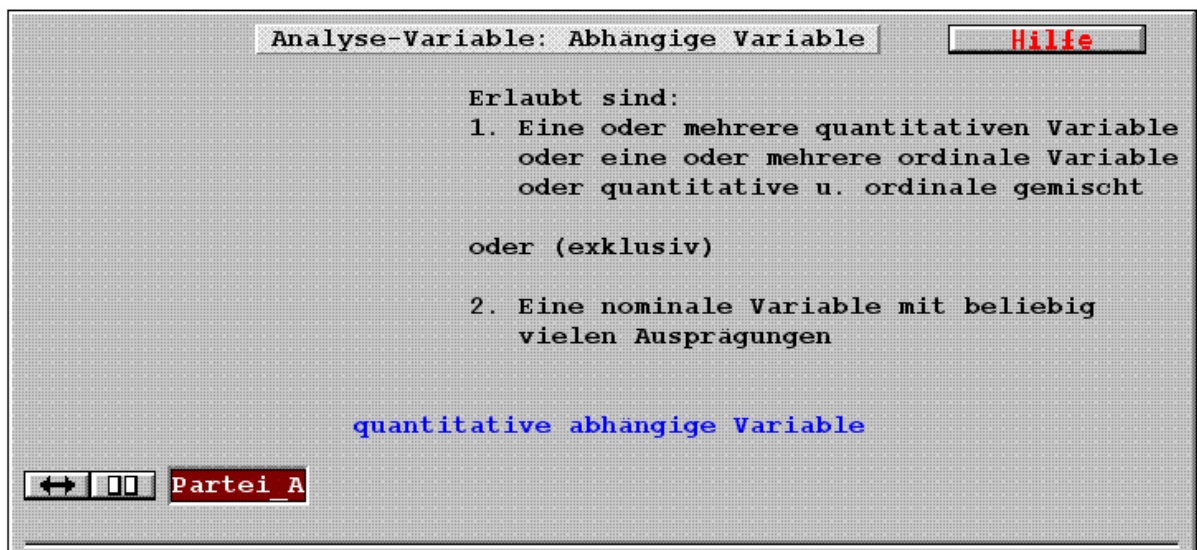
Die Variablen erhalten folgende Namen



Die beiden nominalen Variablen werden in Dummy-Variable aufgelöst: Gemeindetyp in "Stadt". Die redundante (letzte) Dummy "Land" wird nicht gebraucht. Die Dummy-Variable erhält die freie Variablennummer 31.

Geolage wird in die 3 nicht-redundanten Dummies Nord, West und Sued aufgelöst. Sie erhalten die freien Variablennummern 41,42,43. Die letzte Dummy Ost wird nicht gebraucht.

Als abhängige Variable wird der Stimmanteil der Partei_A eingesetzt.



Die unabhängigen Variablen werden in folgender Weise eingegeben

Analyse-Variable: Unabhängige Variable Hilfe

nominale unabhängige Variable Hilfe

0

Interaktionen x. Ordnung zwischen den nominalen unabhängigen Variablen bilden
oder einige ausgewählte Interaktionen bilden
0= keine Interaktionen bilden Hilfe

paarweise Vergleiche (Kontraste) für die nominalen unabhängige Variablen rechnen

quantitative unabhängige Variable Hilfe

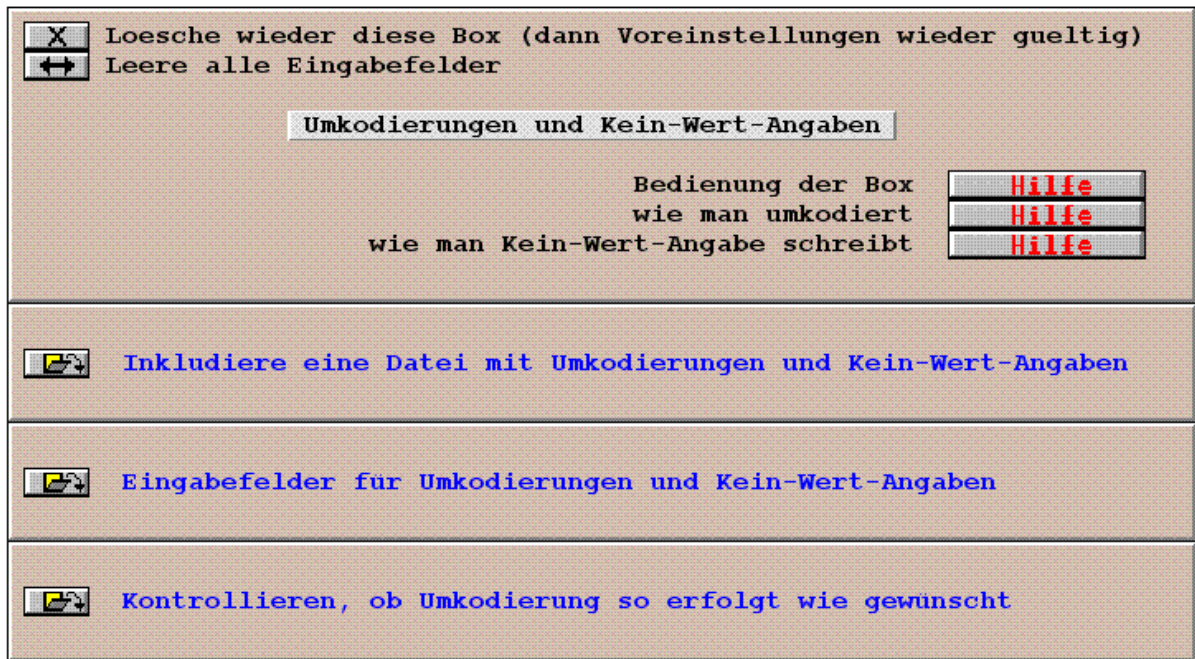
Stadt |Nord,West,Sued| Arbeiter,Jungwaehler,Pensionisten

ordinale unabhängige Variable Hilfe

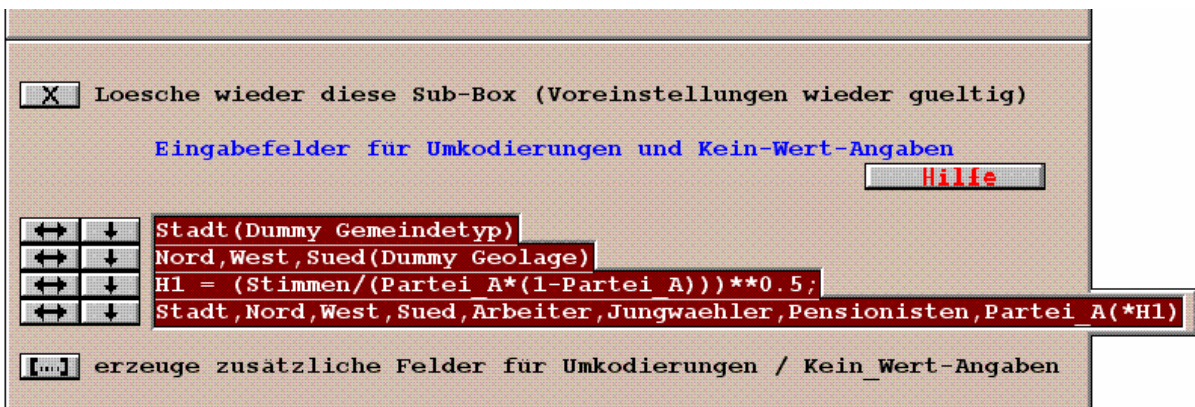
Die 0-1 kodierten Dummies Stadt und Nord,West,Sued werden zusammen mit den anderen unabhängigen Variablen Arbeiter, Jungwaehler und Pensionisten als **quantitative** Variable eingesetzt.

Die 3 Dummies der Geolage sind durch senkrechte Striche zu einer Gruppe eingerahmt. Wir bewirken dadurch, dass für diese 3 Dummies zusammen ein partielles multiples Bestimmtheitsmaß errechnet wird. Siehe dazu Abschnitt P20.11.

Die Unkodierungsbox wird geöffnet. Es erscheint:



Die mittlere Sub-Box wird geöffnet und in folgender Weise ausgefüllt



Die Anweisung `Stadt (Dummy Gemeindetyp)` erzeugt aus der nominalen Variablen die Dummy-Variable "Stadt".

Die 2. Anweisung `Nord, West, Sued (Dummy Geolage)` erzeugt aus Geolage die Dummies Nord, West, Sued.

Mit der 3. Anweisung wird der Gewichtungsfaktor 2 (siehe oben) gebildet und in die Almo-Hilfsvariable H1 gegeben.

In der letzten Anweisung werden dann alle Variable mit H1 gewichtet.

P20.18 Johann Bacher: Interaktionen zwischen nominalen und quantitativen/ordinalen Variablen

Betrachten wir ein Beispiel:

A sei eine unabhängige nominale Variable mit 3 Ausprägungen.

X sei eine quantitative Variable. Sie könnte auch ordinal sein.

Die abhängige Variable y soll bestimmt werden durch A, X und die Interaktion A mit X.

Das ALMO-Syntax-Programm dafür lautet:

```
Vereinbare
Variable = 12;
Anfang
Name1 =A; Name2 =X; Name3=Y;
Name11 =XA1; Name12 =XA2;
Programm=20;

Partial          = X,A aus XA1,XA2;
U_Nominale_V    = A;
Untergrenze A   = 1;
Obergrenze A    = 3;
U_Quantitative_V = X / XA1,XA2;
A_Quantitative_V = y;
Matrix          = Quadratsumme;
Verzichte       = Zellen, Effekte,
                  Vektorausgabe,
                  Matrixausgabe;
Ende_Programmparameter
Lese V1:3 Feld 1,2,2;
H1,2 (Dummy A)
XA1 = X*H1;
XA2 = X*H2;
GP
Gehe_zu Lese
Ende
11223
:
:
*
```

den Variablen werden zuerst Namen gegeben

unabh.nominale Var in 2 Dummies auflösen
Interaktion von x mit den Dummies

Die Vorgehensweise ist also folgende:

1) a. In der Lese-Schleife wird die nominale Variable A in die "notwendigen" Dummies H1 und H2 aufgelöst. Das geschieht durch die Anweisung H1,2 (Dummy A). Da A drei Ausprägungen besitzt, wäre auch eine 3. Dummy zu bilden. Diese ist jedoch redundant.

b. Dann werden die beiden Dummies mit X multipliziert. Es entstehen XA1 und XA2.

Wir sehen:

Besitzt die Untersuchungseinheit die Ausprägung A1, dann ist $XA1 = X$ und $XA2 = 0$,

besitzt sie A2, dann ist $XA1 = 0$ und $XA2 = X$,

besitzt sie A3, dann ist $XA1 = 0$ und $XA2 = 0$.

2) Im Programmparameter-Block muss nun dafür gesorgt werden, dass die beiden Interaktionsvariablen XA1 und XA2 auf die beiden "Haupt"- Variablen A und X keine

auspartiellierende Wirkung haben - jedoch umgekehrt: Das geschieht dadurch, dass in einer PARTIAL-Anweisung A und X aus XA1, XA2 zuvor auspartielliert wird. Zur Wirkung der PARTIAL- Anweisung siehe Abschnitt P20.15 und P19.3.

- 3) Im Programmparameter-Block werden die Interaktionsvariable XA1 und XA2 unter U_QUANTITATIVE_V aufgeführt.
- 4) Ist X eine ordinale Variable, dann müssen auch die Interaktionsvariable XA1 und XA2 als ordinale Variable angegeben werden. Wir müssen also schreiben:

$$U_ORDINALE_V = X / XA1, XA2;$$
 Befinden sich ordinale Variable im Modell, dann müssen die UG und OG angegeben werden. Als Untergrenze für XA1, XA2 wird 0 angegeben, als Obergrenze die maximale Ausprägung von X.

Betrachten wir ein Beispiel, bei dem die beiden nominalen Variablen A und B mit X eine Interaktion bilden - und zusätzlich die Interaktionsvariable AB mit X eine Interaktion bildet. A und B besitzen 3 Ausprägungen.

Die Vorgehensweise ist folgende:

- 1) a. In der Lese-Schleife wird zuerst die nominale Interaktionsvariable AB gebildet.

$$AB = A \text{ MAL } B;$$
 b. Dann werden die notwendigen Dummies von A, B und AB, gebildet. Die Zahl der Dummies von AB ergibt sich aus der Multiplikation der um 1 verringerten Ausprägungen der beteiligten nominalen Variablen also $(3-1) * (3-1) = 4$

$$H1,2(\text{DUMMY A})$$

$$H3,4(\text{DUMMY B})$$

$$H5,6,7,8(\text{DUMMY AB})$$
 c. Die Dummies werden mit X multipliziert. Es entstehen

$$XA1, XA2$$

$$XB1, XB2$$

$$XAB11, XAB12, XAB21, XAB22$$
- 2) Der Programmparameter-Block lautet nun:

```

Programm=20;
Partial = X,A,B aus XA1, XA2, XB1, XB2 /
          X,A,B,AB aus XAB11,XAB12,XAB21,XAB22;

U_Nominale_V      = A,B / AB;
U_Quantitative_V = X /
                  XA1,XA2 | XB1, XB2 /
                  XAB11,XAB12,XAB21,XAB22;

...
...
...
EP
```

Beachte: a. Die PARTIAL-Anweisung wird, gleichgültig, wo sie im Programmparameter-Block steht, immer zuerst ausgeführt. Befindet sich zusätzlich noch eine 2. quantitative Variable Z im Modell, dann können entsprechend auch Interaktionen von Z mit A, mit B und mit AB gebildet werden.

P20.19 Johann Bacher: Überprüfung der Varianzhomogenität

Zwei Annahmen, die in das allgemeine lineare Modell eingehen, sind:

1. Annahme der Homogenität der Regressionskoeffizienten:

Die Regressionskoeffizienten der unabhängigen quantitativen (ordinalen) Variablen sind für jede Merkmalskombination der unabhängigen nominalen Variablen identisch (siehe Abschnitt P20.20).

2. Annahme der Varianzhomogenität:

Die Fehlerstreuung der abhängigen Variablen ist für jede Merkmalskombination der unabhängigen Variablen konstant.

Beide Annahmen lassen sich mit Programm P20 bzw. mit Programm P18 überprüfen. P18 wird im Handbuch, Teil 3 dargestellt.

Bei der Überprüfung der Varianzhomogenität geht man zweistufig vor.

Schritt 1: Mit Programm P18 wird allgemein überprüft, ob Varianzhomogenität vorliegt oder nicht.

Schritt 2: Muss im ersten Schritt die H_0 -Hypothese (=es liegt Varianzhomogenität vor) verworfen werden, lassen sich in einem weiteren Schritt durch die Spezifikation bestimmter Modelle Faktoren der Varianzheterogenität bestimmen. Siehe dazu Handbuch, Teil 3, Abschnitt P18.3.3 und P18.3.4.

Schritt 1 wurde bereits in P18.3.3 und P18.3.4 besprochen. Der zweite Schritt soll nun anhand eines Beispiels beschrieben werden. Das Modell, das gerechnet werden soll, enthält zwei unabhängige nominale Variable A und B. A hat zwei, B drei Ausprägungen. Die abhängige Variable ist y.

P20.19.1 Modelle der Varianzheterogenität

Bezüglich der Varianzheterogenität lassen sich zunächst zwei Modelle unterscheiden.

1. Ein additives Modell:

$$s^2_{ijk} = \tau + \alpha_i + \beta_j + \alpha\beta_{ij} + e_{ijk}$$

s^2_{ijk} ist die k-te Streuung der Zelle (i,j).

2. Ein multiplikatives Modell (das sich leicht in ein lineares überführen lässt):

$$s^2_{ijk} = \tau * \alpha_i * \beta_j * \alpha\beta_{ij} * e_{ijk}$$

Dem Scheffé-Test liegt ein multiplikatives Modell der Varianzheterogenität zugrunde. Der Z-Varianz-Test und der z^2 -Test gehen von einem additiven Modell der Varianzheterogenität aus.

Der Scheffé - Test

Siehe dazu WINER 1972, S. 219.

Jede Zelle (i,j) wird zufällig in weitere Subgruppen (i,j,k) unterteilt ($k=1, \dots, p_{ij}$). Für jede neu gebildete Zelle (i,j,k) wird die Varianz s_{ijk}^2 berechnet. Die Anwendungsvoraussetzungen für den Scheffé-Test sind:

Die Besetzungszahl jeder

- Subgruppe sollte größer 3 sein.
- Die Summe $\sum(p_{ij} - 1)$ sollte größer 10 sein.

WINER (1972, S. 219) empfiehlt weiters, gleichgroße Subgruppen zu bilden. Zur Analyse der Varianzheterogenität wird Programm P20 gerechnet. Die abhängige quantitative Variable ist: $y_{ijk} = \log(s_{ijk}^2)$ bzw. für den gewichteten Scheffé-Test $y_{ijk} = (n_{ijk} - 1) * y_{ijk}$. n_{ijk} ist die Besetzungszahl der Zelle (i,j,k).

Die Logik des Scheffé-Tests und der noch zu besprechenden Tests ist:

Liegt Varianzhomogenität vor, dann sind die durch die nominalen Variablen und deren Interaktionen erklärten Streuungen nicht signifikant. Ist z.B. dagegen die durch A erklärte Streuung signifikant, dann lässt sich die Varianzheterogenität auf Unterschiede der Streuungen hinsichtlich der Variablen A zurückführen.

Z-Varianz-Test

Siehe dazu Overall/Woodward 1974.

$$z_{ij} = \sqrt{\frac{c_{ij} * s_{ij}^2}{MSE}} - \sqrt{c_{ij} - 1}$$

Für jede Zelle (i,j) wird die Größe berechnet.

s_{ij}^2 ist die Streuung der Zelle (i,j)

MSE die geschätzte Gesamtfehlerstreuung

$c_{ij} = (2 + 1/n_{ij}) * (n_{ij} - 1)$ ein Gewichtungsfaktor.

z_{ij} sind die Ausprägungen der abhängigen Variablen Z von Programm P20.

Da das Modell vollständig identifiziert ist, lässt sich die Fehlervarianz nicht mehr schätzen. Sie hat aber - unter H_0 - den Wert 1.0. Zur Überprüfung der Varianzheterogenität muss der Benutzer für jeden Faktor F, das kann eine nominale - oder Interaktionsvariable sein, die mittlere erklärte Streuung $MSS_{yF} = SS_{yF} / (r_F - 1)$ berechnen und mit dem Tabellenwert der F-Verteilung für $(r_F - 1, 0)$ Freiheitsgrade auf Signifikanz testen. Die durch den Faktor F erklärte Streuung SS_{yF} und die Freiheitsgrade $r_F - 1$ sind aus dem Ergebnis von Programm P20 bekannt.

Z²-Test von Levene

Siehe dazu O'BRIEN, 1978.

Der Z²-Test unterscheidet sich von den beiden bisher besprochenen Tests dadurch, dass die ursprünglichen Werte der abhängigen Variablen zur Überprüfung der Varianzheterogenität verwendet werden. Die abhängige Variable ist definiert als $Z_{ijk}^2 = n_{ij} * (y_{ijk} - Y_{ij})^2 / (n_{ij} - 1)$.

Der Mittelwert der (i,j)-ten Zelle ist $Z_{ij}^2 = (Z_{ijk}^2 / n_{ij}) = s_{ij}^2$.

Der Nachteil - neben dem größeren Rechenaufwand - dieses Tests besteht darin, dass die Z_{ijk} je Zelle nicht mehr unabhängig sind und der F-Test einen positiven "bias" aufweist. Bei großen Besetzungszahlen kann dieser aber vernachlässigt werden (O'BRIEN 1978, S. 329).

P20.19.2 Dateneingabe

Das allgemeine Vorgehen besteht darin, dass mit Programm P18 die Varianzen, Mittelwerte und Besetzungszahlen je Zelle durch die Anweisung "OPTION 9=Z_DATEI;" zwischengespeichert werden (P18.3). Anschließend wird mit Programm P20 eine Varianzanalyse gerechnet.

Almo-Syntax-Programm für den Scheffé-Test:

<pre>Vereinbare Variable = 11;</pre>	<p>Speicher für 12 Variable</p>
<pre>Anfang N1 =A; N2 =B; N3=Y;N10 =C; Programm = 18; Nominale_V = A,B,C; Quantitative_V = Y; Untergrenze A,B,C = 3*1; Obergrenze A,B,C = 2,3,2; Option9 = Zwischendatei; Ende_Programmparameter</pre>	<p>den Variablen werden zuerst Namen gegeben C ist die Gruppierungsvariable. Jede Zelle wird in 2 Subgruppen unterteilt. Selbstverständlich können die Personen einer Zelle in mehr als zwei Subgruppen unterteilt werden, sofern die Anwendungsvoraussetzungen für den Scheffé-Test erfüllt sind.</p> <p>Die Daten werden in folgender Reihenfolge geschrieben: A,B,C,Streuung,Besetzung, Mittelwert</p>
<pre>Lese A,B,Y Feld 1,1,2; C(Zufall 0:100;0:50:100=I)</pre>	<p>Jeder Person mit der Ausprägung $A_i B_j$ wird zufällig einer der beiden Subgruppen $A_i B_j C_1$ oder $A_i B_j C_2$ zugeordnet.</p>
<pre>GP Gehe_zu Lese Ende</pre>	
<pre>1107 : *</pre>	
<pre>Anfang Name7 = Slog; Programm = 20; U_Nominale_V = A,B; A_Quantitative_V = Slog; Interaktionen = 2; Verzichte = Zellen, Effekte, Matrixausgabe, Vektorausgabe; Matrix = Quadratsumme; Ende_Programmparameter</pre>	
<pre>Lese A,B,C, V4,5,6 aus Zwischendatei Format frei Leerzu ENDE;</pre>	<p>Aus der Zwischendatei werden die nominalen Variablen A,B,C, die Varianz (=V4), die Besetzungszahl (=V5) und der Mittelwert (=V6) je Zelle eingelesen. Es können natürlich auch andere Variablennummern verwendet werden.</p>

```
Slog = log(V4);
Slog = (V5-1) * log(V4);

GP
Gehe_zu Lese
Ende
```

Es wird ein ungewichteter Scheffé-Test gerechnet, beim gewichteten Scheffé-Test erfolgt hier die Gewichtung nach der angegebenen Formel.

Dateneingabe für den Z-Varianz-Test

Beim Z-Varianz-Test geht man wie beim Scheffé-Test vor. Im ersten Programmparameterblock wird keine Gruppierungsvariable benötigt. Die Leseschleife für Programm P20, in der die Z-Variable berechnet wird, ist:

```
Name11=Z;
H9=1.95;
Lese A,B,C, V4,5,6
aus Zwischendatei
Format frei
leerzu Ende;

H8=(2+1/V4) * (V4-1);

Z = (H8*V3/H9)**0.5 - (H8-1)**0.5;

GP
Gehe_zu Lese
Ende
```

Der Variablen H9 wird der Wert von MSE zugewiesen, dieser wird im Programm P18 bei "OPTION 15=1" berechnet.

Die Gewichtungsfaktoren c_{ij} werden berechnet. Die Variable Z mit den Ausprägungen Z_{ij} wird gebildet.

Dateneingabe für den Z^2 -Test

Beim Z^2 -Test müssen die Mittelwerte und Besetzungszahlen je Zelle in einer Tabelle zwischengespeichert werden, da bei Programm P20 die ursprünglichen Daten verwendet werden. Die Dimension der Tabelle ist $(K,2)$. K ist die Anzahl der Zellen. Die Eingabe lautet:

<pre>Vereinbare Variable = 10; Tabelle_A = 6,2 Anfang N1=A; N2=B; N3=Y; Programm=18; Nominale_V = A,B; Quantitative_V = Y; ... Ende (Daten)</pre>	<p>Speicher für 10 Variable; Tabelle mit 6 Zeilen, 2 Spalten TA = Kurzform von Tabelle_A</p> <p>Die Variablen erhalten Namen Die weitere Eingabe für Programm 18 entspricht dem Scheffé-Test mit der Ausnahme, dass keine Gruppierungsvariable gebildet wird.</p>
<pre>Anfang Lese A,B, V4,5,6 aus Z_Datei Format frei; H1=A mit B; TA(H1,1) =V6; TA(H1,2) =V5; Gehezu Lese Ende</pre>	<p>Für jede Merkmalskombination der nominalen Variablen wird eine Indexvariable gebildet. Die Zellenmittelwerte (=V6) und Besetzungszahlen (=V5) werden in der Tabelle TA zwischengespeichert.</p>
<pre>Anfang Name4=Z2; Programm = 20; U_Nominale_V = A,B; A_Quantitative_V = Z²; Matrix = Quadratsumme; Interaktionen = 2; Verzichte = Zellen, Effekte, Vektorausgabe, Matrixausgabe; Ende_Programmparameter Lese A,B,Y Feld 1,1,2; H1 =A mit B; H2 =TA(H1,1); H3 =TA(H1,2); Z2 =H3*(V5-H2)*(V3-H2)/(H3-1); GP Gehe_zu Lese Ende</pre>	<p>Z2 ist die Z²-Variable Es wird Programm 20 gerechnet</p> <p>Die abhängige Variable ist Z²</p> <p>Die Rohdaten werden eingelesen. Bilden der Indexvariable für A_i, B_j Der Variablen H2 wird der entsprechende Zellenmittelwert, der Variablen die Besetzungszahl der Zelle A_iB_j zugewiesen. Die Variable Z2=Z²_{ijk} wird berechnet</p>

P20.19.3 Interpretation der Ergebnisse

Die erklärten Streuungen der Modelle lassen sich auf die gewohnte Weise interpretieren. Beim Z-Varianz-Test muss der Benutzer die Testgrößen aus den Ergebnissen vom Programm P20 berechnen. Betrachten wir dazu unser Beispiel. Die Werte von SS_y und DF sind aus der Varianzanalyse bekannt.

Ursache	SS _y	DF	MSS _y
AB	0.0370	2	0.0185
A	8.0185	1	8.0185**
B	4.2777	2	2.1334

Die mittlere erklärte Streuung für A ist signifikant, die Varianzheterogenität lässt sich auf die Variable A zurückführen.

P20.20 Johann Bacher: Homogenität der Regressionskoeffizienten

Wie bei der Überprüfung auf Varianzhomogenität wird zweistufig vorgegangen.

Schritt 1: Test auf Homogenität der Regressionskoeffizienten.

Schritt 2: Modellspezifikation.

Beide Schritte sollen wiederum anhand eines Beispiels beschrieben werden. Das Modell, das wir rechnen wollen, ist folgendes:

1. Die abhängige Variable ist Y.
2. Die unabhängigen nominalen Variablen sind A,B und C. A und B besitzen 2 Ausprägungen, C hat 3 Ausprägungen.
3. Die Interaktionen der nominalen Variablen, also AB, AC, BC und ABC.
4. Die Kovariate X.

P20.20.1 Test auf Homogenität der Regressionskoeffizienten

Zwischen den notwendigen Dummies der nominalen Variablen und deren Interaktionen werden mit der Kovariaten multiplikative Variable gebildet. Diese sind:

A1*X		multiplikative Variable 1. Ordnung
B1*X		
C1*X	C2*X	

A1B1*X		multiplikative Variable 2. Ordnung
A1C1*X	A1C2*X	
B1C1*X	B1C2*X	

A1B1C1*X	A1B1C2*X	multiplikative Variable 3. Ordnung
----------	----------	------------------------------------

Diese Variablen werden zusätzlich zur Kovariaten X in das Modell eingebracht.

Die Dateneingabe für den Test ist:

<pre>Anfang N5=Y; N1=A; N2=B; N3=C; N4=X; N11=AB; N12=AC; N13=BC; N14=ABC; N20=A1X; N21=B1X; N22=C1X; N23=C2X; N24=A1B1X; N25=A1C1X; N26=A1C2X; N27=B1C1X; N28=B1C2X; N29=A1B1C1X; N30=A1B1C2X; Programm=20; U_Nominale_V = A, B, C, AB, AC, BC, ABC; U_Quantitative_V = X / A1X B1X C1X,C2X/ A1B1X A1C1X,A1C2X B1C1X, B1C2X / A1B1C1X, A1B1C2X; A_Quantitative_V = Y; Interaktionen = AB ist A mal B,</pre>	<p>Namensgebung abh. Variable nominale Variable Kovariante Interaktionen der nominalen Variablen</p> <p>Es folgen die multiplikativen Variablen</p>
--	---

```

AC ist A mal C,
BC ist B mal C,
ABC ist A mal B mal C;

Partial = A,B,C

aus AB,AC,BC /
A,B,C,AB,AC,BC

aus ABC /
X,A,B,C,AB,AC,BC

aus A1B1X,
A1C1X, A1C2X,B1C1X,B1C2X /
X,A,B,C,AB,AC,BC,ABC

aus A1B1C1X, A1B1C2X;

UG A,B,C=3*1;
OG A,B,C=2,2,3;
Verzichte = Zellen,Effekte,
            Vektorausgabe,
            Matrixausgabe;
Ende_Programmparameter
Lese V1:8 Feld 5*2;
A,B,C,X,Y(0=Kein_Wert)
A1X(Dummy A)
B1X(Dummy B)
C1X,C2X(Dummy C)

H100 =A mal B;
H101 =A mal C;
H102 =B mal C;

A1B1X(Dummy H100)
A1C1X,A1C2X(Dummy H101)
B1C1X,B1C2X(Dummy H102)

H103=A mal B mal C;

A1B1C1X,A1B1C2X(Dummy H103)

A1X,B1X,C1X,C2X,A1B1X,A1C1X,
A1C2X,B1C1X,B1C2X,A1B1C1X,
A1B1C2X ( *X )
GP
Gehe_zu Lese
Ende
(Daten)

```

A1X ist die Dummy von A
B1X ist die Dummy von B
C1X,C2X sind die Dummies von C

H100 enthält die Ausprägungen
für die notwendigen Dummies von AB, H101 von
AC und H102 v. BC.

A1B1X ist Dummy für AB.
A1C1X sind Dummies für AC.
B1C1X,B1 sind Dummies für BC.

H103 enthält die Ausprägungen für die
notwendigen Dummies der Interaktion ABC.
A1B1C1(2)X sind Dummies für ABC.

Durch Multiplikation der Dummies mit der
Kovarioanten X werden die multiplikativen
Variablen berechnet.

Das Vorgehen ist also:

1. In der Leseschleife werden die nominalen und Interaktionsvariablen in ihre notwendigen Dummies aufgelöst.
2. Es werden die multiplikativen Variablen zwischen der Kovariaten und den notwendigen Dummies berechnet.
3. a) Durch die PARTIAL-Anweisung werden Partialvariable gebildet.

Anmerkung:

Da für den Modelltest nur die Schätzung der erklärten Streuungen der Kovariaten und der multiplikativen Variablen entscheidend ist, könnte die PARTIAL-Anweisung weggelassen werden, um Rechenzeit zu sparen. Es genügt die hierarchische Anordnung, die durch Schrägstrich in der Anweisung U_QUANTITATIVE_V=...; hergestellt wird. Wir wollen in diesem Beispiel aber das allgemeine Vorgehen beschreiben, das auch zu einer richtigen Schätzung der erklärten Streuungen der nominalen Variablen führt.

Aus den Interaktionsvariablen AB, AC und BC werden die nominalen Variablen A, B und C herausgenommen, aus der Interaktionsvariablen ABC die nominalen Variablen A, B, C und die - bereits auspartiierten -

Interaktionsvariablen AB, AC und BC. Dieses Vorgehen entspricht soweit dem allgemeinen Modell der Varianz- bzw. Kovarianzanalyse. Da unser Modell zusätzlich Interaktionsvariable zwischen den nominalen Variablen sowie deren Interaktionen mit der quantitativen unabhängigen Variablen enthält, nämlich die multiplikativen Variablen, müssen diese ebenfalls auspartiiert werden: Aus den multiplikativen Variablen A1X, B1X, C1X und C2X werden die Variablen A, B, C und X herausgenommen, aus den multiplikativen Variablen A1B1X, A1C2X, B1C1X und B1C2X die Variablen A, B, C, X und die Interaktionen AB, AC, BC usw.

- b) In der Anweisung U_QUANTITATIVE_V = wird durch den Schrägstrich weiters aus den multiplikativen Variablen A1B1X, A1C1X, A1C2X, B1C1X und B1C2X die multiplikativen Variablen niederer Ordnung, also A1 X, B1X, C1X und C2X herausgenommen. Die Variable X wurde bereits auspartiiert. Wir haben sie dennoch durch einen Schrägstrich von den Variablen A1X, B1X, C1X usw. getrennt, um eine übersichtliche Ausgabe zu erhalten. (Dieser Schrägstrich ist erforderlich, wenn keine PARTIAL-Anweisung verwendet wird. Die Variable X muss in der PARTIAL-Anweisung auspartiiert werden, denn sonst würde im Programm beim gegenseitigen Auspartiierten der nominalen und quantitativen Variablen der direkte Einfluss von X auf die nominalen Variablen unterschätzt und damit die erklärte Streuung der nominalen Variablen überschätzt werden.)

- c) Durch die Auspartiiierung entsteht eine hierarchische Anordnung der Variablen, so dass im Programm selbst folgende Variable gegenseitig auspartiiert werden:

- die unabhängigen nominalen Variablen A, B, C, deren Interaktionen AB, AC, BC, ABC und die quantitative Variable X,
- die Interaktionsvariablen AB, AC, BC, ABC und die multiplikativen Variablen A1X, B1X, C1X, C2X,
- die Interaktionsvariable ABC und die multiplikativen Variablen A1B1X, A1C1X, A1C2X, B1C1X, B1C2X,

usw.

In unserem Beispiel lässt sich unschwer die allgemeine Regel für das Auspartiellieren der multiplikativen Variablen erkennen:

Regel:

Aus den multiplikativen Variablen r-ter Ordnung werden die nominalen Variablen, deren Interaktionen, die quantitativen Variablen und die multiplikativen Variablen bis zur Ordnung r-1 herausgenommen.

4. Die multiplikativen Variablen einer Interaktion der quantitativen Variablen und der entsprechenden nominalen Variablen bzw. deren Interaktionsvariablen werden zu gleichrangigen Gruppen zusammengefasst, also z.B. C1X und C2X. Dadurch werden im Programm für jede Gruppe erklärte Streuungen berechnet. Diese dürfen nicht signifikant sein, wenn die Annahme der Homogenität des Regressionskoeffizienten erfüllt sein soll.
5. Enthält das Modell, das überprüft werden soll, mehrere Kovariaten, werden in der Leseschleife die multiplikativen Variablen für jede Kovariate gebildet und in das Modell einbezogen (siehe Abschnitt P20.18).
6. Als Kovariaten können selbstverständlich ordinale Variable verwendet werden.

Bevor im nächsten Abschnitt die Spezifikation von Modellen mit unterschiedlichen Regressionskoeffizienten für bestimmte Merkmalskombinationen der nominalen Variablen besprochen wird, wollen wir die Ergebnisse unseres Beispiels betrachten.

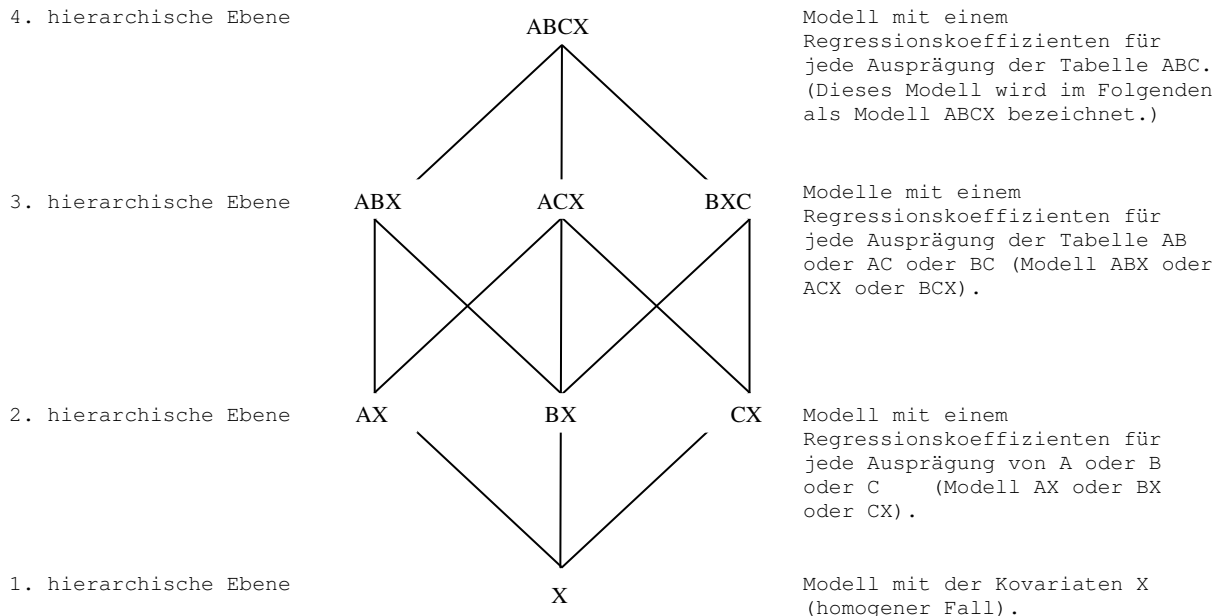
In unserem Beispiel ist die erklärte Streuung der multiplikativen Variablen von B mit X signifikant ($p \leq 0.10$).

Koeffizienten für quant./ordin. Variable
aus univariater Analyse
hinsichtlich der abh. Var. V5 Y

Variable	Regr. Koeff.	St.Abwg.	erklärte Streuung	part. Korr.	F-Wert	Signif. Niveau (1-p)*100
V50 A1B1C1X	0.1432	0.5575	0.0011	0.0422	0.07	20.56%
V51 A1B1C2X	0.2411	0.7456	0.0018	0.0532	0.11	25.37%
Hierarch.Gruppe V50, 51			0.0021	0.0580	0.06	6.07%
V30 A1B1X	-0.2258	0.5329	0.0031	-0.0695	0.18	32.25%
Zusammen V30			0.0031	0.0695	0.18	32.25%
V34 A1C1X	0.0233	0.4400	0.0000	0.0087	0.00	4.31%
V35 A1C2X	0.4560	0.7931	0.0057	0.0941	0.33	42.45%
Zusammen V34, 35						
V40 B1C1X	0.0882	0.5103	0.0005	0.0268	0.03	13.42%
V41 B1C2X	0.1688	0.4905	0.0020	0.0565	0.12	26.76%
Zusammen V40, 41			0.0020	0.0566	0.06	5.78%
Hierarch.Gruppe V30, 34, 35, 40, 41			0.0156	0.1552	0.18	3.42%
V20 A1X	0.5304	0.3590	0.0374	0.2360	2.18	85.55%
Zusammen V20			0.0374	0.2360	2.18	85.55%
V22 B1X	0.6231	0.3396	0.0577	0.2888	3.37	92.88%
Zusammen V22			0.0577	0.2888	3.37	92.88%
V24 C1X	0.1025	0.3743	0.0013	0.0450	0.08	21.80%
V25 C2X	-0.3666	0.4013	0.0143	-0.1485	0.83	36.99%
Zusammen V24, 25			0.0253	0.1958	0.74	48.92%
Hierarch.Gruppe V20, 22, 24, 25			0.1237	0.4040	1.80	85.21%
V6 X	0.0247	0.1450	0.0005	0.0280	0.03	14.01%
Hierarch.Gruppe V6			0.005	0.0280	0.03	14.01%

P20.20.2 Modellspezifikationen: Modelle mit unterschiedlichen Kovariaten je Merkmalskombination der nominalen Variable(n)

Bei der Modellspezifikation ist es zweckmäßig, sich einen hierarchischen Baum aller möglichen Modelle aufzuzeichnen. In unserem Beispiel hat dieser folgende Darstellung:



Die Äste der Baumdarstellung zeigen an, welches Modell einer hierarchischen Ebene in Modellen der übergeordneten Ebene enthalten ist. So ist z.B. das Modell AX ein Submodell der Modelle ABX und ACX, nicht aber von BCX, da sich die Randverteilung von A nicht aus der Tabelle von B mit C ableiten lässt. Allgemein ist also ein Submodell in jenen Modellen der übergeordneten Hierarchieebene enthalten, aus denen sich die Tabelle bzw. die Randverteilung dieses Submodells ableiten lässt.

Im Unterschied zur Teststrategie der Regressionskoeffizienten auf Homogenität sind bei der Modellspezifikation die Modelle der 4. Ebene denen der 3. bis 1. Ebene übergeordnet, Modelle der 3. Ebene denen der 2. und 1. Ebene usw. Die Hierarchieebene wird also umgedreht.

Für die Modellspezifikation gelten folgende Regeln:

- Auf der entsprechenden Hierarchieebene dürfen nur die multiplikativen Variablen für eine gleichrangige Gruppe signifikant sein, also z.B. auf der 3. Hierarchieebene für das Modell ACX nur die multiplikativen Variablen $A1C1X$ und $A1C2X$.
- Alle erklärten Streuungen der multiplikativen Variablen der übergeordneten Hierarchieebene(n) dürfen nicht signifikant sein.
- Die erklärten Streuungen der multiplikativen Variablen von Modellen der niederen Hierarchieebene, die keine Submodelle des gesuchten Modells sind, müssen nicht signifikant sein.

- d) Die erklärten Streuungen der multiplikativen Variablen von Modellen der niederen Ebene, die als Submodelle im gesuchten Modell enthalten sind, können signifikant sein.

In unserem Beispiel sind die vier Bedingungen erfüllt, um das Modell BX mit einem Regressionskoeffizienten für jede Ausprägung von B zu spezifizieren.

- a) Die erklärten Streuungen von C mit X, also C1X und C2X, und von A mit X, also A1X, sind nicht signifikant.
- b) Die erklärten Streuungen der übergeordneten Modelle, also von ABX, ACX, BCX und ABCX sind ebenfalls nicht signifikant.
- c) Da nur mehr das homogene Modell dem Modell BX untergeordnet ist, braucht Regel c nicht überprüft werden.
- d) Die erklärte Streuung des homogenen Modells spielt gemäß Regel d keine Rolle.

Die Eingabe für das Modell BX ist:

```
Vereinbare
Variable = 25;
Anfang
N1=A; N2=B; N3=C;
N22=B1X; N23=B2X;

Programm=20;
U_Nominale_V      = A,B,C;
U_Quantitative_V  = B1X,B2X;
A_Quantitative_V  = Y;
Interaktionen     = 3;
UG A,B,C          = 3*1;
OG A,B,C          = 2,2,3;
Ende_Programmparameter

Lese V1:5 Feld 5*2;
V1:5 (0=KW)
B1X,B2X (*X)
GP
Gehe_zu Lese
Ende
(Daten)
```

den Variablen werden zuerst Namen gegeben

Beachte: In das Modell werden alle multiplikativen Variablen, die durch Multiplikation aller Dummies einer nominalen Variablen bzw. deren Interaktionen mit der Kovariaten entstehen, aufgenommen. Soll z.B. das Modell ABX gerechnet werden, dann muss in der Leseschleife die MIT-Anweisung verwendet werden. Dadurch werden alle multiplikativen Variablen der Interaktion von X mit AB gebildet. In unserem Beispiel also die Variablen A1B1X, A1B2X, A2B1X und A2B2X. Diese Variablen werden in das Modell einbezogen. Die Anweisung in der Leseschleife für das Modell ABX ist:

```
H100 = A MIT B;
A1B1X,A1B2X,A2B1X,A2B2X( DUMMY H100; *X )
```

Kurt Holm

P20.21 Nichtlineare Regression mit dem ALM

Mit Programm 20 können die Parameter von nichtlinearen Funktionen ermittelt werden, wobei folgende Einschränkungen gelten:

- (1) Es ist nur eine unabhängige Variable möglich
- (2) Die nichtlineare Funktion muss durch eine Transformation linearisierbar sein.

Im Folgenden werden wir 5 nichtlineare Funktionstypen behandeln. Es sind dies

- (1) die parabolische und hyperbolische Funktion,
- (2) die Exponentialfunktion
- (3) die allgemeine Exponentialfunktion
- (4) die Gompertz-Funktion
- (5) die logistische Funktion

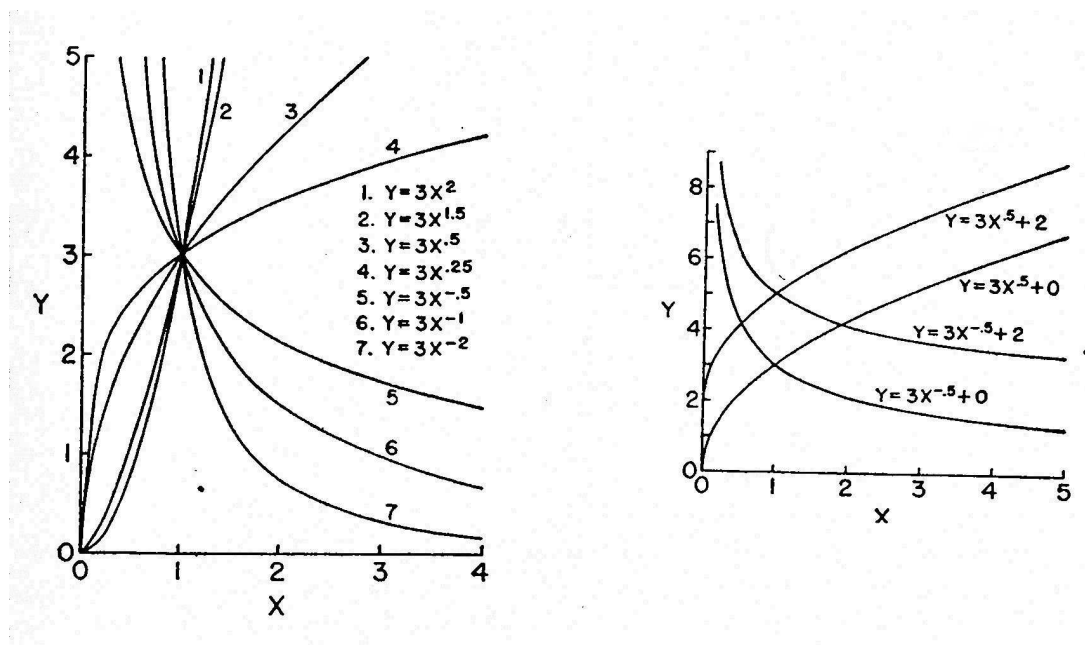
P20.21.1 Die parabolische und hyperbolische Funktion

Ihre Gleichung lautet

$$y = a \cdot x^b + \text{const}$$

wobei a und b Parameter sind, die es zu finden gibt. Die additive Konstante const muss geschätzt werden. Man wird sie zuerst auf .0 setzen.

Betrachten wir einige konkrete Parabeln (entnommen aus D. Lewis, 1960).



P20.21.1.1 Eigenschaften der parabolischen Funktion

Die Funktion ist parabolisch, wenn der Parameter b positiv ist.

Alle Kurven laufen durch die Punkte $(0,0)$ und $(1,a)$ wenn $\text{const} = 0$.

Ist $\text{const} \neq 0$ dann laufen die Kurven durch die Punkte $(0, \text{const})$ und $(1, a + \text{const})$ d.h. die Kurve wird insgesamt um const verschoben.

Die Kurve steigt mit abnehmender Steigung, wenn b kleiner 1 und mit zunehmender, wenn b größer 1.

P20.21.1.2 Eigenschaften der hyperbolischen Funktion

Die Funktion ist hyperbolisch, wenn b negativ ist.

Alle Kurven laufen durch den Punkt $(1,a)$ und haben die Koordinatenachsen x und y als Asymptoten - wenn $\text{const} = 0$.

Ist $\text{const} \neq 0$ dann laufen die Kurven durch den Punkt $(1, a + \text{const})$ und haben die y -Achse und die Gerade $y = \text{const}$ als Asymptoten.

Die Kurve nähert sich umso "früher" ihres Asymptoten an, je kleiner b (je größer der negative Wert von b).

P20.21.1.3 Schätzung der Parameter durch lineare Regression

Die parabolische-hyperbolische Funktion

$$y = a \cdot x^b + \text{const}$$

wird durch Logarithmieren in folgende lineare Funktion transformiert:

$$\log(y - \text{const}) = \log a + b \cdot \log x$$

Wir setzen also

$$y^* = \log(y - \text{const})$$

$$x^* = \log x$$

und rechnen mit Programm 20 eine lineare Regression. Aus ihr erhalten wir b^* als Regressionskoeffizient und a^* als Regressionskonstante. Die gesuchten Parameter a und b der parabolisch-hyperbolischen Funktion erhalten wir dann durch

$$b = b^*$$

$$a = 10^{\log a^*}$$

P20.21.1.4 Eingabe in Almo-Maskenprogramm

Das Almo-Maskenprogramm zur Bestimmung der Parameter a und b ist folgendes:

Prog20mp.Msk
 nicht-lineare Regression:
 Parabel- und Hyperbel-Funktionen
 mit Grafik

Diese haben die Gleichung:

$$Y = a * X^{**}b + const$$

(** = Potenzieren)

wobei a und b Parameter sind, die es zu finden gilt.
 Die additive Konstante 'const' muss geschätzt werden.

Bei der Parabel ist b positiv, bei der Hyperbel negativ
 Alle Parabeln und Hyperbeln laufen durch den Punkt (1,a)

Obige Gleichung kann durch Logarithmieren in folgende lineare Gleichung umgeformt werden:

$$\log(Y-const) = \log a + b*\log X$$

Wir setzen $y = \log(Y-const)$
 $x = \log X$

und rechnen eine Regressionsanalyse, die uns b und log a ausgibt. a erhalten wir dann durch $a = 10^{**}\log a$

Daten aus Don Lewis: Quantitative Methods in Psycholgy,
 McGraw-Hill, 1960, S.53

Was ist ein Kurzprogramm ? -->
 Bedienung -->

1

Speicher fuer x Variable

Vereinbare Variable = ;

2



Option: Weitere Vereinbarungen - nur wenn Almo dazu auffordert

3

Variablennummern

Setzen Sie hier die Variablennummern
 der unabhängigen und der abhängigen
 Variablen ein

Name =x;
 Name =y;

unabhängige Variable
 abhängige Variable

4

Name für Gruppierungsvariable

$(Sie können eine Gruppierungsvariable angeben$
 das ist nicht obligatorisch)



Name

5

Datei aus der gelesen wird

bei Datei-Problemen



Format der Daten



der Datensatz enthält diese Variablen
 Bei Format DIREKT schreiben Sie: alle_U

- 6 Wenn Dateiformat FIX oder Nicht-Standard-FREI
- 7 Sie können eine Gruppierungsvariable angeben
das ist nicht obligatorisch
 Geschlecht
- 8 Option: Ein- und Ausschliessen von Untersuchungseinheiten
- 9 Option: Umkodierungen und Kein-Wert-Angaben
- 10 Option: Spezielle Kein-Wert-Behandlung
- 11 Die additive Konstante 'const' muss geschätzt werden
- 12 **Almo**
Almo = Almo-Grafik ausgeben
0 = nicht
- 13 Verzichte auf nicht bedeutsame Teile der Ergebnis-Ausgabe
1= ja
0= nein
- 14 **zeige**
zeige = Almo-Programmtext zeigen
Editfeld leer = nicht zeigen

Erläuterungen zu den Boxen

Box "Variablennummern"

Variablennummern
Setzen Sie hier die Variablennummern der unabhängigen und der abhängigen Variablen ein

Name **1** =x; # unabhängige Variable #
Name **2** =y; # abhängige Variable #

Die Daten, die analysiert werden, sind folgende:

x	y	G
1	18.85	1
3	22.81	2
5	25.14	2
7	26.64	2
9	28.06	1
11	28.96	2
13	29.87	2
15	30.99	1
16	31.21	1

Die unabhängige Variable x ist V1, hat also die Variablennummer 1. Die abhängige Variable y ist die 2. Variable im Datensatz, hat also die Variablennummer 2. "G" ist die Gruppierungsvariable.

Box 3 und Box 5: Gruppierungsvariable

Name für Gruppierungsvariable
(Sie können eine Gruppierungsvariable angeben das ist nicht obligatorisch) [Hilfe](#)

↔ **Name 3=Geschlecht:männlich,weiblich;**

Sie können eine Gruppierungsvariable angeben
das ist nicht obligatorisch [Hilfe](#)

↔ **Geschlecht**

Beispiel: Die Gruppierungsvariable sei das Geschlecht. Dann werden die Männer und die Frauen im grafischen Streudiagramm (in das auch die Kurve eingezeichnet wird) durch verschiedene Symbole (kleine Kreise, kleine Quadrate) dargestellt. Eine andere Wirkung hat die Gruppierungsvariable nicht.

Box "Datei aus der gelesen wird"

Siehe P0.4.

Box "Option: Ein- und Ausschließen von Untersuchungseinheiten"

Siehe P0.7.

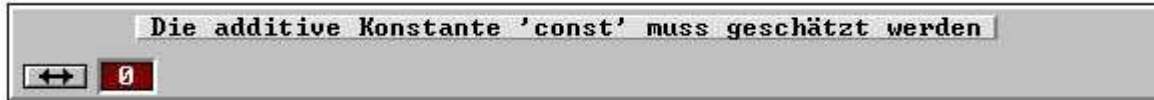
Box "Option: Umkodierungen und Kein-Wert-Angaben"

Siehe P0.5.

Box "Option: Spezielle Kein-Wert-Behandlung"

Siehe sie ausführliche Erläuterung in P20.8.1.1.

Box "Die additive Konstante 'const' muss geschätzt werden"



Die Konstante kann im Programmverlauf nicht geschätzt werden. Der Benutzer muss einen Schätzwert festlegen. In der Regel wird man 0 eingeben.

P20.21.1.5 Das selbst geschriebene Almo-Programm

Siehe dazu das Beispielprogramm Nonlin1a.Alm. Sie finden es im Menü „Almo/Liste aller Almo-Programme“.

P20.21.1.6 Das Ergebnis

Das Maskenprogramm bzw. das äquivalente selbst geschriebene Almo-Programm liefert folgendes Ergebnis (etwas gekürzt)

```
Ergebnisse aus ALMO
-----
Nachfolgend wird die nichtlineare Funktion
      Y = a * X**b + const
in die lineare Gleichung
      log(Y-const) = log a + b*log X
transformiert
Für diese wird eine Regressionsanalyse gerechnet
Der Regressionskoeffizient b und
die Regressionskonstante a* = log a werden berechnet
=====

***** MITTEILUNG
Allgemeines lineares Modell wird mit folgenden Einstellungen gerechnet:

  Analysiert wird die Matrix der Abweichungsquadrate
  Die Streuungen sind Abweichungsquadrate
  Es entstehen nicht-standardisierte Koeffizienten

=====
Alle im Folgenden angegebenen Streuungen und erklarte Streuungen sind
Abweichungsquadratsummen
=====

Gesamtstreuung                                0.041081
=====
Koeffizienten fuer Gesamt-Modell

Durch alle unabh. Variable
erklarte Streuung                             0.041051
Fehlerstreuung                               0.000030
multipler Korrelat.koeff.                    0.999631
F-Wert f. erklarte Streuung                   9481.718733
Freiheitsgrade Nenner = 1
              Zaehler= 7
Signifikanz (1-p)*100                         99.999500 %
=====
```

******* Erläuterung:**

Der multiple Korrelationskoeffizient R und seine Signifikanz können als ein Maß der Anpassung der Kurve an die Datenpunkte betrachtet werden. In unserem Beispiel ist R mit 0.999631 ungewöhnlich hoch. R ist mit 99.9995 eindeutig signifikant.

Koeffizienten fuer quantitat./ordinale Variable aus univariater Analyse
hinsichtlich der abh.Var. V2 y

Variable	Regr. koeff.	Standard fehler	95% Konfidenzbereich nach		erklarte Streuung	part. Korrel.	F-Wert	Signif (1-p)100	df1	df2
			oben	u.unten						
V1 x	0.1826	0.0019	0.0044	0.0411	0.9996	9481.72	100.00	1	7	

******* Erläuterung:**

Der Regressionskoeffizient mit 0.1826 ist bereits der gesuchte Parameter b der parabolisch-hyperbolischen Funktion.

Koeffizienten fuer Konstante

hinsichtlich der abh.Variablen	V2 y
Effekt (Regressionskoeffizient)	1.273160

******* Erläuterung:**

Die Regressionskonstante mit 1.273160 ist noch nicht der gesuchte Parameter a der parabolisch-hyperbolischen Funktion. Aus ihm muss der Anti-log genommen werden. Das geschieht später.

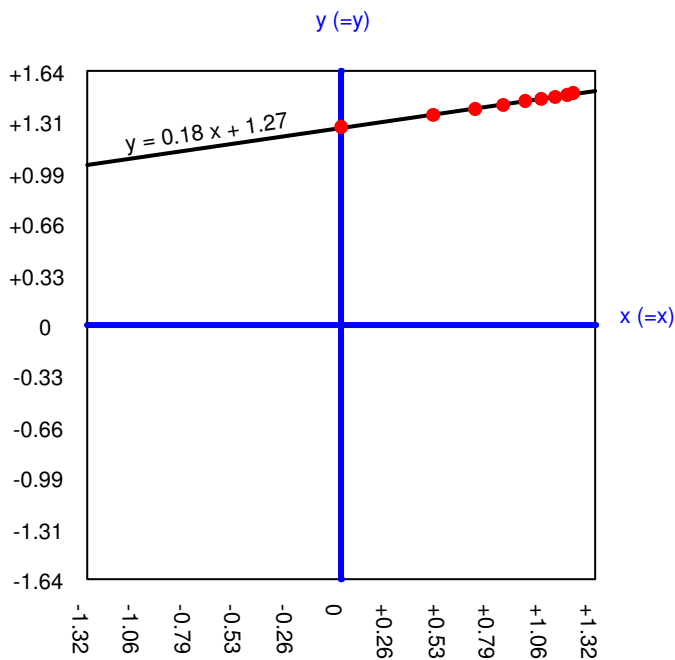
Ergebnisse aus ALMO

Die lineare Gleichung
 $\log(Y-\text{const}) = \log a + b \cdot \log X$
wird im Folgenden grafisch dargestellt
=====

Ergebnisse aus ALMO

Almo erzeugt hier folgende Grafik:

Streudiagramm



***** Erläuterung:

Die Datenpunkte liegen fast exakt auf der Regressionsgeraden.

Ergebnisse aus ALMO

Die Parameter der nichtlinearen Funktion

$$Y = a * X^{**}b + const$$

sind folgende:

Parameter a = 18.75686

Parameter b = 0.1826035

Konstante const = 0

Die Gleichung lautet also:

$$Y = 18.75686 * X^{**}0.1826035 + 0$$

***** Erläuterung:

Die obige Gleichung ist nun die gesuchte parabolisch-hyperbolische Gleichung. Der Parameter a (=18.75686) entstand als Antilog aus der Regressionskonstanten (=1.27316)

Ergebnisse aus ALMO

Die nicht lineare Gleichung

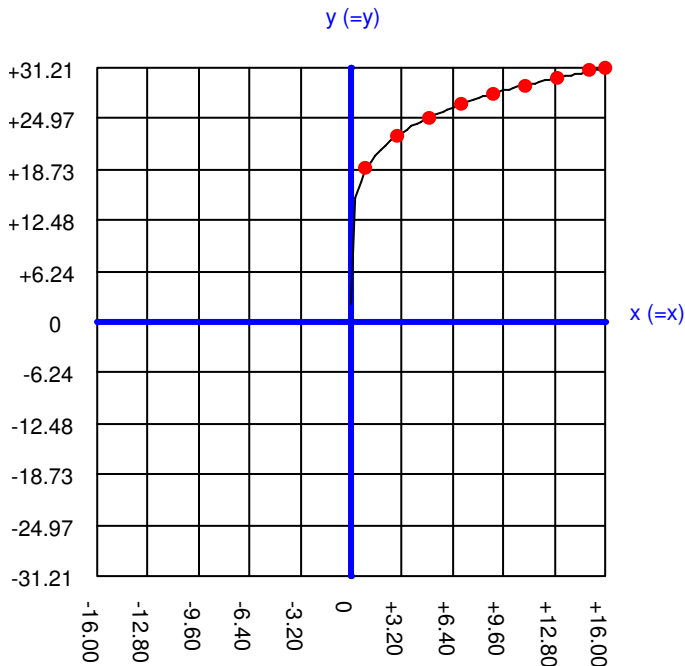
$$Y = a * X^{**}b + const$$

wird im Folgenden grafisch dargestellt

Almo erzeugt hier folgende Grafik:

Streudiagramm

Funktion: $y = 18.7569 \cdot x^{**0.182604} + 0$



***** Erläuterung:

Die Datenpunkte liegen fast exakt auf der parabolisch-hyperbolischen Kurve.

P20.21.2 Die Exponential-Funktion

Die Gleichung für die spezielle Exponentialfunktion lautet

$$(1) y = a \cdot e^{b \cdot x} + \text{const}$$

wobei $e = 2.7183$

Für die allgemeine Exponentialfunktion lautet die Gleichung

$$(2) y = a \cdot b^x + \text{const}$$

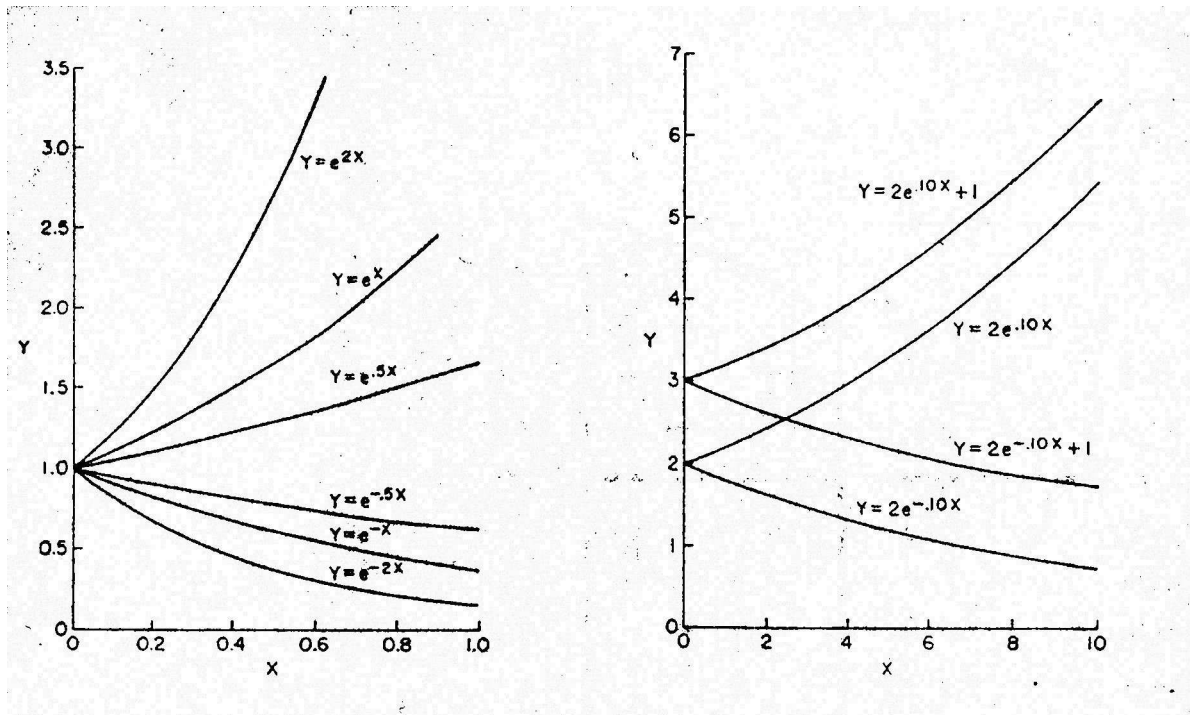
Gleichung (2) kann in (1) überführt werden

$$(2a) y = a \cdot e^{b^* \cdot x}$$

wobei $b^* = \log b / \log e = \log b / 0.4343$

Die Parameter a und b gilt es zu finden. Die additive Konstante const muss geschätzt werden. Man wird sie zuerst auf 0 setzen.

Betrachten wir einige konkrete Kurven der speziellen Exponentialfunktion (entnommen aus D. Lewis, 1960).



Eigenschaften der speziellen Exponentialfunktion

Die Kurve steigt, wenn b positiv ist. Sie fällt, wenn b negativ ist

Die Kurve verläuft durch den Punkt $(0, a + \text{const})$

Ist b negativ, dann ist die Gerade $y = \text{const}$ die Asymptote

Ist b negativ, dann nähert sich die Kurve umso „früher“ ihrer Asymptoten $y = \text{const}$ je kleiner b (je größer der negative Wert von b)

Schätzung der Parameter durch lineare Regression

Die spezielle Exponentialfunktion in (1) wird durch Logarithmieren zu folgender linearen Funktion

$$\log(y - \text{const}) = \log a + \log e \cdot b \cdot x$$

wobei $\log e = 0.4343$

Wir setzen

$$y^* = \log(y - \text{const})$$

und rechnen mit Prog20 eine lineare Regression. Aus ihr erhalten wir b^* als Regressionskoeffizienten und a^* als Regressionskonstante. Die gesuchten Parameter a und b erhalten wir dann durch

$$b = b^* / \log e = b^* / 0.4343$$

$$a = 10^{\log a^*}$$

Die allgemeine Exponentialfunktion linearisieren wir durch Logarithmieren zu

$$\log(y - \text{const}) = \log a + \log b \cdot x$$

Wir setzen

$$y^* = \log(y - \text{const})$$

und rechnen mit Prog20 eine lineare Regression, die uns b^* als Regressionskoeffizienten und a^* als Regressionskonstante liefert. Die gesuchten Parameter a und b erhalten wir durch

$$b = 10^{1 \log b^*}$$

$$a = 10^{1 \log a^*}$$

Eingabe und Ergebnis

Siehe dazu die Maskenprogramme Prog20mq und P20mr, sowie die Beispiel-Programme Nonlin2a.Alm und Nonlin3a.Alm. Die Maskenprogramme sind strukturgleich zu dem in P20.21.1.4 dargestellten und erläuterten Programm Prog20mp für die parabolisch-hyperbolische Funktion. Das Ergebnis gleicht in seiner Struktur dem in Abschnitt P20.21.1.6 dargestellten Ergebnis für die parabolisch-hyperbolische Funktion.

P20.21.3 Die Gompertz-Kurve

Ihre Gleichung lautet

$$y = a \cdot b^{c^x}$$

oder

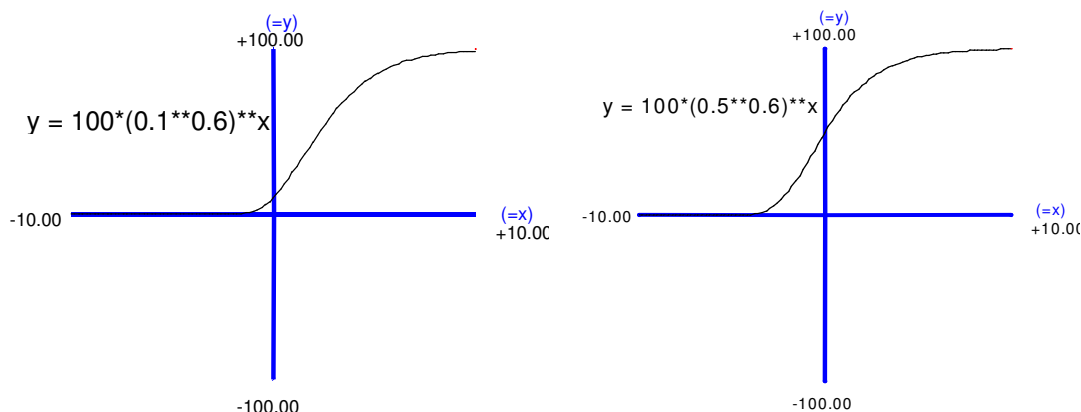
$$(1a) y' = b^{c^x}$$

wobei

$$y' = \frac{y}{a}$$

b und c sind Parameter, die es zu finden gilt. a muss geschätzt werden.

Betrachten wir einige konkrete Gompertz-Kurven:



Eigenschaften der Gompertz-Kurven

Wenn b und c größer 0 und kleiner 1 sind, dann besitzt die Gompertz-Kurve die Asymptote $y=a$ und

dann besitzt die Gompertz-Kurve einen Wendepunkt, wobei dieser im positiven x-Bereich (rechts des Ursprungs) liegt, wenn b größergleich $1/e$ ($=0.3678$) ist und dieser (der Wendepunkt) im negativen x-Bereich (links des Ursprungs) liegt, wenn b kleiner $1/e$ ($=0.3678$) ist.

Bis zum Wendepunkt steigt die Gompertz-Kurve mit zunehmender Steigung, danach mit abnehmender Steigung.

Der Wendepunkt ist:

$$x_w = \frac{-\ln(-\ln b)}{\ln c}$$

Die Gompertz-Kurve ist im Unterschied zur nachfolgend dargestellten logistischen Funktion nicht symmetrisch um den Wendepunkt. Sie erreicht auf der linken Seite $y = 0$ schneller als sie auf der rechten Seite $y = a$ erreicht.

Schätzung der Parameter durch lineare Regression

Die Gompertz-Kurve wird durch Logarithmieren linearisiert zu

$$\log(-(\log y - \log a)) = \log(-\log b) + \log c \cdot x$$

Wir setzen

$$y^* = \log(-(\log y - \log a))$$

und rechnen mit Programm 20 eine lineare Regressionsanalyse, die uns den Regressionskoeffizienten c^* und die Regressionskonstante b^* liefert. Die gesuchten Parameter b und c erhalten wir durch

$$c = 10^{\log c^*}$$

$$b_1 = 10^{\log b^*}$$

$$b = 10^{-\log b_1}$$

Eingabe und Ergebnis

Siehe dazu das Maskenprogramm Prog20ms.Msk und das Beispiel-Programm Nonlin4a.Alm. Die Maskenprogramme sind strukturgleich zu dem in P20.21.1.4 dargestellten und erläuterten Programm Prog20mp für die parabolisch-hyperbolische Funktion. Das Ergebnis gleicht in seiner Struktur dem in Abschnitt P20.21.1.6 dargestellten Ergebnis für die parabolisch-hyperbolische Funktion.

P20.21.4 Die logistische Funktion

Diese hat die Gleichung:

$$(1) \quad Y = 1 / (1/a + e^{-(b+c*x)}$$

Anmerkung: Das negative Vorzeichen vor dem Exponenten $-(b+c*x)$ kann auch weggelassen werden. Die logistische Funktion ist dann einfach um die Senkrechte durch ihren Wendepunkt gedreht.

Durch Umformung erhält man:

$$(2) \quad 1/Y - 1/a = e^{-(b+c*x)}$$

e = die Zahl e (=2.718...)

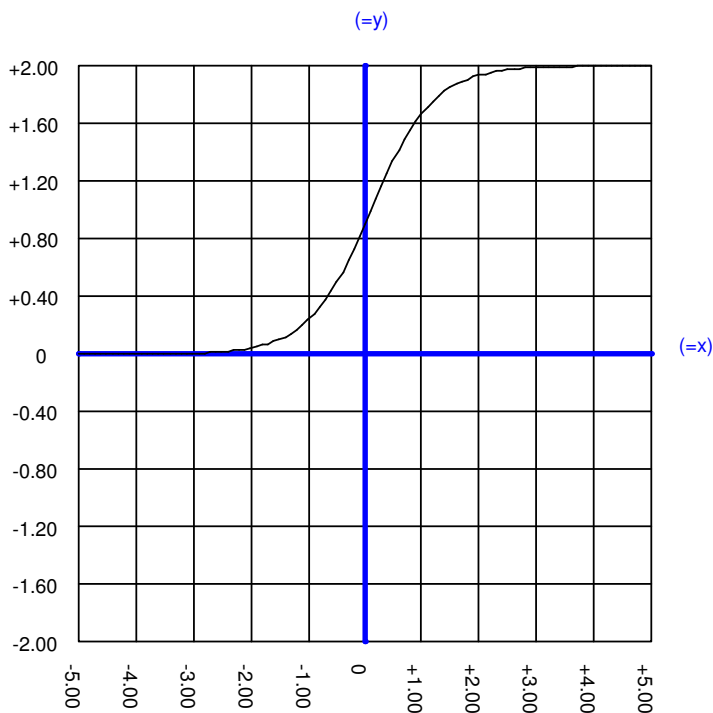
** = Potenzieren

b und c sind Parameter, die es zu finden gilt.

a muss geschätzt werden. Es ist gleich der Obergrenze der sich die logistische Kurve asymptotisch annähert.

Betrachten wir eine konkrete logistische Kurve:

$$Y = 1 / (1/a + e^{**}-(b+c*x)) \text{ mit } a=2, b= 0.5, c= 1,8$$



Eigenschaften der logistischen Funktion:

Der Parameter a bestimmt die Obergrenze, der sich die logistische Funktion annähert. Die logistische Kurve besitzt in halber Höhe einen Wendepunkt, d.h. bei $y=a/2$. Der Parameter b verschiebt die Kurve nach links. Je größer b umso weiter links befindet sich die Kurve. Der Parameter c bestimmt die Steilheit. Je größer der absolute Wert von c umso steiler. Ist das Vorzeichen von c positiv, dann wächst die Kurve von links nach rechts. Ist das Vorzeichen negativ, dann umgekehrt von rechts nach links

Schätzung der Parameter durch lineare Regression

Obige Gleichung (2) kann durch Logarithmieren in folgende lineare Gleichung umgeformt werden:

$$\log(1/Y-1/a) = -0.434*b - 0.434*c * X$$

0.434 ist log e

Wir setzen $y = \log(1/Y-1/a)$

und rechnen eine Regressionsanalyse für dieses Y als abhängige und X als unabhängige Variable, die uns b' und c' ausgibt. b und c erhalten wir dann durch Dividieren mit 0.434 und Vorzeichenumkehr, also $b = -b'/0.434$ und $c = -c'/0.434$

Eingabe und Ergebnisse

Siehe dazu das Maskenprogramm Prog20mw.Msk. Die Maskenprogramme sind strukturgleich zu dem in P20.21.1.4 dargestellten und erläuterten Programm Prog20mp für die parabolisch-hyperbolische Funktion. Das Ergebnis gleicht in seiner Struktur dem in Abschnitt P20.21.1.6 dargestellten Ergebnis der parabolisch-hyperbolischen Funktion

P20.21.5 Literatur zu nichtlineare Regression

Don Lewis: Quantitative Methods in Psychology, McGraw-Hill, New York, 1960

Schlagwortverzeichnis

- Anpassung 88
- Cramers V 27, 41
- Determinante 19
- Diskriminanzanalyse 42
- Effekte 40
- Exponential-Funktion 132
- generalisierte Varianz 18
- Geschachtelte Variable 90
- Gleichrangige Gruppen 65
- Gompertz-Kurve 134
- Gruppenweise hierarchische Regression 62
- Häufigkeitstabelle 39
- hierarchische Gruppe 62
- Hierarchische Regression 60
- Hierarchische Versuchspläne 92
- Homogenität der Regressionskoeffizienten 118
- Hotelling-Lawley Spur 24
- hyperbolische Funktion 124
- Interaktionen zwischen nominalen und quantitativen/ordinalen Variablen 111
- Interaktionsvariablen 73
- kanonische Korrelationsanalyse 27
- Leere Zellen 77
- lineare Wahrscheinlichkeitsanalyse 42
- logistische Funktion 135
- Logit-Analyse 5
- Meßwiederholungen 93
- Minimum-Chi-Quadrate-Schätzung 5
- Multivariate Analyse 18
- Multivariate Effekte 33
- Nichtlineare Regression 124
- parabolische Funktion 124
- PARTIAL-Anweisung 85
- Partialvariable 60, 85
- partielle multiple Korrelation 65
- Pillais Korrelation 25, 41
- Pillais Spur 23
- Polynom 64
- Regressionskoeffizienten 45
- Scheffé - Test 114
- Schrägstriche 62
- senkrechten Strich 65
- Spearman-Brown-Formel der Zuverlässigkeit 105
- SPSS 77
- Tabellenanalyse 35
- Unvollständige Versuchspläne 89
- Varianzheterogenität 113

Varianzhomogenität 113
wiederholte Messungen 93
Wilks Korrelation 25
Wilks Lambda 21, 30
Z²-Test von Levene 114

Zelleneffekte 84
Zufallsvariable 91
Zuverlässigkeit 105
Z-Varianz-Test 114

P20.22 Literatur zum Allgemeinen Linearen Modell

- Aldrich/Nelson:** Linear Probability, Logit and Probit Models, Sage Publications 1984, S. 14 ff.
- Bock, R.D.:** Multivariate Statistical Methods in Behavioral Research, Mc. Graw Hill, 1975 Costner, H.: Criteria for measures of association, in: Am.Soc.Review 1965
- Bortz, J.:** Statistik, Springer Verlag, 1993
- Fahrmeir, L. u. Hamerle, A.:** Multivariate statistische Verfahren, de Gruyter, Berlin, New York 1984
- Gaensslen, H./Schubö, W.:** Einfache und komplexe statistische Analyse, UTB 274, München, Basel 1973
- Harrison, M.J./Mc,Cabe, B.P.:** A Test of Heteroscedasticity Based on Ordinary Least Squares Residuals, in: Journal of the American Statistical Association 1979, S. 494-499
- Hartung/Elpelt:** Multivariate Statistik, 1984, S. 128 ff.
- Holm, Kurt:** Das Allgemeine Lineare Modell, in Holm: Die Befragung 6, UTB 436, Francke Verlag, München 1979
- Holm, Kurt:** Lineare multiple Regression und Pfadanalyse, in Holm: Befragung 5, UTB 435, München, 1977
- Kurth, Horst E.H.:** Fortran-Programm zur Lösung von Coleman- Modellen, in Holm (Hrsg.): Die Befragung 5, Franke, UTB 435, München, 1977
- Levy, K.J.:** An empirical comparison of the z-variance and Box-Scheffé tests for homogeneity of variance, in: Psychometrika, 1975, S. 519-524
- Levy, K.J.:** A Monte Carlo Study of Analysis of Covariance under Violations of the Assumptions of Normality and Equal Regression Slopes, in: EPM 1980, S. 835-846
- Levy, K.J.:** Some multiple Range Tests for Variances, in: EPM 1975, S. 599-604
- Maddala G. S.:** Limited-dependent and qualitative variables in econometrics, Cambridge University Press, 1983.
- O'Brien, R.G.:** Robust Techniques for Testing Heterogeneity of Variance effects in Factorial Designs, in: Psychometrika 1978, S. 327-341
- O'Brien, R.G.;** M.K. Kaiser: Manova method for analyzing repeates measures design: An extensive primer, in: Psychological Bulletin, 1985, Vol. 97, S316 - 333
- Overall, J.E./Woodward, A.J.:** A Simple Test for Heterogeneity of Variance in Complex Factorial Designs, in: Psychometrika 1974, S.311-318
- Rochel, H.:** Planung und Auswertung von Untersuchungen im Rahmen des allgemeinen linearen Modells, Springer Verlag, Berlin, Heidelberg, 1983
- Searle, S. R.:** Linear Models For Unbalanced Data, John Wiley, New York 1987
- Stumpf, Horst:** Das Coleman-Verfahren in Holm (Hrsg.): Die Befragung 5, Francke, UTB 435, München 1977
- Winer, B.J.:** Statistical Principles in Experimental Designs. 2. Aufl., New York 1971
- Winer, B.J., Brown, D.R. and Michels, K.M.:** Statistical Principles in Experimental Design, New York 1991