



# ALM

## Allgemeines Lineares Modell

### Teil 1

Regressionsanalyse  
Varianzanalyse  
Kovarianzanalyse  
Diskriminanzanalyse

**Kurt Holm**

Almo Statistik-System  
[www.almo-statistik.de](http://www.almo-statistik.de)  
[holm@almo-statistik.de](mailto:holm@almo-statistik.de)  
[kurt.holm@jku.at](mailto:kurt.holm@jku.at)

2022

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

Vergleich zu Version von 2019: Nur kleinere Veränderungen und Korrekturen

Der vorliegende Text ist eine überarbeitete Version der 1. Hälfte des Almo-Handbuchs „P20 Das Allgemeine Lineare Modell“. Die 2. Hälfte ist im eigenständigen Dokument "Allgemeines Lineares Modell II.PDF" (Dokument 13a) zusammengefasst. Auch dieses kann heruntergeladen werden. Es enthält insbesondere folgende Themen: (1) Logit- und Probit-Analyse (Kleinste-Quadrate-Schätzung; die ML-Schätzung ist im Dokument 8 enthalten), (2) Multivariate Analysen, (3) Hierarchische Regression, (4) Messwiederholungen, (5) Nichtlineare Regression mit dem ALM, (6) Johann Bacher: Homogenitäts-Tests  
Im Text wird häufig auf das Dokument **P0** Bezug genommen. Dabei handelt es sich um das Almo-Dokument "Arbeiten mit Almo.PDF" (Dokument 0).

## Weitere Almo-Dokumente

Die folgenden Dokumente können alle von der Handbuchseite in <http://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
- 9a. **Bootstrap bei Logit- und Probitanalyse.pdf** (0,8 MB)
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.pdf** (1 MB)
14. **Ereignisanalyse: Sterbetafel-Methode, Kaplan-Meier-Schätzer, Cox-Regression.PDF** (1,5 MB)
15. **Faktorenanalyse.PDF** (1,6 MB)
- 15a. **Bootstrap bei Faktorenanalyse.PDF** (1,7 MB)
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 DAS ALLGEMEINE LINEARE MODELL .....	5
P20.1 Definition des allgemeinen linearen Modells .....	5
P20.2 Die Submodelle des allgemeinen linearen Modells .....	5
P20.3 Auflösung der nominalen Variablen in Dummy-Variable.....	6
P20.4 Die Auflösung der Interaktionen in Dummy-Variable .....	7
P20.5 Mehrere abhängige Variable: Die multivariate Analyse.....	7
P20.6 Die Ergebnisse des allgemeinen linearen Modells.....	8
P20.6.1 Erklärte Streuungen .....	9
P20.6.2 Die Regressionskoeffizienten des allgemeinen linearen Modells .....	10
P20.6.3 Die Korrelationskoeffizienten des allgemeinen linearen Modells.....	10
P20.6.4 Signifikanzkoeffizienten.....	12
P20.6.5 Die Haupt- und Interaktionseffekte .....	13
P20.6.5.1 Inhaltliche Interpretation der Effekte .....	14
P20.6.6 Paarweise Vergleiche (Kontraste) zwischen den Haupteffekten.....	17
P20.6.6.1 Randmittel .....	18
P20.6.7 Die Standardisierung der Daten .....	23
P20.6.8 Ergebnisse bei abhängiger nominaler Variabler	
Diskriminanzanalyse, lineare Wahrscheinlichkeitsanalyse .....	23
P20.6.8.1 Einschränkungen zur linearen Wahrscheinlichkeitsanalyse.....	25
P20.6.9 <b>Hermann Denz</b> : Die Einbeziehung ordinaler Variablen .....	27
P20.7 Die Schätz-Verfahren des Allgemeinen Linearen Modells .....	28
P20.7.1 "Fitting constants I": Die gruppenweise hierarchische Auspartiellierung.....	29
P20.7.1.1 Fitting constants II (SS-Typ II) .....	33
P20.7.2 Das sequentielle Verfahren: Die variablenweise hierarchische Auspartiellierung ..	38
P20.7.2.3 Sonderprogramme und Standard-Programm-Maske Prog20mo .....	40
P20.7.3 Das Verfahren der "weighted squares of means" (SS Typ III).....	42
P20.7.3.1 Der Kalkül des Verfahrens der "weighted squares of means".....	43
P20.7.4 Vergleich der Verfahren .....	45
P20.7.4.1 Vergleich mit SAS und SPSS.....	48
P20.7.5 Berechnung der Effekte .....	49
P20.7.6 Berechnung der paarweisen Vergleiche .....	55
P20.7.7 Leere Zellen und lineare Abhängigkeit .....	56

P20.7.7.1	Almo eliminiert Kovariate wegen linearer Abhängigkeit .....	56
P20.7.7.2	Almo eliminiert Haupt-Dummies wegen linearer Abhängigkeit .....	58
P20.7.7.3	Almo eliminiert Interaktions-Dummies wegen linearer Abhängigkeit. Das Problem der "leeren Zellen" .....	58
P20.7.8	Weitere Eigenschaften von Programm 20 .....	67
P20.8	Die Eingabe in Programm 20 .....	68
P20.8.0	Eingabe in Maskenprogramm Prog20mx .....	68
P20.8.0.1	Erläuterung zu den Eingabe-Boxen von Prog20mx.Msk .....	75
P20.8.1	Eingabe in Maskenprogramm mit Optionen Prog20mo .....	83
P20.8.1.1	Erläuterungen zu den Eingabe-Boxen von Maskenprogramm Prog20mo.Msk .....	86
P20.9	Ausgabe der Ergebnisse .....	118
P20.9.1	Ausgabe bei Varianzanalyse .....	118
P20.9.1.0	Randmittel .....	134
P20.9.1.1	"Wechselnde Werte" für die Interaktionseffekte bei "fitting constants I+II" ..	144
P20.9.1.2	Verschiedene Ergebnisse für Abweichungs-Quadratsummen und Korrelationsmatrix .....	145
P20.9.2	Ausgabe bei Regressionsanalyse .....	145
P20.9.2.2	Maskenprogramm: 2D- Streudiagramm und Regressionsgerade für eine Analyse mit einer unabhängigen Variablen: Prog02mb .....	149
P20.9.2.3	Maskenprogramm: 3D- Streudiagramm und Regressionsebene für eine Analys mit 2 unabhängigen Variablen: Prog02mc .....	155
P20.9.3	Ausgabe bei Kovarianzanalyse .....	157
P20.9.3.1	Prognosewerte und Residuen .....	163
P20.9.3.2	Gewichtete Kleinste-Quadrate .....	168
Anhang:	Kovarianzanalyse mit SPSS und Almo - ein Vergleich .....	169
Schlagwortverzeichnis .....		187
Literatur zum Allgemeinen Linearen Modell .....		188

## P20 Das allgemeine lineare Modell

Siehe auch unsere ausführliche Darstellung des allgemeinen linearen Modells im Almo-Dokument Nr. 15 "Statistische Datenanalyse II", Abschnitt P45.15.

### P20.1 Definition des allgemeinen linearen Modells

Das allgemeine lineare Modell ist ein auf Variable beliebigen Messniveaus verallgemeinertes Regressionsverfahren. Es analysiert den Zusammenhang zwischen einer oder mehreren unabhängigen Variablen beliebigen Messniveaus und einer oder mehreren abhängigen Variablen ebenfalls beliebigen Messniveaus. So können sich beispielsweise auf Seiten der unabhängigen Variablen nominale, ordinale und quantitative befinden - während die abhängige eine nominal-polytome Variable ist. Das allgemeine lineare Modell umfasst als Submodell die klassischen Verfahren der Regressionsanalyse, der Varianzanalyse, der Kovarianzanalyse und der Diskriminanzanalyse - und zwar jeweils in deren univariater als auch multivariater Version.

<b>unabhängige Variable</b>	<b>abhängige Variable</b>
nominale	nominale
quantitative, ordinale (eingeschränkt)	quantitative
gemischt	

Zur Einführung von ordinalen Variablen werden wir noch eine Einschränkung vornehmen müssen (siehe Abschnitt P20.6.9.1). Auch bei den abhängigen Variablen werden einige Einschränkungen notwendig sein (siehe Abschnitt P20.8.).

Programm 20 wurde von Kurt Holm geschrieben.

### P20.2 Die Submodelle des allgemeinen linearen Modells

Zur Bezeichnung dieser Submodelle wollen wir folgende Begriffe verwenden.

Haben wir nur eine abhängige Variable, dann sprechen wir von einer „univariaten“ Analyse. Haben wir mehrere abhängige Variable, dann sprechen wir von einer "multivariaten" Analyse. Zu beachten ist hierbei, dass, wenn wir eine abhängige nominal-polytome Variable mit w Ausprägungen haben, diese in w Dummy-Variable aufgelöst wird. In diesem Falle haben wir es also auch mit einer multivariaten Analyse zu tun.

**Ordinale** Variable sind bei der Kategorie "quantitativ" eingeordnet. Der spezifische Charakter der ordinalen Variablen wird bei der Ermittlung der Streuungsmatrix berücksichtigt, auf die der Regressions-Kalkül des allgemeinen linearen Modells angewendet wird. Siehe P20.6.9.

So erhalten wir nun folgende 12 Submodelle des allgemeinen linearen Modells:

		abhängige Variable			
		eine quantitative	mehrere quantitative	eine nominal dichotome	eine nominal polytome
unabhängige Variable	quantitativ	1	2	3	4
	ordinal				
	nominal				
	gemischt				

unabhängige Variable	nominal	5	6	7	8
	quantitativ + nominal	9	10	11	12

- Modell 1: Univariate Regressionsanalyse
- Modell 2: Multivariate Regressionsanalyse
- Modelle 3 + 4 + 11 + 12: Diskriminanzanalyse (lineare Wahrscheinlichkeitsanalyse)
- Modell 5: Univariate Varianzanalyse
- Modell 6: Multivariate Varianzanalyse
- Modell 7 + 8: Zwei- und mehrdimensionale Tabellenanalyse
- Modell 9: Univariate Kovarianzanalyse
- Modell 10: Multivariate Kovarianzanalyse

**Logit-Modell:** Für die Submodelle 3,4,7,8,11,12 kann nach einer entsprechenden Transformation der abhängigen Variablen ein "Logit"-Modell (als Kleinste-Quadrate-Schätzung) gerechnet werden. Eher üblich ist eine Maximum-Likelihood-Schätzung. Siehe dazu Almo-Dokument Nr. 9 „Logitanalyse“.

**Probit-Modell:** Für die Submodelle 3,7,11 mit dichotomen Variablen als abhängige Variable kann ein Probit-Modell (als Kleinste-Quadrate-Schätzung) gerechnet werden. Ebenfalls enthalten in Almo-Dokument Nr. 9 „Logitanalyse“.

### P20.3 Auflösung der nominalen Variablen in Dummy-Variable

Wenn nominale Variable als unabhängige oder abhängige Variable in die Analyse eingeführt werden, dann werden sie in so genannte Dummy-Variable aufgelöst. Siehe dazu auch Bortz, Kap. 14.1.

Die nominale Variable sei "Beruf" mit den Ausprägungen Arbeiter Angestellter Beamter Selbständiger. Sie wird in 4 Dummy-Variable a1 , a2 , a3 , a4 aufgelöst. Dafür gibt es in Almo 2 Verfahren:

- (1) die 0,1-Kodierung und
- (2) die 0,1,-1 -Kodierung (auch "Effektkodierung" genannt)

	0,1 - Kodierung Dummy-Variable				0,1,-1-Kodierung Dummy-Variable			
	a1	a2	a3	a4	a1	a2	a3	a4
Arbeiter	1	0	0	0	1	0	0	-1
Angestellter	0	1	0	0	0	1	0	-1
Beamter	0	0	1	0	0	0	1	-1
Selbständiger	0	0	0	1	-1	-1	-1	-1

Eine Untersuchungsperson sei Arbeiter. Wir müssen ihr nun in den 4 Nominaldummies Werte zuweisen. Bei der 0,1-Kodierung erhält sie in der Nominaldummy a1 den Wert 1. In den Nominaldummies a2, a3, a4 erhält sie den Wert 0. Beim 0,1,-1 Kodierungsverfahren wird zusätzlich die letzte Dummy a4 auf -1 gesetzt. Ebenso wird die letzte Ausprägung, in unserem Beispiel "Selbständige" durchgehend auf -1 gesetzt. Die Auflösung der nominalen Variablen in Dummy-Variable wird automatisch von unserem Computer-Programm besorgt.

### Redundante Dummies

Wenn wir von einem Probanden wissen, dass er nicht Arbeiter ist und auch nicht Angestellter oder Beamter, dann muss er Selbständiger sein. Bei der 0-1-Kodierung ist also eine Dummy redundant. Im Kalkül des ALM werden nur die nicht-redundanten Dummies verwendet. Das ist notwendig, da sonst *lineare Abhängigkeiten* auftreten, die einen Abbruch der Berechnungen erzwingen würden. Im Almo (wie in fast allen Statistikprogrammen) wird deswegen standardmäßig die letzte Dummy ausgeschlossen.

### P20.4 Die Auflösung der Interaktionen in Dummy-Variable

Wir wollen annehmen, wir hätten nach dem Beruf eine zweite unabhängige nominale Variable, die Schulbildung, mit folgenden Ausprägungen und folgenden Dummy-Variablen  $b_1, b_2, b_3$ .

	0,1 - Kodierung			0,1,-1-Kodierung		
	$b_1$	$b_2$	$b_3$	$b_1$	$b_2$	$b_3$
einfache Schulbildung	1	0	0	1	0	-1
höhere Schulbildung	0	1	0	0	1	-1
Universität	0	0	1	-1	-1	-1

Wenn wir nun die Interaktion zwischen Beruf und Schulbildung in Dummy-Variable auflösen wollen, dann müssen wir zunächst die beiden Sätze von Dummy-Variablen gegeneinander tabellieren. Auf diese Weise entstehen die "multiplikativen Dummies" der Interaktion AB.

	$b_1$	$b_2$	$b_3$
$a_1$	$a_1b_1$	$a_1b_2$	$a_1b_3$
$a_2$	$a_2b_1$	$a_2b_2$	$a_2b_3$
$a_3$	$a_3b_1$	$a_3b_2$	$a_3b_3$
$a_4$	$a_4b_1$	$a_4b_2$	$a_4b_3$

Die Kodierungszahl der jeweiligen multiplikativen Dummy  $a_i b_j$  erhalten wir sehr einfach aus der Multiplikation der Kodierungszahl der Nominaldummies  $a_i$  und  $b_j$ , also  $a_i * b_j$ . Das gilt für beide Kodierungsverfahren. Betrachten wir ein Beispiel: Ein Befragter sei Arbeiter und besitze eine einfache Schulbildung. Dann hat er beim 0,1-Kodierungsverfahren in der multiplikativen Dummy  $a_1 b_1$  den Wert 1 und in allen anderen den Wert 0.

In entsprechender Weise werden auch die multiplikativen Dummies von 3er-Interaktionen und Interaktionen höherer Ordnung gebildet.

### Redundante Interaktionsdummies

In Almo werden bei der 0-1-Kodierung nur die nicht-redundanten Interaktionsdummies verwendet. Sie entstehen aus der Kombination der nicht-redundanten nominalen Dummies.

### P20.5 Mehrere abhängige Variable: Die multivariate Analyse

Die multivariate Analyse wird im vorliegenden Teil 1 nicht behandelt. Bei der multivariaten Analyse ist die abhängige Variable eine Menge von mehreren Variablen. Die unabhängigen Variablen können Nominal- und multiplikative Dummies und/oder quantitative Variable sein.

## **P20.6 Die Ergebnisse des allgemeinen linearen Modells**

Die Koeffizienten, die uns das allgemeine lineare Modell liefert, sind im Wesentlichen:

- 1) die erklärten Streuungen
- 2) die Regressionskoeffizienten
- 3) die Effekte (Haupt- und Interaktionseffekte)
- 4) die Korrelationskoeffizienten
- 5) die Signifikanzkoeffizienten

## P20.6.1 Erklärte Streuungen

Unser Computer-Programm gibt die

- 1) insgesamt durch alle unabhängigen Variablen erklärten Streuungen an,
- 2) die durch Variablengruppen erklärte Streuung, z.B. die durch die Gruppe der quantitativen unabhängigen Variablen erklärte Streuung,
- 3) die durch einzelne Variable erklärte Streuung, z.B. die durch eine einzelne quantitative unabhängige Variable erklärte Streuung.

Zu 2.) Die Variablengruppe, deren erklärte Streuung berechnet und ausgegeben wird, kann sein

- die Gruppe der quantitativen unabhängigen Variablen und die Gruppe der nominalen unabhängigen Variablen.
- Die Variablengruppe kann aber auch durch den Forscher definiert werden. Bei einer Untersuchung zu den Ursachen einer bestimmten Leistung, fasst der Forscher die in der Person begründeten Variablen und die in der sozialen Umwelt begründeten Variablen zu je einer Gruppe zusammen. Deren pauschale erklärte Streuung will er in je einer Zahl ausdrücken. Wie der Benutzer Variablengruppen bilden kann, wird in P20.8.0.1, Unterabschnitt "Variablenhierarchie und gleichrangige Variablengruppen" beschrieben

Bei der Kovarianzanalyse erfolgt beispielsweise folgende Ausgabe (gekürzt):

Die abhängige, zu erklärende Variable sei die Leistung in einem Test. Sie ist quantitativ.

Die unabhängigen quantitativen Variablen sind (1) das Alter der Testperson, (2) ihr Einkommen und (3) die Zahl ihrer zu versorgenden Kinder. Die unabhängigen nominalen Variablen sind das Geschlecht und der Beruf der Testperson.

<b>Streuungsquelle</b>	<b>Streuung</b>
-----	-----
<b>Gesamtstreuung</b>	<b>222.1967</b>
<b>Fehlerstreuung</b>	<b>177.5288</b>
<b>alle unabhängigen Variablen zusammen</b>	<b>44.6680</b>
<b>quantitative Variable zusammen</b>	<b>2.8096</b>
<b>nominale Variable und ihre Interaktionen zusammen</b>	<b>38.5882</b>
<b>quantitative Variable:</b>	
<b>V6 Alter</b>	<b>0.3339</b>
<b>V7 Einkommen</b>	<b>0.0248</b>
<b>V8 Kinderzahl</b>	<b>2.4361</b>
<b>nominale Variable</b>	
<b>V1 Geschlecht</b>	<b>24.4552</b>
<b>V3 Beruf</b>	<b>11.6908</b>
<b>Interaktion Geschlecht*Beruf</b>	<b>4.6516</b>

Zu beachten ist hierbei nun:

- 1) Der Benutzer bestimmt beim Maskenprogramm durch entsprechende Auswahl in der Eingabe-Box "Streuungsmatrix", ob die ausgegebenen Streuungen,
  - (a) Kreuzprodukte sind
  - (b) oder Abweichungs-Quadratsummen
  - (c) oder Varianzen
  - (d) oder Standardabweichungen.

- 2) Wenn die unabhängigen Variablen untereinander korrelieren (und das ist der Normalfall), dann sind die erklärten Streuungen nicht additiv.

Beispiel: Die Fehlerstreuung plus die durch alle unabhängigen Variable zusammen erklärte Streuung ist nicht gleich der Gesamtstreuung. Das wäre nur der Fall, wenn alle unabhängigen Variablen unkorreliert wären.

Beispiel: Die Summe der durch die einzelnen unabhängigen quantitativen Variablen erklärten Streuungen ist kleiner als die durch die Gruppe der unabhängigen quantitativen Variablen insgesamt erklärte Streuung. Auch das wäre nur der Fall, wenn alle quantitativen unabhängigen Variablen unkorreliert wären.

- 3) Generalisierte Streuungen (die in diesem Beispiel nicht vorkommen): Sind mehrere abhängige quantitative Variable vorhanden oder ist die abhängige Variable eine nominale, die im Programm automatisch in Dummies aufgelöst wird, dann werden "generalisierte" Streuungen gerechnet (siehe dazu in Teil II, Abschnitt P20.9.4.1).

### **P20.6.2 Die Regressionskoeffizienten des allgemeinen linearen Modells**

Die Regressionskoeffizienten, die im Rahmen des allgemeinen linearen Modells berechnet werden, sind mit einer Ausnahme identisch mit entsprechenden Koeffizienten, die im Rahmen der klassischen Regressionsanalyse, der klassischen Kovarianzanalyse und der klassischen Diskriminanzanalyse ermittelt werden.

Almo bringt beispielsweise folgende Ausgabe (gekürzt):

<b>Variable</b>	<b>Regress. koeff.</b>
V6 Alter	0.0359
V7 Einkommen	-0.0089
V8 Kinderzahl	-0.2728

Im Falle des Submodells 3 (einfache Diskriminanzanalyse mit einer nominal-dichotomen Variablen als abhängiger Variable) sind sie, um eine Proportionalitätskonstante  $k$  multipliziert, identisch mit den kanonischen Diskriminanzkoeffizienten.

Die Ausnahme ist das Submodell 4 (Diskriminanzanalyse mit einer nominal-polytomen abhängigen Variablen). Hier sind die Regressions-koeffizienten des allgemeinen linearen Modells nicht identisch mit den Diskriminanzkoeffizienten der kanonischen Diskriminanzanalyse. Siehe hierzu auch P20.9.5.2 und unsere Darstellung zur kanonischen Diskriminanzanalyse im Almo-Dokument Nr. 4.

### **P20.6.3 Die Korrelationskoeffizienten des allgemeinen linearen Modells**

Die Korrelationskoeffizienten des allgemeinen linearen Modells sind PRE-Koeffizienten (siehe Costner, 1965, S.344), d.h. Koeffizienten, die die "proportional reduction of error", die proportionale Reduktion des Fehlers in der abhängigen Variablen angeben, die eine einzelne oder eine Untermenge von unabhängigen Variablen bewirkt. Betrachten wir eine Untermenge  $x_1 \dots x_k$  von unabhängigen Variablen. Wir führen nun zwei Analysen durch, eine Analyse mit allen unabhängigen Variablen und eine Analyse, bei der die Untermenge  $x_1 \dots x_k$  aus der Reihe der unabhängigen Variablen herausgenommen wurde. Wir erhalten dabei zwei verschiedene Fehlerstreuungen, die in der abhängigen Variablen  $y$  verbleiben. Die Reduktion, die durch  $x_1 \dots x_k$  in der Fehlerstreuung von  $y$  verursacht wird, ist selbstverständlich.

$$\left| \begin{array}{c} \text{Fehlerstreuung in einer} \\ \text{Analyse ohne } x_i \dots x_k \end{array} \right| \quad \text{minus} \quad \left| \begin{array}{c} \text{Fehlerstreuung in einer} \\ \text{Analyse mit } x_i \dots x_k \end{array} \right|$$

Die proportionale Fehlerreduktion ergibt sich dann dadurch, dass wir die Fehlerreduktion messen an der Fehlerstreuung in einer Analyse ohne  $x_i \dots x_k$ . Wir erhalten somit folgende Formel für den PRE- Koeffizienten:

$$\text{PRE}_{x_i \dots x_k}^2 = \frac{(\text{Fehler ohne } x_i \dots x_k) - (\text{Fehler mit } x_i \dots x_k)}{\text{Fehler ohne } x_i \dots x_k}$$

Die sehr sinnvolle inhaltliche Interpretation des  $\text{PRE}^2$ -Koeffizienten lautet: Wenn wir den unabhängigen Variablen die Untermenge  $x_i \dots x_k$  hinzufügen, dann reduziert sich die Fehlerstreuung in der abhängigen Variablen um einen Anteil von so und so viel.

Wird aus  $\text{PRE}^2$  die Wurzel gezogen, dann erhält man den üblichen PRE-Korrelationskoeffizienten.

Wir können nun verschiedene Konstellationen unterscheiden:

- 1) Die Untermenge  $x_i \dots x_k$  umfasst alle unabhängigen Variablen. Dann ist der PRE-Koeffizient der *multiple* Korrelationskoeffizient  $R_{\text{mult}}$ .
- 2) Die Untermenge  $x_i \dots x_k$  ist tatsächlich eine Untermenge aus einer großen Anzahl unabhängiger Variablen. Dann ist der PRE-Koeffizient ein *partieller multipler* Korrelationskoeffizient.
- 3) Die Untermenge  $x_i \dots x_k$  besteht aus nur einer unabhängigen Variablen. Die Gesamtzahl der unabhängigen Variablen ist zwei oder mehr. Dann ist der PRE-Koeffizient ein *partieller* Korrelationskoeffizient.
- 4) Die Untermenge  $x_i \dots x_k$  besteht aus nur einer unabhängigen Variablen. Diese ist auch insgesamt die einzige unabhängige Variable. Dann ist der PRE-Koeffizient der gewöhnliche Korrelationskoeffizient  $r$ .

Almo gibt bei der Kovarianzanalyse beispielsweise folgende PRE-Koeffizienten aus (gekürzt):

Streuungsquelle	Korrelations- koeffizient
-----	
Gesamtstreuung	
Fehlerstreuung	
alle unabhängigen Variablen zusammen	0.4484
quantitative Variable zusammen	0.1248
nominale Variable und ihre Interaktionen zusammen	0.4226
quantitative Variable:	
V6 Alter	0.0433
V7 Einkommen	-0.0118
V8 Kinderzahl	-0.1163
nominale Variable	
V1 Geschlecht	0.3480

Haben wir zwei oder mehr abhängige Variable (seien diese quantitativ oder die Dummies einer nominalen Variablen), dann sind in die PRE-Formel "generalisierte" Fehlerstreuungen einzusetzen (siehe in Teil II, Abschnitt P20.9.4.1). Die oben unter 1) und 4) angegebenen Koeffizienten sind symmetrisch, d.h. wir können die unabhängigen zu den abhängigen und die abhängigen zu den unabhängigen Variablen machen. Der PRE-Koeffizient bleibt derselbe. Interessant ist dies vor allem im Hinblick auf den Koeffizienten nach 1). Dies ist die Konstellation mit mehreren unabhängigen und einer abhängigen Variablen. Bei einer Umdrehung haben wir dann eine unabhängige Variable und mehrere abhängige Variablen. In die PRE-Formel sind dann generalisierte Streuungen einzusetzen. Trotzdem entsteht derselbe Wert.

Der aus der statistischen Literatur bekannte **Phi-Koeffizient** entsteht als PRE-Koeffizient im Rahmen des Submodells 7 des allgemeinen linearen Modells - für den Sonderfall, dass nur eine dichotome unabhängige und eine dichotome abhängige Variable gegeben ist. Der **punktbiseriale Korrelationskoeffizient** entsteht im Rahmen des Submodells 3 - für den Sonderfall, dass nur eine quantitative unabhängige Variable gegeben ist und die abhängige Variable dichotom ist. Die aus der Varianz- und Kovarianzanalyse bekannten **Eta-Koeffizienten** entstehen als PRE-Koeffizienten im Rahmen unserer Submodelle 5 und 9.

Die aus der linearen Regressionsanalyse abgeleiteten gewöhnlichen, partiellen, partiellen multiplen und multiplen **Produkt-Moment-Korrelationskoeffizienten** entstehen im Rahmen unseres Submodells 1.

Wie Denz in Abschnitt P20.6.8 zeigt, steht auch das **Kendall'sche tau-b** im Rahmen des allgemeinen linearen Modells. Es ist ein PRE-Koeffizient.

In den anderen nicht genannten Submodellen werden dann jeweils gemäß der PRE-Formel Korrelationskoeffizienten gebildet, die in der statistischen Literatur nicht als eigenständige Koeffizienten entwickelt wurden.

Die in der statistischen Literatur angeführten Korrelationskoeffizienten sind zu einem erheblichen Teil nach zwar plausiblen, aber sehr verschiedenen Prinzipien konstruiert. Das allgemeine lineare Modell liefert uns ein einheitliches Prinzip, um nicht zu sagen, eine Theorie für eine Vielzahl von Korrelationskoeffizienten.

Bei quantitativen Variablen kann den gewöhnlichen und den partiellen PRE-Korrelationskoeffizienten noch ein Vorzeichen vorgesetzt werden, indem man das Vorzeichen des Regressionskoeffizienten überträgt. Aus der PRE-Formel ergibt sich nur ein positives Vorzeichen. Es ist jedoch üblich, einen gegenläufigen Zusammenhang (der aus dem Vorzeichen des Regressionskoeffizienten ersichtlich wird) durch ein Minuszeichen zu charakterisieren. Siehe hierzu auch Holm, 1977, Abschnitt 4.9.2

#### **P20.6.4 Signifikanzkoeffizienten**

Wir verwenden in unserem Computer-Programm ausschließlich den F-Wert und den t-Wert. Dabei geben wir noch die Signifikanz  $p$  und  $(1-p)*100$  an (z.B. 95 % Signifikanzniveau). Der F-Wert geht wenn mehrere abhängige Variable vorliegen in eine multivariate Version über (siehe Holm, 1979, Abschnitt 9.4 und hier Abschnitt P20.9.4.1). Also gibt auch noch die **Teststärke** aus. Wir werden diesen Begriff später in P20.9.1 erläutern.

Almo liefert bei der Kovarianzanalyse beispielsweise folgende Ausgabe (gekürzt):

Streuungsquelle	F-Wert	df	Signifikanz p	(1-p)100	Test- stärke
-----					
Gesamtstreuung					
Fehlerstreuung		52			
alle unabhängigen Variablen zusammen	1.6355	8	0.1371	86.2866	0.6544
quantitative Variable zusammen	0.2743	3	0.8442	15.5836	0.0990
nominale Variable und ihre Interaktionen zusammen	2.2606	5	0.0613	93.8725	0.6865
quantitative Variable:					
V6 Alter	0.0978	1	0.757	24.3272	0.0608
V7 Einkommen	0.0073	1	0.930	6.9678	0.0508
V8 Kinderzahl	0.7136	1	0.402	59.8012	0.1317
nominale Variable					
V1 Geschlecht	7.1632	1	0.0097	99.0313	0.7475
V3 Beruf	1.7122	2	0.1888	81.1233	0.3437
Interaktion Geschlecht*Beruf	0.6813	2	0.5149	48.5118	0.1586

### P20.6.5 Die Haupt- und Interaktionseffekte

Effekte sind, vorläufig definiert, die Regressionskoeffizienten der Dummy-Variablen. Wir werden auf die Effekte nochmals in Abschnitt P20.7.5 eingehen. So genannte "Effekte" (Haupt- und Interaktionseffekte) treten nur auf, wenn sich auf Seiten der unabhängigen Variablen auch nominale Variable befinden, also bei Analysen vom Typ der Varianz- und Kovarianzanalyse. Dabei ist es gleichgültig, ob sich auf Seiten der abhängigen Variablen quantitative oder nominale befinden.

Wir wollen ein Beispiel betrachten:

Auf Seiten der unabhängigen Variablen befinden sich:

- 1) die Konstante t
- 2) die nominalen Variablen A mit den Ausprägungen A1, A2  
B mit den Ausprägungen B1, B2, B3
- 3) die Interaktionsvariable AB
- 4) die quantitativen Variablen (=Kovariaten)  $x_1, x_2, x_3$

Die unabhängigen nominalen Variablen bilden folgende Ausprägungskombinationen:

	A <sub>1</sub>	A <sub>2</sub>
B <sub>1</sub>		
B <sub>2</sub>		
B <sub>3</sub>		

**Zellenbesetzung:** Wenn wir im Folgenden von "Zellenbesetzung" sprechen, dann meinen wir damit die Zahl der Untersuchungseinheiten, die sich in den Zellen der obigen Tabelle befinden. Sehr wesentlich ist, (1) ob alle Zellen mit gleich vielen Untersuchungseinheiten besetzt sind oder - was der Normalfall sein dürfte - mit ungleich vielen Untersuchungseinheiten

Nach Auflösung der nominalen und der interaktiven Variablen in 0-1 kodierte Nominal- bzw. multiplikative Dummies erhalten wir dann folgende Gleichung:

$$\begin{aligned}
 (1) \ y' = & \tau * t + && \text{(Konstanteneffekt)} \\
 & \alpha_1 * a_1 + \alpha_2 * a_2 + && \text{(Haupteffekte von A)} \\
 & \beta_1 * b_1 + \beta_2 * b_2 + \beta_3 * b_3 + && \text{(Haupteffekte von B)} \\
 & \alpha\beta_{11} * a_1 b_1 + \alpha\beta_{12} * a_1 b_2 + && \text{(Interaktionseffekte AB)} \\
 & \alpha\beta_{13} * a_1 b_3 + \\
 & \alpha\beta_{21} * a_2 b_1 + \alpha\beta_{22} * a_2 b_2 + \\
 & \alpha\beta_{23} * a_2 b_3 + \\
 & \gamma_1 * x_1 + \gamma_2 * x_2 && \text{(Kovariate)}
 \end{aligned}$$

üblich ist auch die kürzere Schreibweise:

$$(1a) \ y' = \tau + \alpha_i + \beta_j + \alpha\beta_{ij} + \gamma_1 x_1 + \gamma_2 x_2$$

- $y'$  = die vom Modell prognostizierten Werte für die abhängige Variable
- $\alpha_i, \beta_j$  = Effekte der nominalen Variablen A und B
- $\alpha\beta_{ij}$  = Effekte der Interaktionsvariablen AB
- $a_i, b_j$  = Dummy von A bzw. B. Ist entweder 0 oder 1
- $a_i b_j$  = 0-1 Dummy von AB
- $\gamma_k$  = Regressionskoeffizient der quantitativen unabhängigen Variablen  $x_k$
- $x_k$  = unabhängige quantitative Variable

**Zur Notation:** Wir verwenden (kleine) lateinische Buchstaben, um Variable zu bezeichnen. Dabei verwenden wir a, b, c, d... für Dummies und x für Kovariate. Die "Konstantenvariable", die nur aus 1.0 besteht, ist t.

Kleine griechische Buchstaben verwenden wir, um die Effekte und Regressionskoeffizienten zu bezeichnen. So ist beispielsweise  $\alpha_1$  der Effekt für  $a_1$ ,  $\beta_3$  der Effekt für  $b_3$  und  $\gamma_2$  für  $x_2$ .  $\tau$  ist der Konstanteneffekt.

In der 2. Zeile vor Gleichung (1) stehen alle 0-1 kodierten Nominaldummies von A, also  $a_1$  und  $a_2$ . Ihnen sind ihre Haupteffekte  $\alpha_1$  und  $\alpha_2$  beigegeben.

In der 3. Zeile stehen alle 0-1 kodierten Nominaldummies von B, also  $b_1, b_2, b_3$  mit ihren Haupteffekten.

In der 4., 5., 6. Zeile stehen alle 0-1 kodierten multiplikativen Dummies der Interaktion AB, also  $a_1 b_1, a_1 b_2, \dots, a_2 b_3$  mit ihren Interaktionseffekten.

### ***P20.6.5.1 Inhaltliche Interpretation der Effekte***

Die Effekte werden so interpretiert wie Regressionskoeffizienten. D.h. sie sind Koeffizienten, die den direkten Zusammenhang zwischen einer unabhängigen Nominal- bzw. multiplikativen

Dummy und einer abhängigen Variablen angeben, wobei die Wirkung der anderen unabhängigen Variablen eliminiert ist.

Betrachten wir ein konstruiertes Beispiel: In einer Untersuchung über die Erziehungsstrenge von Eltern und deren Ursachen wurden u.a. folgende Variable gemessen:

- |                      |                  |
|----------------------|------------------|
| 1. Wohnort A,        |                  |
| Stadt                | = A <sub>1</sub> |
| Land                 | = A <sub>2</sub> |
| 2. Beruf B           |                  |
| Selbständig          | = B <sub>1</sub> |
| Bauer                | = B <sub>2</sub> |
| Arbeitnehmer         | = B <sub>3</sub> |
| 3. Einkommen         | = x <sub>1</sub> |
| 4. Lebensalter       | = x <sub>2</sub> |
| 5. Erziehungsstrenge | = y              |

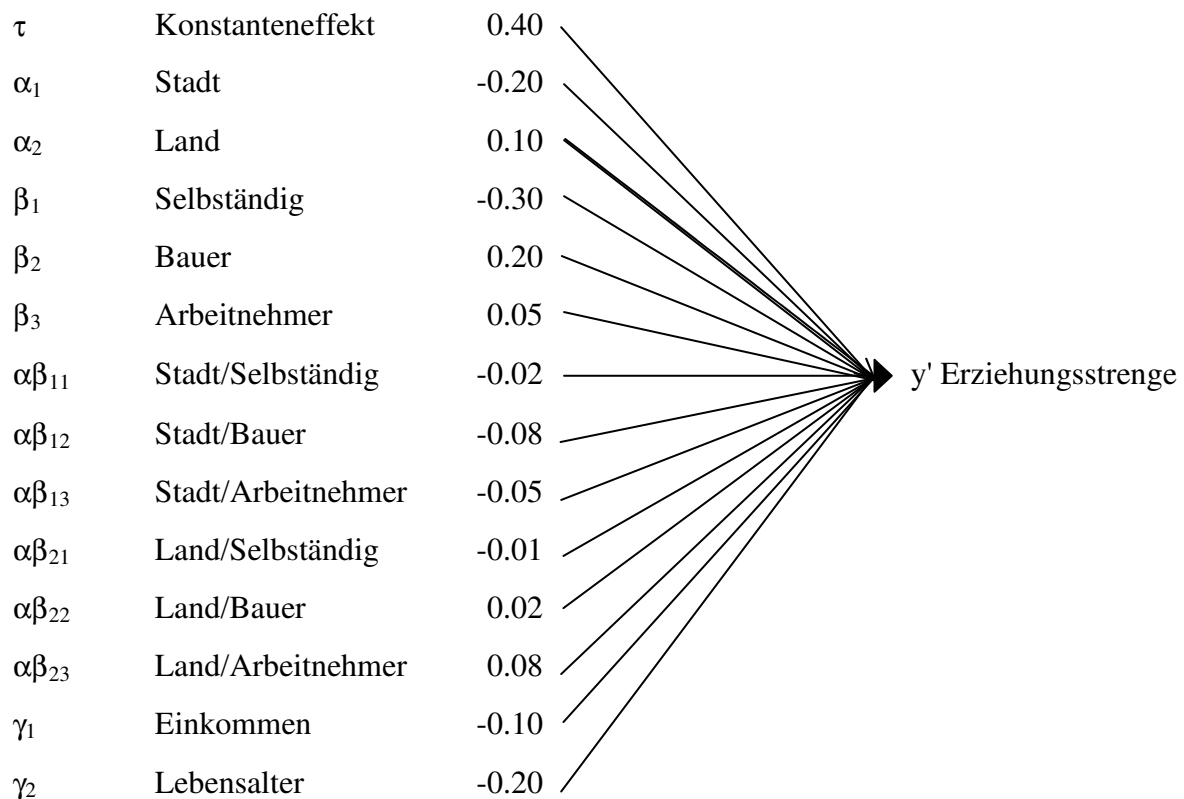
A und B sind zwei nominale unabhängige Variable. x<sub>1</sub> und x<sub>2</sub> sind quantitative unabhängige Variable (=Kovariate), y ist eine quantitative abhängige Variable.

Wir untersuchen also folgendes Modell:

$$\begin{aligned}
 (1) \ y' = & \tau * t + && \text{(Konstanteneffekt)} \\
 & \alpha_1 * a_1 + \alpha_2 * a_2 + && \text{(Haupteffekte von A)} \\
 & \beta_1 * b_1 + \beta_2 * b_2 + \beta_3 * b_3 + && \text{(Haupteffekte von B)} \\
 & \alpha\beta_{11} * a_1 b_1 + \alpha\beta_{12} * a_1 b_2 + && \text{(Interaktionseffekte AB)} \\
 & \alpha\beta_{13} * a_1 b_3 + && \\
 & \alpha\beta_{21} * a_2 b_1 + \alpha\beta_{22} * a_2 b_2 + && \\
 & \alpha\beta_{23} * a_2 b_3 + && \\
 & \gamma_1 * x_1 + \gamma_2 * x_2 && \text{(Kovariate)}
 \end{aligned}$$

y' = das ist der vom Modell prognostizierte Wert in der abhängigen Variablen

Wir erhalten aus einer Analyse nach dem Verfahren der fitting constants I folgende Ergebnisse, die wir gleich graphisch darstellen wollen.



Betrachten wir zunächst den Konstanteneffekt. Wären keine Kovariaten wirksam, dann entspricht  $\tau \cdot t$  genau dem Mittelwert der abhängigen Variablen (über alle Untersuchungseinheiten). Da Kovariate vorhanden sind, wird dieser Wert verschoben.

Die **Regressionskoeffizienten** der Kovariaten sind in der gewohnten Weise zu interpretieren: Je größer das Einkommen, umso geringer die Erziehungsstrenge. Genauer: Wird das Einkommen um eine Einheit erhöht, dann reduziert sich die Erziehungsstrenge um 0.10 Einheiten - besser wäre es zu sagen: dann prognostiziert unsere Gleichung eine um 0.10 Einheiten verringerte Erziehungsstrenge.

Wird das Lebensalter um 1 Einheit erhöht, dann reduziert sich die Erziehungsstrenge sogar um 0.20 Einheiten. Im Vergleich zum Einkommen hat also das Lebensalter einen stärker reduzierenden Einfluss. Wird ein solcher Vergleich angestellt, muss man jedoch berücksichtigen, dass Lebensalter und Einkommen in verschiedenen Einheiten gemessen werden.

Durch Standardisierung aller Variablen können die Einheiten gewissermaßen gleich gemacht werden, so dass ein Vergleich sinnvoll ist. Siehe hierzu P20.6.7 und Holm, 1977, Abschnitt 12.

Wie sind nun die **Effekte der nominalen Variablen** zu interpretieren? Betrachten wir die Nominaldummy  $a_1$  (Stadt). Sie kann in genau derselben Weise interpretiert werden. Wird diese Nominaldummy  $a_1$  um eine Einheit erhöht, dann reduziert sich die Erziehungsstrenge um 0.20. Die Besonderheit ist nun die, dass die Nominaldummy  $a_1$  nur 2 Werte besitzt, nämlich

- 0 = nicht in der Stadt wohnen,
- 1 = in der Stadt wohnen.

Wir müssen also präziser sagen: Wird die Nominaldummy  $a_1$  von 0 auf 1 erhöht, dann wird die Erziehungsstrengung um 0.20 Einheiten reduziert.

Betrachten wir die Nominaldummy  $a_2$  (Land). Wird sie von 0 (=nicht auf dem Land wohnen) auf 1 (=auf dem Land wohnen) erhöht, dann erhöht sich die Erziehungsstrengung um 0.10 Einheiten.

**Interaktionseffekte** sind in der Regel inhaltlich schwer zu interpretieren. Das formale Prinzip der Interpretation bleibt jedoch dasselbe. Betrachten wir die multiplikative Dummy  $a_2b_3$ . Dies sind Personen, die die Merkmalskombination "Land/ Arbeitnehmer" aufweisen. Wird die Dummy  $a_2b_3$  von 0(= nicht Land/Arbeitnehmer) auf 1 (= Land/Arbeitnehmer) erhöht, dann erhöht sich die Erziehungsstrengung um 0.08 Einheiten. Eine sinnvolle Interpretation entsteht in der Regel dadurch, dass wir den Interaktionseffekt vergleichen mit den entsprechenden Haupteffekten. Eine Person, die auf dem Lande wohnt und Arbeitnehmer ist, "besitzt" folgende Effekte:

- $\alpha_2 = 0.10$  d.h. die Tatsache, dass sie auf dem Land wohnt, erhöht ihre Erziehungsstrengung um 0.10 Einheiten,
- $\beta_3 = 0.05$  d.h. die Tatsache, dass sie Arbeitnehmer ist, erhöht ihre Erziehungsstrengung um (nur) 0.05 Einheiten,
- $\alpha\beta_{23} = 0.08$  d.h. die Tatsache, dass sie die Merkmalskombination Land/ Arbeitnehmer besitzt, erhöht ihre Erziehungsstrengung um 0.08.

Von einer Interaktion sprechen wir dann, wenn es alleine das Zusammentreffen bestimmter Merkmale ist, das eine Wirkung auf die abhängige Variable besitzt.

Sind Interaktionen inhaltlich nicht erklärbar, dann - so empfehlen wir - auf sie zu verzichten. Die durch sie erklärte Streuung in der abhängigen Variablen wird dann Teil der Fehlerstreuung, wodurch sich dann allerdings die Signifikanz der Haupteffekte verschlechtert.

### **Effekte und Randmittel**

Die Effekte lassen sich auch aus den Randmitteln, die im nächsten Abschnitt behandelt werden, erklären. Siehe dort.

### **Vergleich der "Effekte" aus Almo und SPSS und SAS**

Wir werden in Abschnitt P20.7.5 dieses Thema ausführlich behandeln

### **P20.6.6 Paarweise Vergleiche (Kontraste) zwischen den Haupteffekten**

Zwischen den beiden Variablen  $a_1$  und  $a_2$  ist nun - auch ohne Standardisierung - ein Vergleich möglich. Wir können unmittelbar die Differenz der beiden Haupteffekte bilden, sie beträgt 0.30 Einheiten. In der Literatur wird diese Differenz "paarweiser Kontrast" oder „paarweiser Vergleich“ genannt. Wir können dann folgende Aussagen machen: Stadt- und Landbewohner "kontrastieren" in ihrer Erziehungsstrengung um 0.30 Einheiten - wobei der mit dem Wohnort korrelierende Einfluss des Berufs, des Einkommens und des Lebensalters, also der Einfluss der anderen unabhängigen Variablen, eliminiert ist.

In dieser entsprechenden Weise können wir nun die Wirkung der Nominaldummies  $b_1$ ,  $b_2$ ,  $b_3$  interpretieren.

Es können nun 3 Vergleiche gebildet werden:

Diff.	Signifikanz
p	(1-p)100

Selbständig - Bauer: 0.50 0.01 99%  
 Selbständig - Arbeitnehmer: 0.35 0.05 95%  
 Bauer - Arbeitnehmer: 0.15 0.20 80%

ALMO gibt zusätzlich noch die Signifikanz des jeweiligen Vergleichs an. Die Differenz Bauer-Arbeitnehmer ist nicht signifikant. Am stärksten kontrastieren Selbständige und Bauern. Wir erachten die paarweisen Vergleiche als eines der wesentlichen Ergebnisse aus dem allgemeinen linearen Modell.

Der beim paarweisen Vergleich verwendete t-Test ist ein "apriori t-Test". Er wird nach einem Verfahren gerechnet, das gelegentlich mit LSD (= least significant difference) bezeichnet wird. Siehe dazu etwa Bortz 1993, S. 240, 249 oder Winer, Brown, Michels 1991, S. 140, 172.

Bei der einfaktoriellen Varianzanalyse, d.h. wenn sich nur eine unabhängige nominale Variable im Modell befindet (die 3 oder mehr Ausprägungen besitzt), gibt Almo auch einen "aposteriori" (oder "post hoc") t-Test aus. Dabei wird das Verfahren des **Scheffé-Tests** gerechnet. Die Ausgabe ist folgende (gekürzt).

**Paarweise Vergleiche (Kontraste) von Faktor A**

	Diff- erenz	LSD				Scheffe-Test		
		t-Wert	Signifikanz p	(1-p)100	Test- staerke	F-Wert	Signifikanz p	(1-p)100
A1 - A2	0.3405	0.5768	0.5658	43.42%	0.0876	0.1663	0.8454	15.46%
A1 - A3	1.2500	1.8649	0.0672	93.28%	0.4507	1.7389	0.1830	81.70%
A2 - A3	0.9095	1.5404	0.1290	87.10%	0.3289	1.1865	0.3126	68.74%

Freiheitsgrade fuer t-Wert (LSD) : 58  
 Freiheitsgrade fuer Scheffe F-Test df1: 2 dfe: 58

Der LSD-t-Test ist ein apriori-Test, der Scheffé-Test ist ein aposteriori-(oder post hoc-) Test. Zu dieser Unterscheidung siehe die oben angegebene Literatur (insbesondere Bortz, 1993).

Die Berechnung der paarweisen Vergleiche zeigen wir in P20.7.6.

**P20.6.6.1 Randmittel**

Wir rechnen mit unseren Testdaten eine Varianzanalyse mit den 3 unabhängigen nominalen Variablen

- A V1 Geschlecht**
- B V2 Wohnort**
- C V4 Schulbildung**

hinsichtlich der abhängigen Variablen V5 Leistung

Der Benutzer kann dieses Beispiel nachrechnen. Dazu muss die Programm-Maske "Prog20\_Randmittel.Alm" verwendet und entsprechend adaptiert werden. Das Programm findet man nach Klick auf den Knopf "alle Progs" am Oberrand des Almofensters.

Almo liefert (beim standardmäßig voreingestellten) Verfahren der weighted squares of means folgende *geschätzte Randmittel*.

**Tabelle 1**

Geschaetzte Randmittel  
hinsichtlich der abhaengigen Variablen Leistung

```

modellreproduzierter
Gesamtmittelwert (Konstante)    3.972222
=====
V1 Geschlecht
  A1 männlich                    3.291667
  A2 weiblich                    4.652778

V2 Wohnort
  B1 Stadt                      4.069444
  B2 Land                       3.875000

V4 Schulbildung
  C1 niedrig                    4.458333
  C2 hoch                      3.486111
=====
2-er Randmittel AB  Geschlecht*Wohnort
  A1 B1                        3.333333
  A1 B2                        3.250000
  A2 B1                        4.805556
  A2 B2                        4.500000
=====
2-er Randmittel AC  Geschlecht*Schulbildung
  A1 C1                        3.583333
  A1 C2                        3.000000
  A2 C1                        5.333333
  A2 C2                        3.972222
=====
2-er Randmittel BC  Wohnort*Schulbildung
  B1 C1                        5.000000
  B1 C2                        3.138889
  B2 C1                        3.916667
  B2 C2                        3.833333
=====
3-er Randmittel ABC  Geschlecht*Wohnort*Schulbildung
  A1 B1 C1                    4.166667
  A1 B1 C2                    2.500000
  A1 B2 C1                    3.000000
  A1 B2 C2                    3.500000
  A2 B1 C1                    5.833333
  A2 B1 C2                    3.777778
  A2 B2 C1                    4.833333
  A2 B2 C2                    4.166667
=====

```

Wie entstehen diese Randmittel ?

Almo gibt (am Beginn der Ergebnisliste) die Zellenmittel der abhängigen Variablen je Leistung aus, die bei der jeweiligen Kombination der nominalen Variablen A, B und C entstehen.

**Tabelle 2**

Zellenmittelwerte der abhängigen Variablen

			Leistung
Geschlecht	Wohnort	Schulbil	
männlich	Stadt	niedrig	4.1667
		hoch	2.5000
	Land	niedrig	3.0000
		hoch	3.5000

weiblich	Stadt	niedrig	5.8333
		hoch	3.7778
	Land	niedrig	4.8333
		hoch	4.1667
Gesamtmittel			3.8852

Mittelwert  
aus Zellenmittelwerten                      3.9722 <--- "modellreproduzierter" Gesamtmittelwert  
Konstante

Wir formen diese Tabelle 2 in folgender Weise um.

**Tabelle 3**

Geschlecht und Wohnort gegen Schulbildung  
Zellenmittelwerte und Randmittel

Geschlecht	Wohnort	Schulbildung		
		niedrig	hoch	
männlich	Stadt	4.1667	2.5000	<b>3.3334</b>
	Land	3.0000	3.5000	<b>3.2500</b>
weiblich	Stadt	5.8333	3.7778	<b>4.8056</b>
	Land	4.8333	4.1667	<b>4.5000</b>
		<b>4.4583</b>	<b>3.4861</b>	

Randmittel für AB Geschlecht\*Wohnort  
=Mittelwert aus 2 Zellenmittelwerten

Randmittel für Schulbildung  
=Mittelwert aus 4 Zellenmittelwerten

Die Werte im Inneren der Tabelle sind die Zellenmittelwerte aus Tabelle 2. Die beiden (fett gedruckten) Werte am unteren Tabellenrand sind jeweils die Mittelwerte aus den 4 Zellenmittelwerten in der Spalte *Schulbildung:niedrig* und *Schulbildung:hoch*. Beim Vergleich mit Tabelle 1 erkennt man, dass sie identisch sind mit den Randmitteln.

Die vier (fett gedruckten) Werte im hinteren Rand der Tabelle 3 sind jeweils die Mittelwerte aus den beiden Zellenmittelwerten einer Zeile. Sie sind die 4 Randmittel der 2-er Kombination AB *Geschlecht\*Wohnort*.

In entsprechender Weise können nun aus Tabelle 2 weitere Tabellen nach dem Muster von Tabelle 3 gebildet werden und dann die Randmittel berechnet werden. Diese Methode ist aber nur anwendbar, wenn nach dem (standardmäßig voreingestellten) Verfahren der "weighted squares of means" (=SS-Typ III) gerechnet wird. Die Randmittel sind bei diesem Verfahren Mittelwerte ; gebildet aus den entsprechend ausgewählten Zellenmittelwerten.

Die Randmittel ABC, also der höchsten Ordnung, sind identisch mit den Prognosewerten für die Probanden. Werden in der Programm-Maske Prognosewerte angefordert, dann gibt Almo in unserem Beispiel aus:

| tatsächlicher Wert | prognostizierter Wert | Residuen |

Datensatz	in der abhangigen	in der abhangigen	(Differenz)
	Variablen V5 Leistung	Variablen V5 Leistung	V5 Leistung
1	4.00000	4.16666	-0.1667
2	5.00000	4.16666	0.83334
3	4.00000	2.50000	1.50000
.	.	.	.
.	.	.	.

Der Proband 1 hat in den Variablen A Geschlecht, B Wohnort, C Schulbildung die Werte 1,1,1. Fur diese Konstellation prognostiziert Almo einen Wert von 4.16666. Das ist das Randmittel fur A1B1C1. Der Proband 3 hat die Werte 1,1,2. Almo prognostiziert 2.5000. Das ist das Randmittel fur A1B1C2.

### Randmittel und Effekte

Das geschatzte Randmittel z.B. fur die Auspragungen mannlich und weiblich des Geschlechts ergeben sich sehr einfach aus den Effekten fur diese beiden plus dem Konstanteneffekt. Almo gibt folgende Effekte aus

```
Effekte von A Geschlecht
A1 mannlich      -0.6806
A2 weiblich       0.6806

Konstanteneffekt  3.9722
```

Die Randmittel sind dann

```
Randmittel Geschlecht:mannlich = -0.6806 + 3.9722 = 3.2916
Randmittel Geschlecht:weiblich =  0.6806 + 3.9722 = 4.6527
```

### Formeln fur Randmittel

Almo berechnet die Randmittel nach allgemein gultigen Formeln, die nicht nur fur das Verfahren der weighted squares of means, sondern auch fur fitting constants gelten. Sie gelten auch, wenn sich zusatzlich Kovariate in der Analyse befinden (Fall der Kovarianzanalyse). Bei den fitting constants mussen allerdings einige Einschrankungen vorgenommen werden.

Wir verwenden fur die Formeln folgende Notation:

```
-----
R = Randmittel
E = Effekt
```

mit  $i, j, k$  werden die Auspragungen der nominalen Variablen A, B, C bezeichnet

```
E(Ai), E(Bj), E(Ck)           = Haupteffekte von A, B, C
E(Ai.Bj), E(Ai.Ck), E(Bj.Ck) = 2-er Interaktionseffekte AB, AC, BC
E(Ai.Bj.Ck)                   = Effekt der 3-er Interaktion ABC
```

```
K = Konstante (bei Varianzanalyse identisch mit Mittelwert der
          abhangigen Variablen)
-----
```

Randmittel der nominalen Variablen A,B,C  $R(A), R(B), R(C)$

$$R(A_i) = K + E(A_i)$$

$$R(B_j) = K + E(B_j)$$

$$R(C_k) = K + E(C_k)$$

2-er Randmittel für Variablenkombination A\*B  $R(AB)$ :

$$R(A_i.B_j) = K + E(A_i) + E(B_j) + E(A_i.B_j)$$

2-er Randmittel  $R(AC)$ :

$$R(A_i.C_k) = K + E(A_i) + E(C_k) + E(A_i.C_k)$$

2-er Randmittel  $R(BC)$ :

$$R(B_j.C_k) = K + E(B_j) + E(C_k) + E(B_j.C_k)$$

3-er Randmittel Variablenkombination A\*B \*C  $R(ABC)$ :

$$R(A_i.B_j.C_k) = K + \underbrace{E(A_i) + E(B_j) + E(C_k)}_{G1} + \underbrace{E(A_i.B_j) + E(A_i.C_k) + E(B_j.C_k)}_{G2} + \underbrace{E(A_i.B_j.C_k)}_{G3}$$

G1 = Gruppe der Haupteffekte

G2 = Gruppe der 2-er Interaktionseffekte

G3 = Gruppe der 3-er Interaktionseffekte

Sind 4 nominale Variable vorhanden, dann sind die Formeln für die Randmittel nach dem Prinzip zu bilden, das bei obigem 3-er Randmittel offenkundig geworden ist. Also berechnet keine Randmittel 5. und höherer Ordnung.

### *Randmittel in Kovarianzanalyse*

Befinden sich Kovariate in der Analyse, dann werden die kovarianzadjustierten Effekte in die Formeln eingesetzt, die Also automatisch im Fall der Kovarianzanalyse berechnet und standardmäßig ausgibt. Auch die Konstante muss "angepasst" werden.

In unserem Beispiel werden noch die beiden Kovariaten hinzugefügt

Alter  
Einkommen

Das geschätzte Randmittel z.B. für die Ausprägung Stadt der Variablen Wohnort ergibt sich dann gemäß folgender Gleichung

$$R(B_j) = E(B_j) + \text{Konstante} + \beta_1 * M_1 + \beta_2 * M_2$$

$\beta_1$  = Regressionskoeffizient der 1. Kovariaten (=Alter)

$\beta_2$  = Regressionskoeffizient der 2. Kovariaten (=Einkommen)

$M_1$  = Mittelwert der 1. Kovariaten (=Alter)

$M_2$  = Mittelwert der 2. Kovariaten (=Einkommen)

Den Ausdruck " $\text{Konstante} + \beta_1 * M_1 + \beta_2 * M_2$ " bezeichnen wir in der Ergebnisliste als "modellreproduzierter Gesamtmittelwert" gelegentlich auch als "an die Kovariaten angepasste" Konstante.

Allgemein gilt also: Die "angepasste" Konstante" ist gleich der Konstanten (wie sie Also ausgibt) plus der Summe der mit ihren Regressionskoeffizienten gewichteten Mittelwerte der Kovariaten.

Beim Verfahren der *fitting constants I* können Randmittel nur für Analysen mit 2 nominalen Variablen A und B und ihre Kombination AB ausgegeben werden. Bei Analysen mit mehr als 2 nominalen Variablen werden nur die Randmittel für die nominalen Variablen, nicht jedoch für die Variablenkombinationen ausgegeben. Der Grund dafür ist, dass in diesen Fällen (bei ungleichen Zellenhäufigkeiten) wechselnde Interaktionseffekte auftreten – die Randmittel somit nicht eindeutig bestimmt werden können.

*Was wird nun durch ein Randmittel ausgedrückt ?*

Wir haben oben gezeigt, dass die Randmittel höchster Ordnung als Prognosewerte für Probanden interpretiert werden können.

Was wird durch ein Randmittel niedriger Ordnung ausgedrückt? Worin besteht der Erkenntnisgewinn, den uns beispielsweise das 2-er Randmittel  $R_{A_1B_1}$  mit 3.3334 anbietet ? In der Formel für das 2-er Randmittel  $R_{(AB)}$  fehlen die Effekte für  $AC, BC, ABC$ . Das Randmittel kann also betrachtet werden als Prognosewerte, bei dem die anderen 2-er Effekte und die höheren Effekte „auspartiielliert“ sind. So könnte definiert werden: Randmittel (unterhalb höchster Ordnung) sind partielle Prognosewerte.

Siehe auch die sehr ausführliche Darstellung zu den Randmitteln in Abschnitt P20.9.1.0. Also berechnet keine Randmittel für Kovarianzanalysen mit 5 und mehr nominalen Variablen.

### **P20.6.7 Die Standardisierung der Daten**

Die Wirkung der Standardisierung besteht darin, dass die Standardabweichung einer Variablen zu ihrer Maßeinheit gemacht wird. Zwei quantitative Variable  $x_1$  und  $x_2$ , die ursprünglich in verschiedenen Maßeinheiten gemessen wurden, werden nach der Standardisierung in derselben Maßeinheit gemessen. Sie sind nunmehr vergleichbar.

Ist die **unabhängige** Variable eine nominale Variable H mit 0-1 kodierten Dummy-Variablen  $h_1, h_2, h_3, \dots, h_w$ , dann wird durch die Standardisierung die Wirkung, die die unterschiedlichen Besetzungszahlen der verschiedenen Ausprägungen von H auf die Haupt- und Interaktionseffekte ausüben, herausgenommen.

### **P20.6.8 Ergebnisse bei abhängiger nominaler Variabler**

#### **Diskriminanzanalyse, lineare Wahrscheinlichkeitsanalyse**

Betrachten wir ein Beispiel: Wir wollen die Bestimmungsgründe der Wahl einer Studienrichtung durch Abiturienten bestimmen. Die abhängige nominale Variable ist die Studienrichtung (Symbol: S) mit den Ausprägungen:

- S1 = Geisteswissenschaft
- S2 = Wirtschaftswissenschaft
- S3 = Naturwissenschaft.

Die Studienrichtung wird in drei 0-1 kodierte Dummies aufgelöst, die den drei Ausprägungen entsprechen.

Die unabhängigen nominalen Variablen sind:

1. Geschlecht (Symbol A): männlich ( $a_1$ ), weiblich ( $a_2$ )
2. Wohnort (Symbol B): Großstadt ( $b_1$ ), Mittel- und Kleinstadt ( $b_2$ ) Land ( $b_3$ ).

Als unabhängige quantitative Variable verwenden wir:

1. das Alter (bei Aufnahme des Studiums). Symbol  $x_1$
2. das Einkommen des Vaters, gemessen in 10 Einkommensstufen.  
Symbol:  $x_2$

In P20.9.5.3 (in Teil II des Almo-Dokuments) ist das Almo-Programm und die Ergebnisse für dieses Beispiel enthalten.

Betrachten wir zunächst die Gleichung für die 0-1 kodierte abhängige Variable  $S_2$ , also für die Wirtschaftswissenschaft

$$(1) S_2' = \tau_2 * t + \alpha_{12} * a_1 + \alpha_{22} * a_2 + \beta_{12} * b_1 + \beta_{22} * b_2 + \beta_{32} * b_3 + \gamma_{12} * x_1 + \gamma_{22} * x_2$$

der 1. Index bezieht sich auf die unabhängige Dummy-Variable, der 2. Index bezieht sich auf die abhängige Variable.

$t$	Konstantenvariable (immer =1)
$\tau_2$	Effekt der Konstanten hinsichtlich der Studienrichtung "Wirtschaftswissenschaften".
$a_1, a_2$	die 0-1 kodierten Dummies des Geschlechts.
$b_1, b_2, b_3$	die 0-1 kodierten Dummies des Wohnorts.
$\alpha_{12}$	der Haupteffekt der unabhängigen Dummy-Variablen "männliches Geschlecht" hinsichtlich der abhängigen Variablen die Wahl der Studienrichtung "Wirtschaftswissenschaft".

$\alpha_{22}$  sowie  $\beta_{12}, \beta_{22}, \beta_{32}$  sind entsprechend definiert.

$x_1, x_2$	das sind die beiden quantitativen unabhängigen Variablen,
$\gamma_{12}, \gamma_{22}$	das sind die Regressionskoeffizienten von $x_1$ und $x_2$ hinsichtlich der abhängigen Variablen $S_2$ .
$S_2'$	vom Modell prognostizierte Wahrscheinlichkeit für "Wirtschaftswissenschaft".

Allgemein lautet die Gleichung:

$$(2) S_i' = \tau_i * t + \alpha_{1i} * a_1 + \alpha_{2i} * a_2 + \beta_{1i} * b_1 + \beta_{2i} * b_2 + \beta_{3i} * b_3 + \gamma_{1i} * x_1 + \gamma_{2i} * x_2$$

für  $i = 1, 2, 3$ ,

wobei 1 = Geisteswissenschaft,  
2 = Wirtschaftswissenschaft,  
3 = Naturwissenschaft.

ALMO liefert uns nun für die drei Gleichungen die Werte der Effekte. Für obige Gleichung 2 erhalten wir folgende Effekte (gerechnet mit Verfahren = fitting constants I):

$$\begin{array}{ll} \tau_2 = 0.35 & \beta_{12} = 0.10 \\ \alpha_{12} = 0.40 & \beta_{22} = 0.05 & \gamma_1 = 0.002 \\ \alpha_{22} = -0.20 & \beta_{32} = -0.05 & \gamma_2 = 0.002 \end{array}$$

Gleichung 1 kann also nunmehr in folgender Form geschrieben werden:

$$3) S_2' = 0.35t_0 + 0.40a_1 - 0.20a_2 + 0.10b_1 - 0.05b_2 + 0.05b_3 + 0.002x_1 + 0.002x_2.$$

Für einen Studenten, der männlich ist (also  $a_1=1$ ,  $a_2=0$ ) und von der Großstadt kommt, (also  $b_1=1$ ,  $b_2=0$ ,  $b_3=0$ ), der im Alter von 20 Jahren sein Studium aufnahm (also  $x_1=20$ ) und dessen Vater in der Einkommensstufe 5 liegt (also  $x_2=5$ ), erhalten wir

$$S'_2 = 0.90;$$

d.h. die Wahrscheinlichkeit, dass er Wirtschaftswissenschaften wählt, ist = 0.90 bzw. 90 %.

Für einen Studenten, der weiblich ist und vom Land kommt, bei dem  $x_1=1$  und  $x_2=2$  ist, erhalten wir

$$S'_2 = 0.14,$$

also eine Wahrscheinlichkeit von 14 %, dass diese Studentin Wirtschaftswissenschaften studiert.

Wir können nun unser Modell dadurch erweitern, dass wir auf der rechten Gleichungsseite die Interaktion von A und B, also die 0-1 kodierten multiplikativen Dummies  $a_1b_1$ ,  $a_1b_2$ ,  $a_1b_3$ ,  $a_2b_1$ ,  $a_2b_2$ ,  $a_2b_3$ , als unabhängige Variable einführen. Am Prinzip der Interpretation würde sich nichts ändern.

Das Prinzip unserer Interpretation ist also folgendes: Ist die abhängige Variable eine nominale, so erhalten wir so viele Gleichungen, wie die abhängige Variable Ausprägungen besitzt. Jede Gleichung prognostiziert dann die **Wahrscheinlichkeit**, dass die betreffende Ausprägung gegeben ist. Siehe hierzu auch unsere Darstellung der Diskriminanzanalyse in P20.9.5.2 und in P29.2.

Am Vorzeichen der Effekte erkennen wir, ob die betreffende Ausprägung die Wahrscheinlichkeit für die Wahl der Wirtschaftswissenschaften als Studienrichtung begünstigt oder benachteiligt. "männlich" erhöht die Wahrscheinlichkeit für Wirtschaftswissenschaft, "weiblich" verringert sie.

Bei den Regressionskoeffizienten der quantitativen Variablen bedeutet ein positives Vorzeichen: Je mehr  $x$  umso höher die Wahrscheinlichkeit für  $S_i$  - und ein negatives: je mehr  $x$ , umso geringer die Wahrscheinlichkeit für  $S_i$ .

An der absoluten Höhe der Effekte und der Regressionskoeffizienten erkennen wir die Bedeutsamkeit der betreffenden Variablen. So ist in unserem Beispiel offenkundig, dass "Großstadt" ( $b_1$ ) bedeutsamer ist als "Mittel- und Kleinstadt". Hat eine unabhängige nominale Variable nur 2 Ausprägungen (wie das Geschlecht), dann ist ein solcher Vergleich nicht möglich. Es ist nicht sinnvoll zu sagen, dass "männlich" bedeutsamer ist als "weiblich".

Ein Vergleich zwischen den unabhängigen quantitativen Variablen  $x_1$  und  $x_2$  ist eigentlich nur nach Standardisierung der Variablen möglich (wenn wir mit "MATRIX=KORRELATION;" rechnen). Siehe dazu Holm, 1979, Abschnitt 11 und Holm, 1977, Abschnitt 12.

ALMO liefert auch die durch die unabhängigen nominalen und quantitativen Variablen erklärten generalisierten Streuungen und (daraus abgeleitet) die partiellen PRE-Korrelationskoeffizienten. Diese Koeffizienten sind pauschale Koeffizienten, die die Wirkung einer unabhängigen Variablen hinsichtlich der (zusammengefassten) Menge der Dummies der abhängigen nominalen Variablen in einem einzigen Zahlenwert ausdrücken. Siehe dazu unser Beispiel im Abschnitt P20.9.4.

#### ***P20.6.8.1 Einschränkungen zur linearen Wahrscheinlichkeitsanalyse***

Die Anwendung des ALM auf dichotome und polytome Zielvariable schafft einige Probleme.

Diese sind:

1. Das Modell kann Wahrscheinlichkeiten prognostizieren, die außerhalb des Bereichs 0 bis 1 liegen. Es kann beispielsweise prognostizieren, dass die Wahrscheinlichkeit für Wirtschaftswissenschaften  $p=1.08$  (also 108%) ist.
2. Es besteht modellbedingte Varianz-Heteroskedastizität mit der Folge, dass die Schätzer für die Parameter der ursächlichen Variablen zwar unverzerrt und konsistent, aber nicht mehr effizient sind. Das bedeutet, dass die Standardfehler der Effekte und Regressionskoeffizienten der ursächlichen Variablen nicht minimal sind, mit der Folge, dass die Signifikanzüberprüfung mit t- und F-Test nicht korrekt ist. Siehe dazu die ausführliche Darstellung bei Aldrich/Nelson (1984, S. 12ff) und Urban (1993, S. 17ff), sowie Urban (1982, Abschnitt 3.1 und 3.1.1).

Auf das 1. Problem sind wir im Almo-Dokument Nr. 25 „Statistische Datenanalyse II“, Abschnitt P45.15.1.6 ausführlich eingegangen. Es zeigt sich, dass die Reproduzierungs- bzw. Prognosefähigkeit des Modells dadurch nicht beeinträchtigt wird.

Das 2. Problem kann im Rahmen des ALM durch die „gewichtete Kleinste-Quadrate-Schätzung“ gelöst werden. Dieses Verfahren geht auf Goldberger (1964) zurück. Man nimmt dabei allerdings in Kauf, dass die Reproduzierbarkeit dieser Modellvarianten schlechter ist. D.h. die Fähigkeit des Modells, die Untersuchungseinheiten einer Ausprägung der Zielvariablen richtig zuzuweisen, ist schlechter als bei der normalen Kleinste-Quadrate-Lösung. Wir bieten im Programm Prog20mo die gewichtete Kleinste-Quadrate-Schätzung für nominal-dichotome Zielvariable als Option an. Siehe dazu die Erläuterungen zur Optionsbox „gewichtete Analyse“ in P20.8.0.1. In Prog45gw im Almo-Dokument Nr. 25 „Statistische Datenanalyse II“, Abschnitt P45.15.2.2 und mit Prog20mq bieten wir ein Programm an, das eine gewichtete Kleinste-Quadrate-Schätzung für nominal-polytome Zielvariable leistet.

Als beste Alternative zum ALM für dichotome und polytome Zielvariable wird das **Logit-Modell** empfohlen. Wir haben es im Almo-Dokument Nr. 9 „Logitanalyse“, Abschnitt P22 und im Almo-Dokument Nr. 25 „Statistische Datenanalyse II“, Abschnitt P45.7.6 dargestellt. Das Logit-Modell leidet zwar nicht unter diesen beiden Problemen, seine Ergebnisse sind aber nicht so einsichtig interpretierbar wie die des ALM. Dies gilt insbesondere für polytome Zielvariable.

Eine weitere Möglichkeit, nominale Zielvariable zu analysieren, besteht in der (kanonischen) Diskriminanzanalyse. Einige Autoren (etwa Urban, 1993) vertreten die Auffassung, dass bei der kanonischen Diskriminanzanalyse auf Seiten der ursächlichen Variablen nur quantitative aber nicht nominale Variable (bzw. deren 0-1 kodierte Dummies) zulässig sind - was natürlich die praktische Anwendung dieses Verfahrens sehr einschränkt. Außerdem entstehen, bei der auf nominal-polytome Zielvariable angewandten kanonischen Diskriminanzanalyse, Koeffizienten, die inhaltlich nicht interpretierbar sind. Die kanonische Diskriminanzanalyse ist als Prog29 und Prog27 in Almo enthalten und im Almo-Dokument Nr. 4 „Kanonische Analysen“ ausführlich dargestellt.

## Hermann Denz

### P20.6.9 Die Einbeziehung ordinaler Variablen

Der Programmteil in unserem Computer-Programm, der es ermöglicht ordinale Variable einzuführen wurde von Hermann Denz programmiert.

Zu den folgenden Ausführungen siehe auch die ausführliche Darstellung von H. Potuschak im Handbuch, Teil 3, Anhang oder im Almo-Dokument Nr. 5 „Korrelation“.

Sobald sich auch nur eine ordinale Variable auf Seiten der unabhängigen oder der abhängigen Variablen befindet, entspricht unsere Vorgangsweise ungefähr der Berechnungsweise des Kendall'schen tau-b. Sie ist folgende (siehe dazu Holm, 1979, Abschnitt 6 und Denz, in Holm (Hrsg.), 1979, Abschnitt 5):

Wir bilden Paare von Untersuchungspersonen und zwar so, dass jede Untersuchungsperson mit allen anderen (N-1) Untersuchungspersonen zusammengebracht wird. Es entstehen so  $(N^2-N)/2$  Paare. Für jedes Paar wird ermittelt, welche Differenz es in der jeweiligen Variablen aufweist.

Unsere neue Datenmatrix besteht nun nicht mehr aus individuellen Untersuchungspersonen, sondern aus den Paaren (als Zeilen) und aus den Variablen (als Spalten), wobei für diese nicht mehr die Werte der individuellen Untersuchungspersonen, sondern die Wertedifferenz für ein Paar eingetragen wird.

Die Wertedifferenz für ein Paar ij wird nun bei ordinalen Variablen in folgender Weise ermittelt: Ist die Untersuchungsperson i in der betreffenden Variablen größer als die Untersuchungsperson j, dann wird dem Paar ij der Wert +1 zugewiesen. Ist die Untersuchungsperson i kleiner als j, dann erhält das Paar ij den Wert -1. Sind i und j gleich groß, dann erhält das Paar ij den Wert 0.

Bei quantitativen Variablen wird die Differenz des Werts der Untersuchungsperson i in der betreffenden Variablen minus dem Wert der Untersuchungsperson j verwendet; desgleichen bei den 0-1 kodierten Dummies.

Die Interpretation der Regressionsgleichung (mit mindestens einer unabhängigen ordinalen Variablen) richtet sich immer nach dem Niveau der abhängigen Variablen:

Wenn die abhängige Variable quantitativ ist, dann ist der durch die rechte Gleichungsseite prognostizierte Wert der abhängigen Variablen  $Y'$  eine Differenz zwischen zwei Untersuchungspersonen i und j.

Ist die abhängige Variable, die 0-1 kodierte Dummy-Variablen einer nominalen Variablen, dann gibt der prognostizierte Wert  $Y'$ , der zwischen 0 und 1.0 liegen wird, die Wahrscheinlichkeit an, dass die beiden Untersuchungspersonen i und j in der abhängigen Variablen ungleich sind. Es sei daran erinnert, dass 0 Gleichheit des Paares bedeutet.

Ein Wert von beispielsweise  $Y' = 0.7$  bedeutet, dass i und j mit 70%-iger Wahrscheinlichkeit in den abhängigen Dummy-Variablen verschieden sind. Die Regressionsgleichung ist als lineare Wahrscheinlichkeitsfunktion zu interpretieren (siehe hierzu Holm, 1979, Abschnitt 13).

Ist die abhängige Variable eine 1,0,-1 kodierte ordinale Variable, dann wird der prognostizierte Wert  $Y'$  zwischen +1 und -1 liegen. Er ist als Wahrscheinlichkeit des Abweichens von 0, d.h. der Wahrscheinlichkeit der Ungleichheit zu interpretieren, wobei das Vorzeichen die Richtung der Ungleichheit angibt. Ein Wert von beispielsweise  $Y' = +0.7$  besagt, dass i mit 70%-iger Wahrscheinlichkeit in y größer ist als j. Ein Wert von  $Y' = -0.3$  besagt, dass i mit 30%-iger Wahrscheinlichkeit kleiner ist als j. Auch in diesem Fall ist die Regressionsgleichung eine lineare Wahrscheinlichkeitsfunktion.

Die Interpretation der Regressionskoeffizienten und Effekte ist folgende: Sie geben an, wie stark die Ungleichheit zwischen den Untersuchungspersonen  $i$  und  $j$  in der unabhängigen Variablen die Ungleichheit von  $i$  und  $j$  in der abhängigen Variablen bestimmt.

**Beachte:** Die Berechnung parametrischer Signifikanztests (F-Test, t-Test) wie sie in Prog20 automatisch durchgeführt werden, ist bei unabhängigen ordinalen Variablen problematisch.

## P20.7 Die Schätz-Verfahren des Allgemeinen Linearen Modells

In den folgenden Abschnitten betrachten wir den Fall, dass sich unabhängige nominale Variablen im Modell befinden. Dabei dürfen sich zusätzlich noch unabhängige quantitative Variable (=Kovariate) im Modell befinden (Fall der Kovarianzanalyse). Befinden sich nur unabhängige quantitative Variable im Modell, dann ist das ALM identisch mit der üblichen Regressionsanalyse. Deren Theorie und Kalkül setzen wir als bekannt voraus – so dass wir diese hier nicht darstellen werden.

### Terminologie

Wir verwenden im folgenden Text auch immer wieder kürzere Begriffe für einige lange und umständliche Begriffe

Unabhängige nominale Variable werden auch als **Faktoren** bezeichnet.

Unabhängige quantitative Variable werden auch als **Kovariate** bezeichnet.

Der im ALM verwendete Begriff "Faktor" darf nicht verwechselt werden mit dem aus der Faktorenanalyse.

Die Faktoren (=unabhängigen nominalen Variablen) A,B,C,... korrelieren miteinander, wenn die Zellen, d.h. ihre Merkmalskombinationen mit ungleichen Häufigkeiten besetzt sind. Dies ist in der nicht-experimentellen Forschung der Normalfall.

Wenn wir also die Wirkung, die eine einzelne Variable, beispielsweise A, auf die abhängige Variable  $y$  ausübt, ermitteln wollen, dann müssen wir alle Faktoren A,B,C,... gegeneinander "auspartiellieren". Der Sachverhalt wird noch dadurch weiter kompliziert, dass auch die Interaktionen AB, AC,... ABC bei ungleichen Zellenhäufigkeiten korrelieren und dass die unabhängigen quantitativen Variablen (=die Kovariaten)  $x_1, x_2, x_3, \dots$  mit den Faktoren und ihren Interaktionen in der Regel ebenfalls korrelieren.

### Auspartiellierung und Anpassung

Wenn unabhängige Variable untereinander korrelieren, dann müssen sie gegeneinander auspartielliert werden. In der Literatur sind die unschönen Begriffe "Auspartiellierung" oder "Auspartialisierung" zu finden, nicht selten auch der "schönere" Begriff der "Anpassung". Wir haben uns entschieden, die beiden Begriffe der *Auspartiellierung* und der *Anpassung* zu verwenden. Mit "Auspartiellierung" wollen wir beim Leser eine Assoziation zum geläufigeren Begriff der "partiellen Korrelation" herstellen. Denn das Auspartiellieren von unabhängigen Variablen aus abhängigen ist nichts anderes als das Bilden von ganz bestimmten partiellen Korrelationen zwischen ihnen.

Auspartiellierung und Anpassung sind komplementäre Begriffe. Wenn wir beispielsweise formulieren

"A wird aus B *auspartielliert*", dann heißt das "B wird an A *angepasst*".

In ALMO werden 5 Methoden der Auspartiellierung und damit 5 Verfahren für das ALM angeboten.

1. Das Verfahren der "fitting constants I". Es wird eine gruppenweise hierarchische Auspartiellierung durchgeführt.
2. Das Verfahren der „fitting constants II“. Es ist eine Variante von fitting constants I. Es ist identisch mit „SS Typ II“ aus dem GLM von SAS und SPSS.
3. Das Verfahren der "weighted squares of means". Es wird eine gegenseitige Auspartiellierung aller Effekte vorgenommen. Dieses Verfahren ist identisch mit SS Typ III in SAS und SPSS. Es ist das Standardverfahren in Almo und SPSS.
4. Das sequentielle Verfahren. Es wird eine variablenweise hierarchische Auspartiellierung durchgeführt. Dieses Verfahren ist identisch mit SS Typ I in SAS und SPSS.
5. Die vom Benutzer beliebig selbst bestimmte Auspartiellierung. Sie wird in Abschnitt P20.13 im 2. Teil des Handbuchs beschrieben.

Bei gleichen Zellenhäufigkeiten erbringen die Methode 1 bis 4 dieselben Ergebnisse. Wir wollen hier gleich eine **Empfehlung** aussprechen:

Wenn nicht besondere Gründe dafür sprechen, dann sollte man mit dem Verfahren der „weighted squares of means“ rechnen. Es ist als Standard-Verfahren in den Almo-Programm-Masken Prog20mo und Prog20mx voreingestellt. Auch in SPSS ist es unter der Bezeichnung SS-Typ III das Standardmodell. Der Begriff "weighted squares of means“ geht zurück auf Yates, der dieses Verfahren im Ansatz schon 1934 entwickelte. Der Begriff wird heute bei SPSS und SAS nicht mehr verwendet.

*Der Leser, der sich für die teilweise sehr komplexen Ausführungen zu den anderen vier Methoden nicht interessiert, sei empfohlen, die folgenden Abschnitte zu überspringen und bei Abschnitt P20.7.3 (ca. S.42) weiter zu lesen. Dort wird das Standardverfahren des ALM ausführlich erläutert.*

## **P20.7.1 Das Verfahren der "fitting constants I"**

### **Die gruppenweise hierarchische Auspartiellierung**

Der Begriff der "fitting constants" wurde (vermutlich) erstmalig von J.E.Overall & D.K.Spiegel in ihrer Arbeit "Concerning least squares analysis of experimental data" (1969) verwendet. Der Titel verdeutlicht schon, um was es geht. Es geht darum, ein Rechenverfahren zu entwickeln, das für nicht-orthogonale Analysen eine Kleinste-Quadrate-Lösung bereitstellt. Nicht-orthogonale Analysen sind Analysen mit ungleichen Zellenhäufigkeiten bzw. unbalancierten Zellenhäufigkeiten.

Um es vorwegzunehmen: Das Verfahren der "fitting constants" besitzt einige Einschränkungen, auf die weiter unten eingegangen wird. Sie treten nicht auf, wenn gleiche Zellenhäufigkeiten gegeben sind und - auch bei ungleichen Zellenhäufigkeiten - wenn keine Interaktionen höher als 2. Ordnung in die Analyse einbezogen werden.

Die unabhängigen Variablen werden nach einem bestimmten Schema gegeneinander auspartielliert. Betrachten wir ein Beispiel mit 3 Kovariaten  $x_1$ ,  $x_2$ ,  $x_3$  und 3 nominalen Variablen A, B, C. Sie bilden folgenden Gruppen von unabhängigen Variablen:

<b>Gruppe der Kovariaten</b>	bilden die unabhängigen quantitativen Variablen $x_1$ , $x_2$ , $x_3$ ,
<b>Gruppe 1</b>	bilden die Dummies der Faktoren A, B, C,
<b>Gruppe 2</b>	bilden die Dummies der Zweier-Interaktionen AB, AC, BC

### Gruppe 3

bildet die Dummies Dreier-Interaktion ABC.

Gruppe 1 bis 3 werden von den 0-1 kodierten Dummies der Faktoren bzw. deren Interaktionen gebildet (dabei werden zur Vermeidung linearer Abhängigkeiten bestimmte Dummies herausgenommen - siehe dazu Abschnitt P20.3 und P20.4.

#### ***P20.7.1.0.1 Fitting constants I bei Varianzanalyse.***

Wir wollen zunächst den einfachen Fall der Varianzanalyse betrachten. Die unabhängigen Variablen sind nominal. Kovariate befinden sich nicht im Modell.

Betrachten wir das Beispiel einer Varianzanalyse mit den 3 Faktoren A,B,C und ihren Interaktionen. Die Faktoren (=unabhängigen nominalen Variablen) sind also folgende

A    B    C / AB AC BC / ABC

Die Schrägstriche trennen die drei hierarchischen Gruppen. Der Einfachheit halber wollen wir annehmen, dass die beiden Faktoren A und B je 2 Ausprägungen besitzen und C 3 Ausprägungen. A und B besitzen als nur eine nicht-redundante Dummy, a1 und b1, C hingegen zwei, c1 und c2. Aus den Kombinationen der Dummies entstehen dann die Interaktionsdummies 2. Ordnung a1b1, a1c1, a1c2, b1c1, b1c2 und 3.Ordnung a1b1c1, a1b1c2. In Dummies aufgelöst liegt folgendes Modell vor:

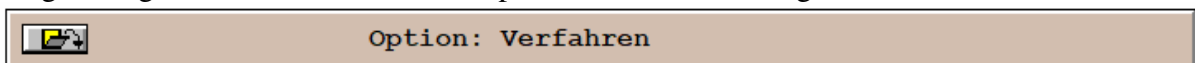
Gruppe 1				/	Gruppe 2						/	Gruppe 3	
A	B	C		/	AB	AC		BC		/	ABC		
				/						/			
		-----		/		-----		-----		/	-----		
a1	b1	c1 c2		/	a1b1	a1c1 a1c2		b1c1 b1c2		/	a1b1c1 a1b1c2		

Die Gruppen 1 bis 3 werden hierarchisch auspartielliert. Aus den Dummies der Gruppe 3 (den Dreier-Interaktionen) wird das herausgenommen, was die Gruppe 1 und 2 in ihnen erklären. Aus den Dummies der Gruppe 2 (den Zweier-Interaktionen) wird das herausgenommen, was die Gruppe 1 in ihnen erklärt. Zurück bleiben in Gruppe 2 und 3 somit Partial-Dummies. Gruppe 1 bleibt "unangetastet". In den Gruppen 1, 2 und 3 wird nun noch gruppenintern gegenseitig auspartielliert, d.h. innerhalb einer Gruppe werden die Dummies gegenseitig auspartielliert.

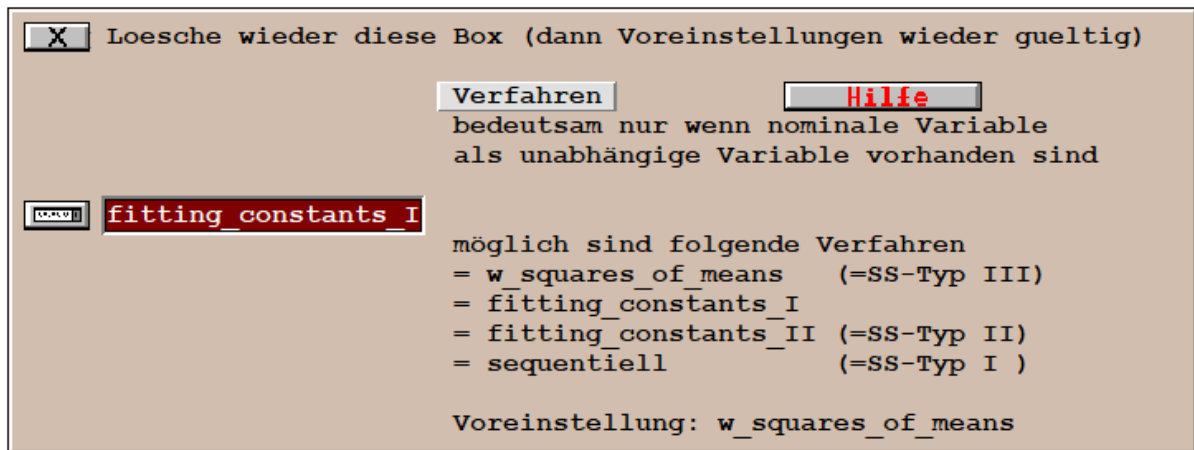
Das Prinzip ist also: Aus der Gruppe X werden alle Gruppen, die hierarchisch vor ihr stehen, auspartielliert. Danach werden die so entstandenen Partial-Dummies gruppenintern gegenseitig auspartielliert. Abschließend wird dann errechnet, welche Streuungen die in Partial-Dummies zerlegten Faktoren in der abhängigen Variablen erklären. Wir nennen das hier beschriebene Verfahren, das der *gruppenweisen hierarchischen Auspartiellierung*. Verzichtet der Forscher auf Interaktionen, dann befindet sich im Modell nur die Gruppe 1 der Faktoren (ohne deren Interaktionen). In diesem Fall wird nur gruppenintern auspartielliert.

#### ***Almo-Programm-Masken.***

Varianzanalysen nach dem Verfahren der fitting constants I werden mit der Programm-Maske Prog20mo gerechnet. Dazu muss die Optionsbox "Verfahren" geöffnet werden.



geöffnet:



### ***P20.7.1.0.2 Fitting constants I bei Kovarianzanalyse.***

Sind Kovariate vorhanden (Fall der Kovarianzanalyse) dann mussen die Gruppe der Kovariaten und die Gruppen der Faktoren aneinander angepasst werden. Dafur gibt es zwei Vorgehensweisen. Wir nennen die eine *die einmalige Anpassung* und die andere *die hierarchische Anpassung*. Die beiden liefern unterschiedliche Ergebnisse. Wir beschreiben zuerst die einmalige Anpassung.

#### ***Die einmalige Anpassung von Kovariaten und Faktoren***

In einem 1. Schritt findet eine *gegenseitige* Auspartiellierung der Gruppe der Kovariaten einerseits und der Gruppen der Dummies andererseits statt. Die Gruppen 1 bis 3 werden aus der Gruppe der Kovariaten auspartielliert. D.h. aus den Variablen der Gruppe der Kovariaten wird das herausgenommen, was die Dummies der anderen Gruppen 1 bis 3 in ihnen erklaren. Sind 3 Kovariate  $x_1, x_2, x_3$  vorhanden dann bleiben die "Partialvariablen"  $x^*_1, x^*_2, x^*_3$  zuruck. Diese werden dann anschlieend gruppenintern gegeneinander auspartielliert. Danach konnen die Regressionskoeffizienten der Kovariaten gegenuber der abhangigen Variablen  $y$  bestimmt werden und die Streuung errechnet werden, die die Kovariaten in der abhangigen Variablen  $y$  erklaren.

Im 2. Schritt wird umgekehrt aus den Dummies der Gruppen 1 bis 3 das herausgenommen, was die Kovariaten (also  $x_1, x_2, x_3$ ) in ihnen erklaren. Zuruck bleiben "Partial-Dummies" (die zwar noch dichotom sind, aber nicht mehr die Werte 0-1 besitzen).

Danach werden die Gruppen 1 bis 3 hierarchisch auspartielliert. Wie schon beschrieben, wird aus den Dummies" der Gruppe 3 (den Dreier-Interaktionen) das herausgenommen, was die Gruppe 1 und 2 in ihnen erklaren - usw. Zu beachten ist, dass die Dummies der drei Gruppen bereits an die Kovariaten angepasst sind. Die gruppenweise hierarchische Auspartiellierung wird also auf Partial-Dummies angewendet. Durch sie werden somit Partial-Dummies 2. Ordnung erzeugt. Abschliessend kann dann errechnet werden, welche Streuungen die so in Partial-Dummies 2. Ordnung zerlegten Faktoren in der abhangigen Variablen erklaren.

#### ***Die hierarchische Anpassung von Kovariaten und Faktoren***

Im 1. Schritt werden alle Dummies der drei Gruppen aus den Kovariaten  $x_1, x_2, x_3$  auspartielliert. Zuruck bleiben die "Partialvariablen"  $x^*_1, x^*_2, x^*_3$  der Kovariaten. Diese werden dann anschlieend gruppenintern gegeneinander auspartielliert. Danach werden fur die Kovariaten die Regressionskoeffizienten gegenuber der abhangigen Variablen  $y$  bestimmt

und die Streuungen errechnet, welche die Kovariaten in der abhängigen Variablen  $y$  erklären. Der Vorgang ist derselbe wie bei der einmaligen Anpassung - mit dem Unterschied, dass nicht gegenseitig die Kovariaten aus allen Dummies der drei Gruppen auf ein Mal auspartiielliert werden. Das geschieht separat für jede der drei hierarchischen Gruppen.

1. Zuerst werden die Dummies der Gruppen 1 und 2 plus die Kovariaten  $x_1, x_2, x_3$  aus den Dummies der 3. Gruppe auspartiielliert. Für die so entstandenen Partial-Dummies der 3. Gruppe werden dann die Effekte und die erklärten Streuungen hinsichtlich der abhängigen Variablen  $y$  ermittelt.
2. Dann werden die Dummies der Gruppe 1 plus die Kovariaten aus der 2. Gruppe auspartiielliert. Für die so entstandenen Partial-Dummies der 2. Gruppe werden dann die Effekte und die erklärten Streuungen hinsichtlich der abhängigen Variablen  $y$  ermittelt.
3. Und schließlich werden die Kovariaten aus den Dummies der Gruppe 1 auspartiielliert. Für die so entstandenen Partial-Dummies werden dann die Effekte und die erklärten Streuungen hinsichtlich der abhängigen Variablen  $y$  ermittelt.

#### ***Vergleich der Ergebnisse bei einmaliger und hierarchischer Anpassung***

Die Ergebnisse, die Almo für die beiden Anpassungsmethoden liefert, sind verschieden - mit einer Ausnahme: Die Ergebnisse für die letzte Gruppe sind immer identisch. In unserem Beispiel ist dies die Gruppe 3 der Interaktion ABC.

Die Frage ist nun, welche Anpassung ist vorzuziehen. Das Verfahren der fitting constants I folgt, wie das oben im Abschnitt für die Varianzanalyse ausführlich dargestellt wurde, der Logik der *gruppenweisen hierarchischen Auspartiiellierung* der Dummies. Es scheint deswegen sinnvoll zu sein, diese Logik auch für die Anpassung zu übernehmen, d.h. die Dummies auch hierarchisch an die Kovariaten anzupassen. Dafür müssen die besonderen Programm-Masken **ProgFI\_2** bis **ProgFI\_5** verwendet werden. Die Standard-Programm-Maske Prog20mo darf nicht verwendet werden. Das gilt aber nur für den Fall der Kovarianzanalyse bei der Interaktionen ermittelt werden sollen.

#### ***Almo-Programm-Masken.***

Die Standard-Programm-Maske **Prog20mo** verwendet beim Verfahren der fitting constants I die *einmalige* Anpassung der Kovariaten an die Faktoren. Sie muss verwendet werden, wenn eine Kovarianzanalyse ohne Interaktionen gerechnet werden soll. In diesem Fall sind die beiden Anpassungsmethoden identisch.

Die besonderen Programm-Masken **ProgFI\_2** bis **ProgFI\_5** verwenden die *hierarchische* Anpassung. Sie müssen verwendet werden wenn eine Kovarianzanalyse mit Interaktionen gerechnet werden soll. Man findet sie nach Klick auf den Knopf "Verfahren/Allgemeines lineares Programm" oder "alle Progs" am Oberrand des Almo-Fensters.

#### ***Problem und Einschränkungen der fitting constants I***

Die „fitting constants I“ haben ein Gebrechen.

Für Analysen mit beliebig vielen Faktoren und deren Interaktionen und beliebig vielen Kovariaten können zwar für alle die erklärten Streuungen und die daraus abgeleiteten Koeffizienten (partielle Korrelationen, F-Wert, Signifikanz, Teststärke) ermittelt werden.

Bei ungleichen (bzw. nicht-balancierten) Zellenhäufigkeiten und 3 und mehr Faktoren und ihren Interaktionen sind die Interaktionseffekte jedoch nicht eindeutig ermittelbar. Es treten "wechselnde" Interaktionseffekte auf. Die Folge davon ist, dass keine Prognosewerte für die

Probanden ermittelt werden können. Das ist eine erhebliche Einschränkung und bedeutet das Aus für das Verfahren der fitting constants I bei dieser Konstellation.

Werden diese Interaktionseffekte benötigt, dann muss nach dem Verfahren der "weighted squares of means" gerechnet werden. Zu den "wechselnden " Interaktionseffekte siehe Abschnitt P20.6.5.1 und P20.9.1.1. Kovariate sind immer in beliebiger Zahl zulässig.

Bei Analysen mit beliebig vielen Faktoren aber ohne Interaktionen oder bei Analysen mit maximal 2 Faktoren und ihrer Interaktion und mit oder ohne Kovariate liefert das Verfahren jedoch eine eindeutige Kleinste-Quadrate-Lösung. Alle Effekte und alle sonstigen bedeutsamen Koeffizienten werden korrekt als Kleinste-Quadrate-Schätzer errechnet und ausgegeben - auch die Prognosewerte für die Probanden. Kovariate sind immer in beliebiger Zahl zulässig.

### ***P20.7.1.1 Fitting constants II (SS-Typ II)***

Dieses Verfahren ist in den Statistiksystemen SAS und SPSS unter der Bezeichnung **SS Type II** enthalten.

Das Verfahren kann als eine Erweiterung der "fitting constants I" betrachtet werden. Gegen die "fitting constants I" kann vorgebracht werden, dass die Haupteffekte verzerrt sind, weil Interaktionseffekte noch in ihnen enthalten sind. Hier bietet es sich an, aus den Haupteffekten jene Interaktionseffekte herauszunehmen, an denen sie selbst nicht beteiligt sind.

#### ***P20.7.1.1.1 Varianzanalyse mit fitting constants II***

Betrachten wir das Beispiel einer Varianzanalyse mit den 3 Faktoren A,B,C und ihren Interaktionen AB, AC, BC. Die unabhängigen Variablen sind also folgende

A    B    C    /    AB   AC   BC   /    ABC

Das wäre das Schema für die fitting constants I. Die Schrägstriche trennen die drei hierarchischen Gruppen.

A ist nicht beteiligt an der Interaktion BC. Um die durch A in der abhängigen Variablen Y erklärte Streuung  $SS_A$  zu ermitteln wird folgende Analyse gerechnet:

A    B    C    **BC** / AB   AC / ABC

Soll die durch A in Y erklärte Streuung gewonnen werden, dann muss neben B und C auch BC aus A auspartiiert werden. Das geschieht praktischerweise dadurch dass BC in die 1. Gruppe eingefügt wird. Aus A werden dann die anderen Variablen dieser 1. Gruppe auspartiiert. Die in den Gruppen 2 und 3 verbleibenden Variablen werden für die Berechnung von  $SS_A$  nicht gebraucht. Sie sind hierarchisch nachgeordnet.

In entsprechender Weise werden dann auch die Streuungen von B und C ermittelt. Betrachten wir **B**. Die 1. hierarchische Gruppe ist dann

A    B    C    **AC**

Die Interaktion AC, an der B nicht beteiligt ist muss in die 1. Gruppe aufgenommen werden, wenn die durch B erklärte Streuung ermittelt werden soll. Aus B werden dann die anderen Variablen in der 1. Gruppe, also A, C, AC auspartiiert.

Wenn 4 Faktoren A,B,C,D und ihre Interaktionen vorhanden sind, dann muss folgende Vorgehensweise angewendet werden.

Die Variablen sind

A B C D / AB AC AD BC BD CD / ABC ABD ACD BCD / ABCD

Das ist das Schema bei fitting constants I. Um die durch A in der abhängigen Variablen Y erklärte Streuung zu ermitteln, muss folgende Analyse gerechnet werden:

A B C D **BC BD CD BCD** / AB AC AD / ABC ABD ACD / ABCD

Von Interesse ist hier wieder die 1. Gruppe. Gruppe 2, 3 und 4 werden für A nicht gebraucht

Die Besonderheit ist, dass die Interaktionen (aller Ordnungen), an denen A nicht beteiligt ist, in die hierarchische Gruppe der Faktoren aufgenommen werden - das sind die 2-er Interaktionen BC, BD, CD und die 3-er Interaktion BCD. Entsprechend ist dann vorzugehen, wenn für B, C und D die erklärte Streuung berechnet werden soll.

Diese Vorgehensweise muss auch angewendet werden, wenn die Streuung ermittelt werden soll, die in Y durch die einzelnen 2-er Interaktionen AB, AC etc. erklärt wird .

Betrachten wir AB. In den nachgeordneten hierarchischen Gruppen 3 und 4 ist AB nicht beteiligt an der 3-er Interaktion ACD und BCD. Die durch **AB** in Y erklärte Streuung  $SS_{AB}$  ergibt sich aus folgender Analyse:

A B C D / AB AC AD BC BD CD **ACD BCD** / ABC ABD ACD / ABCD

Zuerst wird die übergeordnete 1. Gruppe aus der 2. Gruppe (so wie sie im Schema angegeben ist) auspartiielliert. Dann werden aus AB die anderen Variablen der 2. Gruppe, also AC,AD,BC,BD,CD und ACD,BCD auspartiielliert. Danach kann dann die durch AB erklärte Streuung errechnet werden. Die hierarchisch nachgeordneten Gruppen 3 und 4 werden nicht gebraucht. Die gleiche Wirkung entsteht, wenn gleich "in einem Aufwasch" die Variablen der 1. Gruppe plus die anderen Variablen der (erweiterten) 2. Gruppe aus AB auspartiielliert werden.

Die Besonderheit ist also, dass die 3-er Interaktion BCD, an der AB nicht beteiligt ist, in die hierarchische Gruppe der 2-er Interaktionen aufgenommen wird. Soll z.B. die durch die Interaktion **AC** erklärte Streuung  $SS_{AC}$  ermittelt werden, dann besteht die Gruppe der 2-er Interaktionen aus

AB AC AD BC BD CD **ABD BCD**

#### ***P20.7.1.1.2 Kovarianzanalyse mit fitting constants II***

Das Verfahren der fitting constants II ist selbstverständlich auch auf den Fall der Kovarianzanalyse (bei der die Kovariaten  $x_1, x_2, x_3 \dots x_m$  eingeführt werden) anwendbar.

Kovariate und Faktoren und deren Interaktionen müssen aneinander angepasst werden.

Wir betrachten wieder das Beispiel mit 3 Faktoren A,B,C und ihren Interaktionen. In einem 1. Schritt werden die Gruppen 1 bis 3 der Faktoren (genauer: deren Dummies) aus der Gruppe der Kovariaten auspartiielliert. Zurück bleiben die Partialvariablen  $x^*_1, x^*_2, x^*_3, \dots$  der Kovariaten. Diese werden dann gruppenintern gegeneinander auspartiielliert. Danach können die Regressionskoeffizienten der Kovariaten gegenüber der abhängigen Variablen y bestimmt werden und die Streuung errechnet werden, die die Kovariaten in der abhängigen Variablen y erklären.

Danach müssen umgekehrt die 3 Gruppen der Faktoren an die Kovariaten angepasst werden. Wie oben für fitting constants I bereits ausgeführt, gibt es zwei Möglichkeiten der Anpassung, die *einmalige* und die *hierarchische*. In Almo ist nur die *hierarchische Anpassung* für die fitting constants II realisiert.

Betrachten wir nochmals die 2-er Interaktion **AB**. Um die Streuung zu ermitteln, die sie in der abhängigen Variablen  $y$  erklärt, müssen zuvor die Kovariaten aus der Gruppe der 2-er Interaktionen und der hierarchisch davor stehenden Gruppe auspartielliert werden. Die Kovariaten werden also auspartielliert aus

A B C D    AB AC AD BC BD CD **ACD BCD**

nicht jedoch aus den hierarchisch nachgeordneten Gruppen der 3-er und 4-er Interaktionen. Erst danach wird die oben beschriebene Vorgehensweise auf die Faktoren und ihre Interaktionen, die jetzt Partialvariable sind (die Kovariaten sind ja aus ihnen "auspartielliert"), angewendet.

### **Problem und Einschränkungen der fitting constants II**

Die fitting constants II haben daselbe Gebrechen wie das fitting constants I. Bei ungleichen (bzw. nicht-balancierten) Zellenhäufigkeiten und mehr als 2 Faktoren und ihren Interaktionen sind die Interaktionseffekte nicht ermittelbar. Siehe die Ausführungen oben. Bei Analysen ohne Interaktionen oder mit maximal zwei Faktoren und ihrer Interaktion und mit oder ohne Kovariate liefert das Verfahren jedoch eine eindeutige Kleinste-Quadrate-Lösung. Alle Effekte und alle sonstigen bedeutsamen Koeffizienten werden korrekt als Kleinste-Quadrate-Schätzer errechnet und ausgegeben.

### **Almo-Programm-Masken für fitting constants II**

Das Verfahren der fitting constants II kann mit dem Standard-Maskenprogramm **Prog20mo** gerrechnet werden. Dazu muss die Optionsbox "Verfahren" geöffnet werden.

Für 2, 3 und 4 Faktoren (und beliebig viele Kovariate) sind noch vier Sonder-Programm-Masken, **ProgFII2**, **ProgFII3**, **ProgFII4**, **ProgFII5**, in Almo enthalten. Man findet sie nach Klick auf den Knopf "Verfahren/Allgemeines lineares Programm" oder "alle Progs" am Oberrand des Almo-Fensters. Diese Programme verwenden spezielle Interaktionsvariable und die „Partial“-Anweisung aus der Almo-Programmiersprache. Siehe dazu Abschnitt P20.8.1, „Eingabebox: Option:Verfahren“ und die ausführliche Darstellung im 2. Teil dieses Handbuchs, Abschnitt P20.14 und P20.15.

Mit Prog20mo liefert Almo ein sehr sparsames Ergebnis. Es besteht aus der erklärten Streuung, deren Signifikanz und den partiellen Korrelationen. Im Fall der Kovarianzanalyse werden noch die erklärten Streuungen und die Regressions- und partiellen Korrelationskoeffizienten der Kovariaten ausgegeben. Die Sonderprogramme liefern dazu noch die Effekte der 1. Gruppe, d.h. die Haupteffekte.

### **Vergleich fitting constants I und II**

Das im vorausgehenden Abschnitt P20.7.1 dargestellte Verfahren der fitting constants I und das hier dargestellte der fitting constants II, erbringt das gleiche Ergebnis.

1. wenn die Interaktionen nicht miteinbezogen werden. Dies gilt auch für den Fall der Kovarianzanalyse. Man sollte dann die Programm-Maske Prog20mo einsetzen. Sie liefert ein umfangreiches Ergebnis.

- Sind nur die beiden nominalen Variablen A,B und ihre Interaktion AB und keine Kovariaten vorhanden, dann sind die Ergebnisse gleich. In diesem Fall sollte mit Prog20mo und Verfahren fitting constants I gerechnet werden, weil dann eine umfangreiche, vollständige Ergebnisliste von Almo erzeugt wird.  
Sind Kovariate vorhanden, dann entstehen identische Ergebnisse, wenn fitting constants I mit dem Sonderprogramm Prog20FI\_2 und fitting constants II mit dem Sonderprogramm Prog20FII2 gerechnet wird. Beide verwenden die hierarchische Anpassung von Kovariaten und nominalen Variablen. Tatsächlich sind die beiden Programme identisch.
- Liegen gleiche Zellenhäufigkeiten vor, dann erbringen alle Verfahren, also auch das sequentielle und das „weighted squares of means“-Verfahren das gleiche Ergebnis.

### Vergleich mit SPSS

In SPSS ist das Verfahren der fitting constants I nicht enthalten, jedoch das Verfahren der fitting constants II. Es wird dort als Modell **SS-Typ II** bezeichnet.

Der Leser rechne in Almo das Programm Prog20FII3. Er findet es nach Klick auf den Knopf "alle Progs" am Oberrand des Almo-Fensters. Gerechnet wird dabei eine Analyse mit 3 Faktoren und 2 Kovariaten.

Danach rechne er das entsprechende SPSS-Programm (das die gleichen Faktoren und Kovariaten verwendet). Es ist unter dem Namen "SPSS\_ABC\_Mod\_II.sps" als Syntaxprogramm im Almo-Ordner TESTDAT enthalten. Es verwendet die Datei "Testdat.sav", die sich ebenfalls im Almo-Ordner TESTDAT befindet. Zuerst werden die Daten und danach das Syntaxprogramm in SPSS geladen.

Das SPSS-Modell **SS-Typ II** wird nur und ausschließlich dazu verwendet, die durch die Faktoren und Kovariaten erklärten Streuungen und die aus diesen abgeleiteten Koeffizienten zu ermitteln. Sie werden unter der Überschrift "Tests der Zwischensubjekteffekte" ausgegeben. Die Ergebnisse stimmen mit den entsprechenden aus Almo exakt überein.

Ergebnis aus SPSS (etwas gekürzt)

Tests der Zwischensubjekteffekte					
Quelle	Quadratsumme SS-Typ II	F	Sig.	Partielles Eta-Quadrat	Beobachtete Trennschärfe
Konstanter Term	122,024	33,300	,000	,415	1,000
alter	,154	,042	,839	,001	,055
einkomm	,092	,025	,875	,001	,053
geschlec (FakA)	22,922	6,255	,016	,117	,688
wohnort (FakB)	,746	,204	,654	,004	,073
beruf (FakC)	14,703	2,006	,146	,079	,394
geschlec * wohnort (IntAB)	,104	,028	,867	,001	,053
geschlec * beruf (IntAC)	3,849	,525	,595	,022	,131
wohnort * beruf (IntBC)	5,928	,809	,451	,033	,180
geschl*wohn*beruf (IntABC)	1,000	,136	,873	,006	,070
Korrigiertes Modell	49,971	1,049	,424	,225	,536
Fehler	172,226				
Korrigierte Gesamtvariation	222,197				

\*\*\* Beobachtete Trennschärfe wird in Almo "Teststärke" genannt.

Ergebnis aus ALMO (zusammengefasst und etwas gekürzt)

alle Streuungen sind Abweichungs-Quadratsummen

Variable	alle Var.	FakA	FakB	FakC	IntAB
----------	-----------	------	------	------	-------

Gesamtstreuung	222.196721				
Fehlerstreuung	172.225813				
erklärte Streuung	49.970908	22.921588	0.745803	14.703453	0.104280
Korrelat. quadriert	0.224895	0.117458	0.004312	0.078658	0.000605
F-Wert erklärte Streuung	1.048995	6.255245	0.203528	2.006268	0.028458
Signifikanz: p	0.424285	0.015926	0.653595	0.143961	0.866896
Teststaerke von F	0.535818	0.687838	0.072665	0.393710	0.053135

Variable	IntAC	IntBC	IntABC	Alter	Einkomm
erklärte Streuung	3.848534	5.927614	1.000245	0.1537	0.0923
Korrelat. quadriert	0.021857	0.033273	0.005774	0.030*	0.023*
F-Wert erklärte Streuung	0.525128	0.808816	0.136482	0.042	0.025
Signifikanz: p	0.600336	0.455176	0.869075	0.840	0.874
Teststaerke von F	0.131446	0.180034	0.069778	0.0546	0.0528

\* nicht quadriert

Alle anderen Koeffizienten, die SPSS noch ausgibt (Parameter, Kontraste, Randmittel usw.) werden in SPSS nicht nach dem Modell SS-Typ II sondern nach dem Standard-Modell SS-Typ III (also dem Verfahren der weighted squares of means) berechnet. Das führt zu störenden Inkonsistenzen, wie wir im Folgenden zeigen.

In Almo werden auch diese Koeffizienten (allerdings nicht für die Interaktionen) nach dem Kleinste-Quadrate-Kalkül der fitting constants II berechnet. Will der Benutzer aus Almo jedoch dieselben Koeffizienten wie aus SPSS erhalten, dann muss er eine zweite Analyse nach dem Verfahren der weighted squares of means rechnen.

Betrachten wir ein Beispiel: Almo und SPSS geben z.B. für die Variable Geschlecht bzw. FakA eine erklärte Streuung von 22.922 aus. Siehe die beiden obigen Tabellen. Für den paarweisen Vergleich von FakA gibt Almo ebenfalls 22.922 aus. Das muss bei einem dichotomen Faktor so sein.

Ergebnis aus ALMO (gekürzt)

Paarweise Vergleiche (Differenzkontraste) von FakA

	Differenz	Standard- fehler	erklärte Streuung	t-Wert (LSD)	Signif. p	Test- staerke
A1 - A2	-1.3312	0.5323	22.9216	2.5010	0.0159	0.6885

SPSS gibt hingegen für den äquivalenten *Differenzkontrast* von Geschlecht 24.654 aus. Dieser Wert entsteht aus einer Analyse nach dem Verfahren der der weighted squares of means bzw SS-Typ III.

Ergebnis aus SPSS (gekürzt)

Paarweise Vergleiche für Geschlecht FakA

geschlec	geschlec	Mittlere Differenz	Standard Fehler	Sig.b
1,00	2,00	-1,400*	,540	,013
2,00	1,00	1,400*	,540	,013

Tests auf Univariate

	Quadratsumme	df	F	Sig.

Kontrast	24,654	1	6,728	,013
Fehler	172,226	47		

## P20.7.2 Das sequentielle Vefahren (SS Typ I)

### Die variablenweise hierarchische Auspartiellierung

Dieses Verfahren ist in den Statistiksystemen SAS und SPSS unter der Bezeichnung **SS Type I** enthalten.

Wird "Verfahren" auf „sequentiell" gesetzt, werden nicht, wie oben beschrieben, die Gruppen 1 bis 3 hierarchisch auspartielliert, sondern die Variablen. Wir nennen dies das Modell der "variablenweisen Hierarchie".

#### P20.7.2.1 Varianzanalyse mit dem sequentiellen Verfahren

Wir betrachten wieder das Beispiel mit 3 nominalen Variablen A,B,C und AB, AC, BC als 2-er Interaktionen und ABC als 3-er Interaktion. Es wird folgende Hierarchie der Variablen gebildet:

$$A / B / C / AB / AC / BC / ABC$$

Die Schrägstriche symbolisieren die hierarchische Abstufung. Die Variablen werden als Gruppen betrachtet. Eine Gruppe besteht also aus den Dummies einer Variablen. Diese Gruppen werden dann hierarchisch auspartielliert. Die Folge ist, dass aus A überhaupt nichts mehr "herausgenommen" wird und aus B nur die Wirkung von A. Lediglich aus der letzten der nominalen Variablen, aus C, werden die beiden anderen nominalen Variablen A und B "herausgenommen" - so dass dasselbe Ergebnis entsteht, wie beim "fitting constants I" Modell. Die Wirkungen, die die 3 Faktoren A, B, C auf die abhängigen Variablen haben, sind somit nicht vergleichbar - es sei denn, man will absichtlich eine Hierarchie dieser Variablen errichten. Will man voll auspartiellierte Koeffizienten für die Faktoren A, B, C, so muss man 3 Analysen rechnen, wobei jede Variable einmal als letzte Variable eingesetzt wird (oder man rechnet gleich ein "fitting constants I"-Modell. Bei gleichen Zellenhäufigkeiten sind die Faktoren und ihre Interaktionen unkorreliert. Das normale Modell der "gruppenweisen Hierarchie" und das der "variablenweisen Hierarchie" erbringen dann dieselben Ergebnisse.

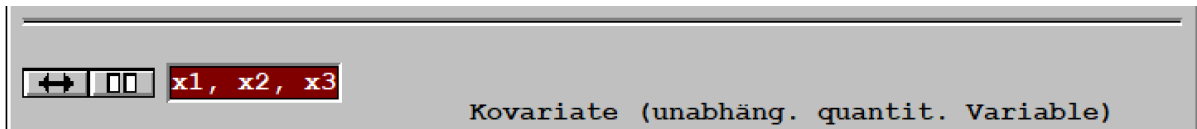
#### P20.7.2.2 Kovarianzanalyse mit dem sequentiellen Verfahren

Befinden sich Kovariate im Modell, so werden diese, äquivalent dem Verfahren der fitting constants I bzw. II, gegenseitig mit den nominalen Variablen und allen ihren Interaktionen auspartielliert.

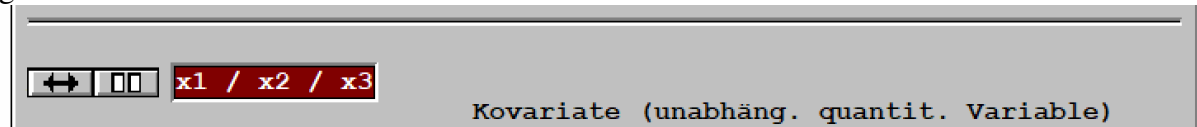
Wie oben für fitting constants I und II bereits ausgeführt, gibt es zwei Möglichkeiten der Anpassung von Kovariaten und nominalen Variablen, die *einmalige* und die *hierarchische*. In der Programm-Maske Prog20mo ist die *einmalige* Anpassung realisiert. In den Sonderprogrammen ProgSq\_2 bis ProgSq\_4 ist die *hierarchische* Anpassung realisiert.

Ähnlich wie beim Verfahren der fitting constants I und II entspricht die hierarchische Anpassung eher der Logik des sequentiellen Verfahrens. Wir empfehlen deswegen im Fall der Kovarianzanalyse die Sonderprogramme vorzuziehen. Auch SPSS verwendet standardmäßig die hierarchische Anpassung.

Sind 2 oder mehr Kovariate vorhanden, dann können sie hierarchisch oder gleichrangig angeordnet werden. In den oben genannten Maskenprogrammen werden, bei gleichrangiger Anordnung, die Kovariaten durch einen Beistrich getrennt. In unserem Beispiel muss in der Eingabebox "Analyse-Variable: Unabhängige Variable" eingetragen werden:



Sollen die Kovariaten hierarchisch geordnet werden, dann wird durch einen Schrägstrich getrennt



Wird hierarchisch angeordnet, dann bestimmt auch der Platz der Kovariaten das Ergebnis, das Almo für die Kovariaten ausgibt. Es ist nicht belanglos, wo die Kovariate in der Hierarchie steht.

Die Ergebnisse, die Almo für die Kovariaten ausgibt, sind davon abhängig, ob (1) die Anordnung der Kovariaten hierarchisch oder gleichrangig ist und (2) wenn sie hierarchisch ist, welchen Platz in der Hierarchie die jeweilige Kovariate einnimmt. Die Ergebnisse für die nominalen Variablen allerdings bleiben unbeeinflusst von der hierarchischen oder gleichrangigen Anordnung und auch vom Platz der jeweiligen Kovariaten in der Hierarchie.

### Almo-Programm-Masken für sequentielles Verfahren

Das sequentielle Verfahren kann mit dem Standard-Maskenprogramm **Prog20mo** gerechnet werden. Dazu muss die Optionsbox "Verfahren" geöffnet werden. Für den Fall der Kovarianzanalyse sollte diese Programm-Maske nicht verwendet werden. Die Kovariaten und nominalen Variablen werden *einmalig* aneinander angepasst. Das widerspricht der Logik der variablenweise hierarchischen Auspartiellerung, die beim sequentiellen Verfahren angewendet wird.

Sind keine Kovariaten vorhanden (Fall der Varianzanalyse) dann kann Prog20mo problemlos angewendet werden.

Besser geeignet für den Fall der Kovarianzanalyse sind die drei Sonderprogramm-Masken **ProgSq\_2**, **ProgSq\_3**, **ProgSq\_4** für 2, 3 und 4 Faktoren und beliebig viele Kovariate. Sie sind auch für die Varianzanalyse einsetzbar. Man findet sie nach Klick auf den Knopf "Verfahren/Allgemeines lineares Programm" oder "alle Progs" am Oberrand des Almo-Fensters. In den Sonderprogrammen werden die nominalen Variablen hierarchisch an die Kovariaten angepasst. Die Kovariaten können untereinander gleichrangig oder auch hierarchisch angeordnet werden.

### Vergleich mit SPSS

In SPSS ist das sequentielle Verfahren unter der Bezeichnung "Modell **SS-Typ I**" enthalten. Im Fall der Kovarianzanalyse werden Kovariate und Faktoren hierarchisch angepasst. Die Kovariaten werden als hierarchisch gereiht betrachtet. Der Platz, den die jeweiligen Kovariate in der hierarchischen Reihung einnimmt ist also für das Ergebnis bedeutsam. Die erklärten Streuungen und die von diesen abgeleiteten Koeffizienten werden unter der Überschrift "Tests der Zwischensubjekteffekte" ausgegeben. Die Ergebnisse stimmen mit den entsprechenden aus Almo exakt überein.

Wir haben bereits beim Verfahren der fitting constants II (SS-Typ II) ausgeführt, dass alle anderen Koeffizienten, die SPSS noch ausgibt (Parameter, Kontraste, Randmittel usw.) nicht nach dem Modell SS-Typ I sondern nach dem Standard-Modell SS-Typ III (also dem Verfahren der weighted squares of means) berechnet. Das führt, wie gezeigt wurde, zu störenden Inkonsistenzen. In Almo werden auch diese Koeffizienten nach dem Kleinst-

Quadrate-Kalkül des sequentiellen Verfahrens berechnet. Will der Benutzer aus Almo jedoch dieselben Koeffizienten wie aus SPSS erhalten, dann muss er eine zweite Analyse nach dem Verfahren der weighted squares of means rechnen.

### **P20.7.2.3 Sonderprogramme und Standard-Programm-Maske Prog20mo**

Almo enthält diese Schätzverfahren

```

weighted squares of means (SS-Typ III) (voreingestellt)
fitting constants I
fitting constants II      (SS-Typ II)
sequentiell Verfahren    (SS-Typ I)

```

Will der Benutzer ein Allgemeines Lineares Programm (ALM) rechnen dann wird er standardmäßig die Programm-Maske Prog20mo verwenden und mit dem (voreingestellten) Verfahren "weighted squares of means" rechnen. Dieses Verfahren wird auch SS-Typ III genannt. Die Programm-Maske Prog20mx entspricht dem Prog20mo, bietet dem Benutzer aber keine Optionen, die ihm zusätzliche Informationen liefern könnten.

Prog20mo bietet dem Benutzer jedoch auch an, mit den eher seltener verwendeten Verfahren der "fitting constants I und II sowie dem sequentiellen Verfahren zu arbeiten. Befinden sich auf Seiten der unabhängigen Variablen zusätzlich zu den nominalen Variablen (den Faktoren) auch noch quantitative (die Kovariaten) - dies ist der Fall der Kovarianzanalyse - dann treten bei der Programm-Maske Prog20mo Komplikationen auf. Prog20mo verwendet in diesem Fall die in Abschnitt P20.7.1.0.2 dargestellt "einmalige Anpassung" der Faktoren an die Kovariaten. "Modellgerechter" ist jedoch die "hierarchische Anpassung". Diese wird in den Sonderprogrammen eingesetzt.

Unsere **Empfehlung** lautet:

1. Analysen mit dem (voreingestellten) Verfahren der "weighted squares of means" werden immer mit Prog20mo gerechnet.
2. Soll eines der drei anderen Schätzverfahren eingesetzt werden, dann gilt: Die Regressionsanalyse und die Varianzanalyse werden mit Prog20mo gerechnet, die Kovarianzanalyse mit den Sonderprogrammen.
3. Ausnahme: Kovarianzanalysen ohne Interaktionen werden mit Prog20mo gerechnet.
4. Analysen mit nur *einem* nominalen Faktor werden immer mit Prog20mo gerechnet.

Da unter bestimmten Bedingungen einmalige und hierarchische Anpassung identisch sind, kann von dieser pauschalen Empfehlung abgewichen werden. Differenzierter sind also die Empfehlungen in nachfolgender Tabelle.

Dies sind die Sonderprogramme. Man findet sie nach Klick auf den Knopf "Verfahren/ Allgemeines lineares Modell" oder auf den Knopf "alle Progs" am Oberrand des Almo-Fensters.

```

*1 Sonder-Programme für fitting constants I
ProgFI_2 für Varianz- und Kovarianzanalysen mit 2 Faktoren und Interaktionen
ProgFI_3                                     3                               Interaktionen
ProgFI_4                                     4                               Interaktionen
ProgFI_5                                     5                               Interaktionen

*2 Sonder-Programme für fitting constants II
ProgFII2 für Varianz- und Kovarianzanalysen mit 2 Faktoren und Interaktionen
ProgFII3                                     3                               Interaktionen
ProgFII4                                     4                               Interaktionen
(alle 3 Prog mit zusätzlicher Ausgabe der Haupteffekte)

```

\*3 Sonder-Programme für sequentielles Verfahren  
**ProgSq\_2** für Varianz- und Kovarianzanalysen mit 2 Faktoren und Interaktionen  
**ProgSq\_3** 3 Interaktionen  
**ProgSq\_4** 4 Interaktionen

In der folgenden Tabelle wird unterschieden zwischen Varianzanalyse bei der auf Seiten der unabhängigen Variablen nur nominale Variable (Faktoren) vorhanden sind und Kovarianzanalyse bei der zusätzlich noch quantitative Variable (Kovariate) dazu kommen.

Und es wird unterschieden, ob eine Analyse gerechnet werden soll (1) ohne Interaktionen oder (2) mit allen bis zur höchsten Interaktion ("saturiertes Modell") oder mit reduzierten Interaktionen, d.h. Interaktion unterhalb der höchsten Ordnung.

Die abhängige Variable kann (1) quantitativ/ordinal oder (2) nominal dichotom oder polytom sein. (3) Auch multivariate Analysen mit mehreren abhängigen quantitativen/ordinalen Variablen sind möglich.

Varianzanalyse =====	keine Interaktionen -----	alle Interaktionen -----	reduzierte Interaktionen -----
w.squares of means (SS-Typ III)	Prog20mo	Prog20mo	Prog20mo
fitting constants I	Prog20mo	Prog20mo und *1)	Prog20mo und *1)
fitting constants II (SS-Typ II)	Prog20mo	Prog20mo und *2)	Prog20mo und *2)
sequentiell (SS-Typ I)	Prog20mo	Prog20mo und *3)	Prog20mo und *3)
Kovarianzanalyse =====			
w.squares of means (SS-Typ III)	Prog20mo	Prog20mo	Prog20mo
fitting constants I	Prog20mo	*1)	*1)
fitting constants II	Prog20mo + *2)	Prog20mo + *2)	Prog20mo + *2)
sequentiell	*3)	*3)	*3)

Beispiel: 1. Gerechnet werden soll eine Kovarianzanalyse ohne Interaktionen nach dem Verfahren der fitting constants I. Verwendet werden sollte die Programm-Maske Prog20mo.

2. Gerechnet werden soll eine Kovarianzanalyse mit allen Interaktionen nach dem Verfahren der fitting constants II. Verwendet werden können beide, die Standard-Programm-Maske Prog20m oder eines der Sonderprogramme vom Typ \*2. Die Ausgabe aus den beiden Programmen ist etwas verschieden. Bei \*2 werden die Haupteffekte und die Interaktionseffekte höchster Ordnung ausgegeben, die bei Prog20mo fehlen.

Die Programm-Masken der Sonderprogramme sind gegenüber denen von Prog20mo bzw. Prog20mx geringfügig verschieden. Betrachten wir Maske ProgFI\_4 und dort die Eingabebox für die Variablennamen.

Variablenamen

Datei der Variablenamen
Hilfe

".\Testdat\Varnamen.nam"

zeige

zeige = Namensdatei in Output zeigen  
 leer = nicht zeigen

---

Freie Namensfelder
Hilfe

erzeuge zusätzliche Namensfelder für weitere Variable

---

Setzen Sie hier die Variablennummern  
der 4 unabhängigen nominalen Variablen (Faktoren) ein  
(entsprechend ihrem Platz im Datensatz)

Name	1	=FakA
Name	2	=FakB
Name	3	=FakC
Name	4	=FakD

Die Namen für die unabhängigen nominalen Variablen (Faktoren) werden vom Programm vorgegeben. Sie lauten bei einer Analyse mit 4 Faktoren "FakA, FakB, FakC, FakD". Der Benutzer muss im unteren Abschnitt der Box in die rot unterlegten Eingabefelder die Nummern der Variablen einsetzen, die er als FakA, FakB, ... etc. in seiner Analyse verwenden will. In der Regel ist die Nummer identisch mit dem Platz der Variablen im Datensatz. Die Variablenamen für die abhängige Variable und die Kovariaten bestimmt der Benutzer. Er kann sie im Abschnitt Freie "Namensfelder" eintragen oder in eine "Datei der Variablenamen" einschreiben oder beides.

### P20.7.3 Das Verfahren der "weighted squares of means" (SS Typ III)

Die gegenseitige Auspartiellierung

Dies ist das Standardverfahren des Allgemeinen linearen Modells.

Wir betrachten unser Beispiel mit 3 Faktoren A, B, C. Wenn sich im zu analysierenden Modell die unabhängigen nominalen Variablen A,B,C, ihre Interaktionen 2. Ordnung AB, AC, BC und ihre Interaktion 3. Ordnung ABC befinden, so werden diese nun voll gegeneinander auspartielliert. Es existiert keine Hierarchie. Die Haupteffekte von A werden z.B. auch an die Interaktionen angepasst und nicht nur umgekehrt, wie bei der Methode der "fitting constants".

Almo arbeitet in diesem Falle mit der 0,1,-1 - Kodierung der Dummy-Variablen und nicht, wie in den beiden anderen Methoden mit der 0,1-Kodierung. Die Ergebnisse, die bei dieser Vorgehensweise entstehen, stimmen mit denen des von Yates 1934 entwickelten Verfahrens der "weighted squares of means" überein. Bei SAS und SPSS wird das Verfahren mit "SS-Typ III" bezeichnet und auch dort als Standardverfahren empfohlen

Befinden sich Kovariate im Modell, dann werden diese, wie bei den anderen beschriebenen Verfahren gegenseitig mit den nominalen Variablen und ihrer Interaktionen auspartielliert.

**P20.7.3.1 Der Kalkül des Verfahrens der "weighted squares of means" (SS Typ III)**

Der Kalkül der "weighted squares of means" wird in Almo - so wie der der "fitting constants" und des sequentiellen Modells - als Regressionskalkül mit Dummy-Variablen ausgeführt. Der Kalkül der "weighted squares of means" ist jedoch sehr anschaulich darstellbar wenn man den Kalkül von Yates verwendet (siehe Winer, 1971, S. 417).

Betrachten wir zunächst ein einfaches Modell mit A und B als unabhängigen nominalen Variablen und ihrer Interaktion AB.

Die Daten seien folgende:

Daten

	B1	B2
A1	23.5 23.7	28.7
A2	8.9	5.6 8.9
A3	10.3 12.54	13.6 14.6

In der Zelle A1B1 befinden sich 2 Probanden mit den Werten 23.5 und 23.7 in der abhängigen Variablen etc.

Tabelle der Mittelwerte

	B1	B2	$\bar{A}_i$
A <sub>1</sub>	23.6	28.7	26.15
A <sub>2</sub>	8.9	7.25	8.075
A <sub>3</sub>	11.4	14.1	12.75
$\bar{B}_k$	14.63	16.68	$\bar{G}=15.6583$

Zunächst werden die Zellenmittelwerte  $\overline{A_i B_k}$  errechnet und aus diesen der Gesamtmittelwert  $\bar{G}=15.6583$ . Auch die Ausprägungsmittelwerte  $\bar{A}_i$  und  $\bar{B}_k$  werden aus den jeweiligen Zellenmittelwerten errechnet.

Die Haupteffekte ergeben sich dann gemäß

$$\alpha_i = \bar{A}_i - \bar{G}$$

$$\beta_k = \bar{B}_k - \bar{G}$$

Die Interaktionseffekte entstehen aus

$$\alpha\beta_{ik} = \overline{A_i B_k} - \alpha_i - \beta_k - \bar{G}$$

Der Effekt der Konstante ist

$$\tau = \bar{G}$$

so erhalten wir beispielsweise für

$$\alpha_1 = \bar{A}_1 - \bar{G} = 26.15 - 15.6583 = 10.49$$

$$\beta_1 = \bar{B}_1 - \bar{G} = 14.63 - 15.6583 = -1.025$$

$$\alpha\beta_{11} = \overline{A_1B_1} - \alpha_1 - \beta_1 - \overline{G} = 23.6 - 10.49 + 1.025 - 15.6583 = -1,525$$

$$\tau = \overline{G} = 15.6583$$

Wir sehen, dass die Effekte inhaltlich sehr sinnvoll interpretierbar sind. Der **Konstanteneffekt** ist gleich dem Mittelwert aus den Zellenmittelwerten. Der **Haupteffekt** ist gleich dem Mittelwert aus den Zellenmittelwerten der betreffenden Ausprägung minus dem **Konstanteneffekt**. Der **Interaktionseffekt** ist gleich dem Zellenmittelwert minus dem Konstanteneffekt und minus den betreffenden Haupteffekten. Wichtig ist festzuhalten, dass für die Effekte die unterschiedlichen Zellenhäufigkeiten nicht berücksichtigt werden. Insofern ist das Verfahren der "weighted squares of means" bezogen auf die Daten nicht als "Kleinst-Quadrate"-Schätzung zu bezeichnen.

Die durch A erklärte Streuung ergibt sich gemäß

$$SS_A = \left( \sum h_i \alpha_i^2 - \frac{(\sum h_i a_i)^2}{\sum h_i} \right)$$

Die Summierung erfolgt von  $i=1$  bis  $i=q$

wobei  $q =$  Zahl der Ausprägungen von A (in unserem Beispiel:3)

$h_i =$  harmonisches Mittel der Zellenhäufigkeiten für die Ausprägung  $A_i$ .

Die Streuung für B ergibt sich äquivalent.

In Almo wird **nicht** nach diesem Kalkül gerechnet. In Almo wird die 0,1,-1-Dummy-Kodierung verwendet und dann alle Dummies gegeneinander auspartielliert. Die 0,1, -1-Dummy-Kodierung entspricht den Restriktionen

$$\sum \alpha_i = 0 \quad (\text{summiert über } i)$$

$$\sum \beta_k = 0 \quad (\text{summiert über } k)$$

$$\sum_k \alpha_i \beta_k = 0 \quad (\text{summiert über } k)$$

$$\sum_i \alpha_i \beta_k = 0 \quad (\text{summiert über } i)$$

Diese Restriktionen sind für die Lösung notwendig. Die beiden Restriktionen für die Interaktionseffekte sagen aus, dass diese, über die Spalten summiert, 0 ergeben und ebenso über die Zeilen summiert auch 0 ergeben.

Der Kalkül der "fitting constants I und II" und des sequentiellen Verfahrens kann nicht so anschaulich dargestellt werden. Wichtig ist es festzuhalten, dass die Effekte bei der "fitting constants I und II" nicht aus den Zellenmittelwerten alleine, sondern auch unter Berücksichtigung der Zellenhäufigkeiten berechnet werden. Das hat Vorteile und Nachteile.

*Vorteil:*

Die Effekte werden bei "fitting constants I und II" nach dem bewährten Prinzip der *Kleinsten Quadrate* berechnet. Starke Disproportionalität der Zellenhäufigkeiten werden berücksichtigt.

*Nachteil:*

(1) Bei ungleichen Zellenhäufigkeiten und mehr als 2 unabhängigen nominalen Variablen sind die Interaktionseffekte nicht mehr eindeutig bestimmbar. Es treten „wechselnde Werte“ auf. Siehe P20.6.5.1.

(2) Die Ergebnisse sind stichprobenabhängig. Damit ist folgendes gemeint: In der experimentellen Forschung wird man bemüht sein, gleiche bzw. balancierte Zellenhäufigkeiten zu erreichen. Dabei können zufällig ungleiche Zellenhäufigkeiten entstehen (weil z.B. einige Untersuchungseinheiten ausfallen). Dieser Fall ist unproblematisch. Das Verfahren der "weighted squares of means" ist hier angebracht. In der nicht-experimentellen Forschung können teilweise sehr stark korrelierende unabhängige nominale Variable und demzufolge sehr stark disproportionale Zellenhäufigkeiten auftreten. Betrachten wir folgendes Beispiel: Als unabhängige nominale Variable werden Schulbildung und Beruf verwendet. Aus einer Zufallsstichprobe ergeben sich folgende Zellenhäufigkeiten:

		Beruf		
		manuell	administrativ	kreativ
Schulbildung	Volks- und Hauptschule	124	73	2
	Gymnasium	7	23	7
	Universität	2	12	4

Hier stellt sich die Frage, ob es sinnvoll ist - wie für die "weighted squares of means" notwendig - einen Zellenmittelwert einmal aus 124 Personen und einmal aus 2 Personen zu ermitteln und daraus dann Effekte zu errechnen. Hier scheint es sinnvoll zu sein, das Verfahren der "fitting constants I oder II" anzuwenden.

#### P20.7.4 Vergleich der Verfahren

Wir wollen zuerst den Begriff der gleichen oder balancierten Zellenhäufigkeiten klären. Betrachten wir folgende Zellenhäufigkeiten:

	B1	B2	$n_i$
A <sub>1</sub>	5	10	15
A <sub>2</sub>	15	30	45
A <sub>3</sub>	10	20	30
$n_k$	30	60	$n=90$

Hier gilt für jede Zellenhäufigkeit

$$n_{jk} = \frac{n_i \cdot n_k}{n}$$

Für die Zelle A<sub>1</sub>B<sub>1</sub> gilt beispielsweise

$$5 = \frac{15 \cdot 30}{90}$$

Obige Tabelle besteht aus "balancierten" (oder proportionalen) Zellenhäufigkeiten. Wenn wir im Folgenden von "gleichen" Zellenhäufigkeiten sprechen, dann ist dieser Fall der balancierten Zellenhäufigkeiten immer mit eingeschlossen.

1. Bei gleichen (oder balancierten) Zellenhäufigkeiten sind die Ergebnisse der 4 Verfahren identisch.
2. Bei nur *einer* unabhängigen nominalen Variablen, beliebig vielen Kovariaten und ungleichen Zellenhäufigkeiten sind die Streuungen und erklärten Streuungen, die die 4 Verfahren liefern, gleich. Die Effekte und Kontraste sind jedoch verschieden.

3. Werden bei gleichen aber auch ungleichen Zellenhäufigkeiten die Interaktion nicht miteinbezogen, dann sind die Ergebnisse aus den „fitting constants I“ und den „fitting constants II“ identisch.
4. Sind bei gleichen oder ungleichen Zellenhäufigkeiten nur 2 nominale Variable A und B und ihre Interaktion AB vorhanden, dann sind die erklärten Streuungen „fitting constants I“ und den „fitting constants II“ identisch. Wenn sich noch Kovariate im Modell befinden, dann gilt dies bei ungleichen Zellenhäufigkeiten nur, wenn fitting constants I mit dem Sonderprogramm ProgFI\_2 gerechnet wird (siehe Abschnitt P20.7.1.0.2 über hierarchische Anpassung).
5. Werden (bei ungleichen Zellenhäufigkeiten) die Interaktionen nicht miteinbezogen, dann sind zwischen "fitting constants I" bzw. „fitting constants II“ und "weighted squares of means" zwar die Effekte  $i, k, \dots$  verschieden, die durch die unabhängigen nominalen Variablen A, B, ... erklärten Streuungen sind jedoch gleich. Ebenso sind die paarweisen Vergleiche gleich.
6. Bei ungleichen Zellenhäufigkeiten und 2 oder mehr unabhängigen nominalen Variablen A, B, ... und Einschluss der Interaktionen sind die Ergebnisse der 4 Verfahren verschieden - mit Ausnahme von Punkt 4 und mit Ausnahme der durch die Interaktion höchster Ordnung erklärten Streuung. Diese ist bei allen 4 Verfahren gleich.
7. Die durch das Gesamtmodell erklärte Streuung ist bei allen 4 Verfahren gleich.
8. Die Fehlerstreuung (die in Almo residual als verbleibende nicht erklärte Streuung ermittelt wird) ist demzufolge auch gleich.

All diese Aussagen gelten auch für den Fall, dass sich Kovariate im Modell befinden (mit Ausnahme von Punkt 4). Für den Almo-Benutzer stellt sich die Frage, welches Verfahren er wählen soll. Eine eindeutige Antwort ist hier nicht möglich. Wir wollen folgende

**Empfehlung** geben:

1. Bei gleichen Zellenhäufigkeiten liefern alle Verfahren das gleiche Ergebnis. Wir empfehlen hier "Verfahren=w\_squares\_of\_means;" zu setzen, weil dabei in Almo die übersichtlichste Ausgabe entsteht.
2. Das "sequentielle" Verfahren wird man nur anwenden, wenn man aus inhaltlichen, theoretischen Gründen eine sequentielle Hierarchie der Variablen unterstellen möchte.
3. Bei ungleichen Zellenhäufigkeiten und Ausschluss der Interaktionen sind "weighted squares of means" und die beiden "fitting constants"-Verfahren gleichwertig. Die erklärten Streuungen sind zwar dieselben, die Effekte werden bei den beiden "fitting constants" als "Kleinste-Quadrate"-Schätzer ermittelt. Ist nur *eine* unabhängige nominale Variable vorhanden (und eventuell auch noch Kovariate) dann sollte man die fitting constants I verwenden.
4. Bei ungleichen Zellenhäufigkeiten und 2 unabhängigen nominalen Variablen A und B und ihrer Interaktion AB sind "weighted squares of means" und die beiden "fitting constants"

gleichwertig. Sie erbringen allerdings bei ungleichen Zellenhäufigkeiten nicht dieselben Ergebnisse.

5. Bei ungleichen Zellenhäufigkeiten und mehr als 2 unabhängigen nominalen Variablen und ihren Interaktionen muss "weighted squares of means" oder „fitting constants II“ gewählt werden, da bei den "fitting constants I" bei den Interaktionseffekten "wechselnde Werte" auftreten – es sei denn, man möchte sich diese ausgeben lassen.
6. Bei (einigen) leeren Zellen ist eher das Verfahren der fitting constants I angebracht, das dann ohne Interaktion gerechnet werden muss.

Aus dieser komplizierten Liste von Empfehlungen lässt sich eine „General-Empfehlung“ für den Benutzer ableiten:

*Wenn Sie jetzt nicht mehr durchblicken, dann verwenden Sie das Verfahren der „weighted squares of means“ (SS Typ III). Bei jeglicher Konstellation können Sie dabei keinen (zu großen) Fehler machen !!*

Wir wollen an einem Beispiel die Unterschiede der Verfahren darstellen. Wir betrachten ein Modell mit A und B und ihrer Interaktion AB als unabhängigen nominalen Variablen.

Wir rechnen folgendes Almo-Programm, das als Syntaxprogramm "Ungleich.Alm" in Almo vorhanden ist. Sie finden das Programm durch Klick auf den Knopf "alle Progs" am Oberrand des Almo-Fensters. Die Daten sind folgende

A	B	Y
1	1	23.5
1	1	23.7
1	2	28.7
2	1	8.9
2	2	5.6
2	2	8.9
3	1	10.3
3	1	12.5
3	2	13.6
3	2	14.6

Wir erhalten u.a. folgende Ergebnisse:

(Beachte, dass in unserem Beispiel mit nur 2 Faktoren und deren Interaktion „fitting constants“ und „fitting constants II“ dasselbe Ergebnis erbringen.)

	weighted squares of means	fitting constants I + II	sequentiell
-----	-----	-----	-----
Gesamtstreuung	528.8610	528.8610	528.8610
gesamte erklärte Streuung	520.4760	520.4760	520.4760
Fehlerstreuung erklärte Streuung	8.3850	8.3850	8.3850
durch A	479.1079	499.1203	494.0310
B	9.4556	10.7143	10.7143

Effekte von A	AB	15.7307	15.7307	15.7307
	A1	10.4917	10.6271	10.2700
	A2	-7.5833	-7.5871	-7.2300
	A3	-2.9083	-2.2800	-2.2800
Effekte von B	B1	-1.0250	-1.0714	wechselnde Werte
	B2	1.0250	1.0714	wechselnde Werte
Effekte von AB	AB11	-1.5250	-0.9857	-0.9857
	AB12	1.5250	1.9714	1.9714
	AB21	1.8500	2.5286	2.5286
	AB22	-1.8500	-1.2643	-1.2643
	AB31	-0.3250	-0.2786	-0.2786
	AB32	0.3250	0.2786	0.2786

Die "wechselnden Werte" für die Effekte von B beim "sequentiellen" Verfahren sind folgende:

		wechselnde Werte von B
B1 wenn	A1	1.0743
	A2	1.4286
	A3	1.0714
B2 wenn	A1	1.4286
	A2	0.7143
	A3	1.0714

Folgende Feststellungen können getroffen werden:

1. Gesamte erklärte Streuung, Fehlerstreuung und die durch die Interaktion (höchster Ordnung) erklärte Streuung sind bei allen Verfahren gleich.
2. Die durch die letzte unabhängige nominale Variable (in unserem Beispiel: B) erklärte Streuung ist bei "fitting constants" und "sequentiellen" Verfahren gleich.
3. Die durch A, B und AB erklärte Streuung kann beim "sequentiellen" Verfahren, aber nicht bei den beiden anderen, additiv zur "gesamten erklärten Streuung" zusammengefasst werden, d.h. es gilt beim sequentiellen Verfahren:

$$SS_t = SS_A + SS_B + SS_{AB} + SS_E$$

$$528.861 = 494.031 + 10.7143 + 15.7307 + 8.3850$$

$SS_t$  = Gesamtstreuung  
 $SS_A, SS_B, SS_{AB}$  = durch A bzw. B bzw. AB erklärte Streuung  
 $SS_E$  = Fehlerstreuung

#### ***P20.7.4.1 Vergleich mit SAS und SPSS***

In SAS und SPSS werden folgende Schätz-Verfahren zur Berechnung der erklärten Streuungen unterschieden:

1. SS Typ I: Ist identisch mit dem "sequentiellen" Verfahren in Almo. Im Falle der Kovarianzanalyse muss in Almo allerdings mit den Sonderprogrammen ProgSq\_2 bis ProgSq\_4 gerechnet werden. Siehe dazu die ausführliche Darstellung in Abschnitt

P20.7.2.2.

2. SS Typ II: Ist identisch mit dem Verfahren der "fitting constants II" in Almo. Siehe dazu die ausführliche Darstellung in Abschnitt P20.7.1.1
3. SS Typ III: Ist identisch mit dem Verfahren der "weighted squares of means" in Almo.
4. SS Typ IV: SPSS empfiehlt, dieses Verfahren zu verwenden, wenn leere Zellen vorhanden sind. Die Ergebnisse, die das Verfahren liefert, sind nicht selten drastisch anders, wie wenn mit Typ III gerechnet wird. Sind keine leeren Zellen vorhanden, dann ist SS Typ IV identisch mit SS Typ III. Die Art und Weise, wie Almo das Problem leerer Zellen behandelt, wird in nachfolgendem Abschnitt P20.7.7 ausgeführt.

Für die GLM-Prozedur von SPSS gilt: Unabhängig davon, welchen SS Typ der Benutzer vorgibt, es werden immer dieselben Parameter (und die aus ihnen abgeleiteten Koeffizienten) berechnet. Wir werden darauf gegen Ende von Abschnitt P20.7.5 zurückkommen.

### P20.7.5 Berechnung der Effekte

Betrachten wir als Beispiel eine Varianzanalyse mit 2 unabhängigen nominalen Variablen (Faktoren) A und B und einer abhängigen quantitativen Variablen y. Siehe Beispielprogramm Gleich.Alm in Almo. Sie finden es durch Klick auf das Menü „Almo/Liste aller Almo-Programme“.

Die Daten sind folgende:

	B1		B2	
A1	23.5	23.7	29.5	28.7
A2	4.8	8.9	5.6	8.9
A3	10.3	12.5	13.6	14.6

A besitzt 3 Ausprägungen, B besitzt 2. In der Zelle A1B1 sind 2 y-Werte enthalten: 23.5 und 23.7.

Die Besonderheit dieser Datenmatrix ist: Die Zellenhäufigkeiten sind **gleich** (jeweils 2).

Almo liefert u. a. folgende Mittelwerte:

		Zellenmittelwerte		
		B1	B2	Mittelwerte von A1 A2 A3
A1		23.6	29.1	26.35
A2		6.85	7.25	7.05
A3		11.4	14.1	12.75
Mittelwerte von B1 B2		13.95	16.8167	15.3833

Beispiel: Der Mittelwert für A1 ist  $\bar{A}_1 = 26.35$ . Er ergibt sich als Mittelwert der 4 y-Werte für A<sub>1</sub>, aber auch (da wir gleiche Zellenhäufigkeiten haben) als Mittelwert aus den 2 Zellenmittelwerten 23.6 und 29.1.

Wir wollen hier gleich vorwegnehmen: Bei ungleichen Zellenhäufigkeiten (und dem Verfahren der "weighted squares of means") werden die Ausprägungsmittelwerte  $\bar{A}_i$  aus den 2 Zellenmittelwerten errechnet und nicht aus den einzelnen y-Werten.

Der Gesamtmittelwert über alle 12 y-Werte ist  $\bar{G} = 15.3833$

Mit der Varianzanalyse bestimmen wir folgende Gleichung

$$(1) y = \alpha_i + \beta_j + \alpha\beta_{ij} + \tau$$

$\tau$  = Konstanteneffekt

$\alpha_i$  = die Haupteffekte von A

$\beta_j$  = die Haupteffekte von B

$\alpha\beta_{ij}$  = die Interaktionseffekte der Interaktionen AB

Almo liefert uns folgende Effekte

$$\tau = 15.3833$$

$$\alpha_1 = 10.9667$$

$$\alpha_2 = -8.3333$$

$$\alpha_3 = -2.6333$$

$$\beta_1 = -1.4333$$

$$\beta_2 = 1.4333$$

$$\alpha\beta_{11} = -1.3167$$

$$\alpha\beta_{12} = 1.3167$$

.

.

$$\alpha\beta_{32} = -0.0833$$

Wir erkennen

1. Der Konstanteneffekt  $\tau$  ist gleich dem Gesamtmittelwert

$$(2) \tau = \bar{G}$$

2. Ein Haupteffekt ist gleich dem Ausprägungsmittelwert minus dem Gesamtmittelwert (bzw. Konstanteneffekt)

$$(3a) \alpha_i = \bar{A}_i - \bar{G}$$

$$(3b) \beta_j = \bar{B}_j - \bar{G}$$

$\bar{A}_i, \bar{B}_j$  = Ausprägungsmittelwert von  $A_i$  bzw.  $B_j$

Beispiel:

$$\alpha_1 = \bar{A}_1 - \bar{G} = 26.35 - 15.3833 = 10.9667$$

Wir können den Begriff "Haupteffekt" nun definieren:

Der Haupteffekt  $a_i$  ist das Ausmaß, in dem der mittlere  $y$ -Wert unter der Ausprägung  $A_i$  vom Gesamtmittelwert verschoben wird.

Betrachten wir ein konkretes Beispiel:  $y$  sei das Einkommen.

A sei der Beruf mit den 3 Ausprägungen

A1=Unternehmer,

A2=Arbeiter

A3=Angestellter.

Unternehmer zu sein hat den "Effekt", dass das eigene Einkommen um  $a_1=10.9667$  Geldeinheiten vom Gesamtmittelwert in positiver Richtung verschoben wird.

Arbeiter bzw. Angestellter zu sein hat den Effekt, dass das Einkommen um  $a_2= -8.3333$

bzw.  $a_3 = -2.633$  Geldeinheiten vom Gesamtmittelwert in negativer Richtung verschoben wird.

3. Ein Interaktionseffekt  $ij$  ist gleich dem Zellenwert minus dem Gesamtmittelwert und der beiden Haupteffekten

$$(4) \alpha\beta_{ij} = \overline{AB}_{ij} - (\overline{G} + \alpha_i + \beta_j)$$

Beispiel:

$$\alpha\beta_{11} = 23.6 - (15.3833 + 10.9667 - 1.4333) = -1.3167$$

Wenn der Zellenmittelwert vollständig durch den Gesamtmittelwert plus den beiden Haupteffekten erklärt wird, dann ist die Interaktion = 0

Nun wollen wir den  $y$ -Wert einer Untersuchungseinheit prognostizieren, die die Ausprägungen  $A_1$  und  $B_1$  besitzt. Wir setzen in Gleichung 1 ein und erhalten

$$y' = \tau + \alpha_1 + \beta_1 + \alpha\beta_{11}$$

$$y' = 15.3833 + 10.9667 - 1.4333 - 1.3167 \\ = 23.6$$

D.h. wir prognostizieren den Zellenmittelwert  $\overline{AB}_{11}$ .

Wir bezeichnen die Art und Weise, wie in Punkt 1 bis 3 die Effekte beschrieben wurden, als die "klassische" Interpretation der Effekte – klassisch deswegen, weil sie so auf R.A. Fisher zurückgeht.

Aus den Haupteffekten lassen sich schnell die **paarweisen Vergleiche** ermitteln. Wir geben Sie nur für A aus.

	Differenz	Standard- fehler	erklärte Streuung	t-Wert (LSD)	Signifikanz p	(1-p)100	Test- stärke
A1 - A2	19.3000	1.1941	744.9800	16.1630	0.0001	99.99%	1.0000
A1 - A3	13.6000	1.1941	369.9200	11.3895	0.0002	99.98%	1.0000
A2 - A3	-5.7000	1.1941	64.9800	4.7735	0.0036	99.64%	0.9765

Freiheitsgrade fuer t-Wert: 6

### Wie Almo rechnet

Almo rechnet nicht so anschaulich, wie hier dargestellt. Die nominalen Variablen werden in Dummies aufgelöst (beim Verfahren der "fitting constants" in 0,1 kodierte Dummies, beim Verfahren der "weighted squares of means" in 0, 1, -1 kodierte Dummies – siehe P20.4). Dann wird eine Regressionsanalyse gerechnet. Dabei entstehen Regressionskoeffizienten, die *noch nicht* identisch mit den Effekten sind.

Wird mit dem Verfahren der "**weighted squares of means**" gerechnet (siehe P20.7.3), dann entstehen folgende Regressionskoeffizienten

für die Dummy	a1	10.9667
	a2	-8.3333
	a3	0

Die ersten beiden sind identisch mit den Haupteffekten  $\alpha_1$  und  $\alpha_2$ . Der Effekt  $\alpha_3$  ergibt sich aus

$$\begin{aligned}\alpha_3 &= 0 - (10.9667 - 8.3333) \\ &= -2.6333\end{aligned}$$

Demzufolge ist die Summe aller Haupteffekte = 0

Allgemein gilt (beim Verfahren der "weighted squares of means"): Die Effekte – außer dem der letzten Dummy – sind gleich den Regressionskoeffizienten

$$(5) \alpha_i = r_i$$

Der Effekt der letzten Dummy  $\alpha_z$  ergibt sich aus

$$(5b) \alpha_z = r_z - \sum_{i=1}^{z-1} r_i$$

$z$  = Zahl der Dummies (hier : 3)

$\alpha_z$  = Effekt der letzten Dummy

$r_i$  = Regressionskoeffizient für Dummy  $a_i$

$r_z$  = Regressionskoeffizient der letzten Dummy (ist = 0)

### Vergleich mit SAS und SPSS

Das SPSS-Syntax-Programm kann nachgerechnet werden. Es ist unter dem Namen "Gleich.sps" im Ordner TESTDAT enthalten. Dort befinden sich auch die Daten unter dem Namen "Gleich.sav".

SAS und SPSS erbringen beim Verfahren der "weighted squares of means" (=SS-Typ III) unter der Bezeichnung "Parameter" folgende Koeffizienten

für "Intercept" (Konstante)	14.1
für $a_1$	15.0
$a_2$	-6.85
$a_3$	0
für $b_1$	-2.7
$b_2$	0
für $a_1b_1$	-2.8
$a_1b_2$	0
$a_2b_1$	2.3
$a_2b_2$	0
$a_3b_1$	0
$a_3b_2$	0

Wir setzen in Gleichung 1 für  $i = 1$  und  $j = 1$  ein

$$\begin{aligned}y' &= \tau + \alpha_i + \beta_j + \alpha\beta_{ij} \\ &= 14.1 + 15.0 - 2.7 - 2.8 \\ &= 23.6\end{aligned}$$

Es wird also der Zellenmittelwert  $\overline{A_1B_1}$ , so wie es sein muss, prognostiziert – obwohl die Parameter völlig andere Werte als die von Almo ausgegebenen Effekte haben.

Wir erkennen also:

Die Koeffizienten sind relativ beliebig. Sie müssen jedoch so geschätzt werden, dass sie (eingesetzt in Gleichung 1) den Zellenmittelwert reproduzieren.

Beide Berechnungsweisen, die von SPSS durch die "Parameter"-Schätzung und die von Almo durch "Effekte"-Schätzung sind legitim.

Die Parameter von SAS bzw. SPSS können inhaltlich nicht sinnvoll interpretiert werden.

1. Der Parameter des "Intercept" (Konstante) ist gleich dem Zellenmittelwert  $\overline{AB}_{32}$ , allgemein: gleich dem Mittelwert der Zelle im rechten unteren Eck in der Tabelle der Zellenmittelwerte (bzw. dem Mittelwert der Kombination der letzten Ausprägungen der nominalen Variablen).
2. Der Parameter für  $a_1$  ist gleich der Differenz des Zellenmittelwert  $\overline{AB}_{12}$  – Intercept, also  
 $29.1 - 14.1 = 15.0$

Der Parameter für  $a_2$  ist gleich  $\overline{AB}_{22}$  – Intercept, also

$$7.24 - 14.1 = 6.85$$

(siehe dazu Littell, Freund, Spector: SAS System for Linear Models, SAS Institute, 1991, S. 164)

Die Parameter aus SAS und SPSS dürfen also nicht inhaltlich wie die "klassischen Effekte" bzw. wie die Effekte aus Almo interpretiert werden.

### "Klassische" Effekte bei SPSS

Die klassischen Effekte, die Almo ausgibt, können auch in SPSS ausgegeben werden. In der Dialogbox "Kontraste" werden für die nominalen Variablen Abweichungskontraste angefordert. In der Ergebnisausgabe werden unter der Überschrift „Kontrastergebnisse (K-Matrix)“ dann die Effekte ausgegeben. Der Effekt der letzten (redundanten) Dummy wird dort weggelassen. Er muss errechnet werden als negative Summe der anderen Effekte, so dass die Summe aller Effekte gleich 0 ist. Interaktionseffekte allerdings können nicht als Abweichungskontraste ermittelt werden.

### Paarweise Vergleiche bei SPSS

Die paarweisen Vergleiche (so wie sie in Almo errechnet werden) können bei SPSS dadurch angefordert werden dass man in der Dialogbox "Optionen" die Randmittel für die nominalen Variablen anfordert und das Kästchen "Haupteffekte vergleichen" anklickt.

### Berechnung der Effekte bei ungleichen Zellenhäufigkeiten

Betrachten wir folgende Datenmatrix.

	B1		B2	
A1	23.5	23.7	28.7	
A2	8.9		5.6	8.9
A3	10.3	12.5	13.6	14.6

Sie finden das Almo-Programm zu dieser Datenmatrix unter dem Namen "Ungleich.Alm" (als Syntaxprogramm) oder "Ungleich2.Alm" (als Maskenprogramm) durch Klick auf den Knopf "alle Progs" am Oberrand des Almo-Fensters. Die Daten für ein SPSS-Programm findet man im Almo-Ordner TESTDAT unter dem Namen "Ungleich.sav".

Wenn wir mit Almo, SAS und SPSS rechnen, dann erhalten wir folgende Ergebnisse (wir betrachten der Einfachheit halber nur die Effekte von A):

	Almo weighted squares of means	Almo fitting constants	SPSS	SAS
r <sub>1</sub>	10.4917	12.9071	14.6000	14.600
r <sub>2</sub>	-7.5833	-5.3071	-6.8500	-6.850
r <sub>3</sub>	0	0	0	0
α <sub>1</sub>	10.4917	10.6271	-	-
α <sub>2</sub>	-7.5833	-7.5871	-	-
α <sub>3</sub>	-2.9083	-2.2800	-	-
τ	15.6583	15.0300	14.1000	14.100

r<sub>1</sub>, r<sub>2</sub>, r<sub>3</sub> = bei Almo: Regressionskoeffizienten der Dummies a<sub>1</sub>, a<sub>2</sub>, a<sub>3</sub>  
 bei SAS und SPSS: "Parameter"

α<sub>1</sub>, α<sub>2</sub>, α<sub>3</sub> = Haupteffekte von A<sub>1</sub>, A<sub>2</sub>, A<sub>3</sub>

τ = Konstante

Die erklärten Streuungen sind folgende:

	Almo weighted squares of means	Almo fitting constants I+II	SPSS SS Typ III	SPSS SS Typ II
SSA	479.108	499.120	479.108	499.120
SSB	9.456	10.714	9.456	10.714
SSAB	15.731	15.731	15.731	15.731

SSA = durch A erklärte Streuung

SSB = durch B erklärte Streuung

SSAB = durch die Interaktion AB erklärte Streuung

In Abschnitt P20.7.3.1 haben wir gezeigt, wie die Effekte nach dem Verfahren der "weighted squares of means" in einer sehr anschaulichen Weise auch bei ungleichen Zellenhäufigkeiten berechnet werden können.

Die Almo-Effekte beim "fitting constants" Verfahren werden gemäß (6) bzw. (7) errechnet. Sie sind echte Kleinste-Quadrate-Schätzer. Sie stimmen nicht überein mit denen, die beim Verfahren der "weighted squares of means" ermittelt werden. Die Unterschiede sind allerdings nicht sehr groß. Je ungleicher die Zellenhäufigkeiten sind, umso größer werden jedoch die Unterschiede.

Die Koeffizienten aller 4 Varianten ergeben, eingesetzt in Gleichung 1, den jeweiligen Zellenmittelwert. Die Prognosewerte und Residuen, die Almo und SPSS liefern, sind für Almo und SPSS für alle Verfahren gleich.

### P20.7.6 Berechnung der paarweisen Vergleiche

Die paarweisen Vergleiche für die Variable A und B ergeben sich einfach als Differenz der Effekte, also

$$k_1 = \alpha_1 - \alpha_2$$

$$k_2 = \alpha_1 - \alpha_3$$

$$k_3 = \alpha_2 - \alpha_3$$

$$k_4 = \beta_1 - \beta_2$$

#### Paarweise Vergleiche bei SPSS und SAS

Die paarweisen Vergleiche (so wie sie in Almo errechnet werden) können bei SPSS dadurch angefordert werden dass man in der Dialogbox "Optionen" die Randmittel für die nominalen Variablen anfordert und das Kästchen "Haupteffekte vergleichen" anklickt

Die Varianz eines Vergleichs (Kontrast)  $k_i$  wird in Almo folgendermaßen berechnet:  
Es wird eine Kontrastmatrix  $\mathbf{K}$  gebildet. Sie hat folgende Gestalt

		a1	a2	a3	b1	b2
Kontrast	k1= $\alpha_1 - \alpha_2$	1	-1	0	0	0
	k2= $\alpha_1 - \alpha_3$	1	0	-1	0	0
	k3= $\alpha_2 - \alpha_3$	0	1	-1	0	0
	k4= $\beta_1 - \beta_2$	0	0	0	1	-1

Der Effekt, der subtrahiert wird, erhält  $-1$ , der andere  $+1$  und der nicht beteiligte  $0$ . Die Spalten, die den redundanten Dummies  $a_3$  und  $b_2$ , also die 3. und 5. Spalte werden aus  $\mathbf{K}$  herausgenommen.

Anmerkung: Diese Kontrastmatrix darf nicht verwechselt werden mit der Kontrastmatrix im SPSS-Kalkül.

Weiterhin wird eine Streuungsmatrix  $\mathbf{Q}$  der Dummies von A und B, also  $a_1, a_2, a_3, b_1, b_2$  gebildet. Auch bei ihr wird die Zeile und Spalte, die den redundanten Dummies  $a_3$  und  $b_2$  entspricht, herausgenommen. Es wird die quadratische Form  $q$  gebildet.

$$q = \mathbf{K}_i \cdot \mathbf{Q}^{-1} \cdot \mathbf{K}_i'$$

$\mathbf{K}_i$  = das ist die Zeile  $i$  des Kontrastes  $\mathbf{K}_i$  aus der Matrix  $\mathbf{K}$

$\mathbf{Q}^{-1}$  = Inverse der Streuungs-Matrix (der Abweichungsquadratsummen) der nicht-redundanten Dummies  $a_1, a_2, b_1$

Die Varianz ist dann

$$s^2_{k_i} = q \cdot SS_e / dfe$$

und der t-Wert

$$t = \frac{k_i}{\sqrt{s^2_{k_i}}}$$

$s^2_{k_i}$  = Varianz des Kontrastes  $i$

$SS_e$  = Fehlerstreuung des Gesamtmodells

$dfe$  = Nenner-Freiheitsgrade des Gesamtmodells

Bei der einfaktoriellen Varianzanalyse vereinfacht sich die Berechnung der Varianz des Kontrast  $K_i$  erheblich.

$$s^2_{k_i} = \left( \frac{1}{n_u} + \frac{1}{n_v} \right) \cdot SS_e / dfe$$

$n_u, n_v$  = Häufigkeit der (kontrastierten) Ausprägungen u bzw. v der Variablen A (siehe dazu Bortz 1993, S. 246, 247)

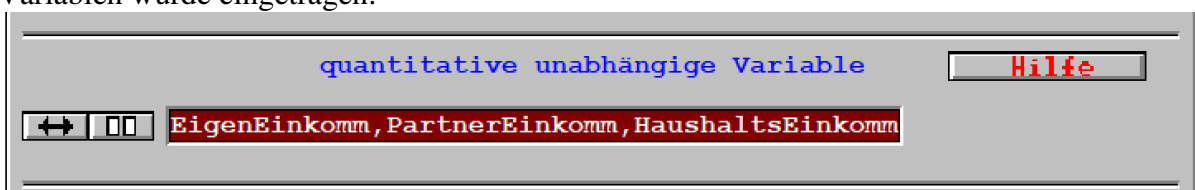
Die hier beschriebene Berechnungsmethode wird als LSD-Methode bezeichnet, dabei bedeutet LSD = least significant difference. Siehe auch unsere Ausführungen in Abschnitt P20.6.6.

### P20.7.7 Leere Zellen und lineare Abhängigkeit

In einer Regressions-, Varianz- oder Kovarianzanalyse können lineare Abhängigkeiten auftreten. Dieser Fall wird auch als "(Multi-) Kollinearität" bezeichnet. Eine lineare Abhängigkeit entsteht, wenn eine Variable mit einer anderen Variablen mit 1.0 oder nahe 1.0 korreliert oder wenn sie mit einer Linearkombination mehrere anderer Variable mit 1.0 oder nahe 1.0 korreliert. Almo erkennt eine lineare Abhängigkeit, wenn die Korrelation  $\geq 0.99996$  ist. Über eine Option kann dieser Schwellenwert verändert werden. Variable, von denen Almo erkennt, dass sie die lineare Abhängigkeit verursachen, werden eliminiert - genauer: ihre Korrelationen mit allen anderen unabhängigen Variablen und mit der abhängigen Variablen werden auf 0 gesetzt. Lineare Abhängigkeiten können bei (1) Kovariaten, (2) bei den Hauptdummies und (3) bei den Interaktionsdummies auftreten. Bei (1) und (3) treten Probleme auf.

#### P20.7.7.1 Almo eliminiert Kovariate wegen linearer Abhängigkeit

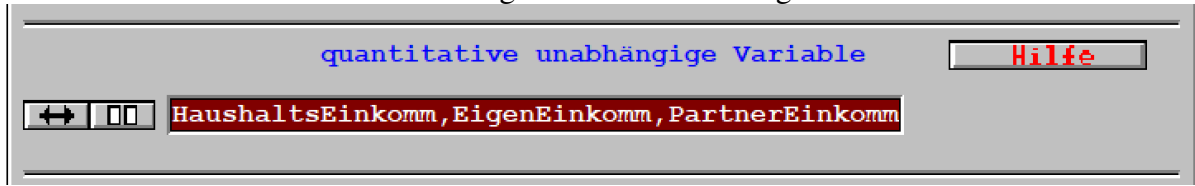
Betrachten wir ein Beispiel: In einer Regressionsanalyse oder auch Kovarianzanalyse werden die Einkommen von Ehemann und Ehefrau sowie das Haushaltseinkommen als Kovariate eingesetzt. Das Programm zu diesem Beispiel findet man durch Klick auf den Knopf "alle Progs" am Oberrand des Almo-Fensters unter dem Namen "LinAbh\_Kovar.Alm". Wir zeigen hier nur Ausschnitte aus diesem Programm. In die Eingabebox für die unabhängigen Variablen wurde eingetragen:



Das HaushaltEinkommen ist jedoch (bei 99,99.. % aller Probanden) die Summe von Eigen- und PartnerEinkommen. Das HaushaltEinkommen korreliert also mit 1.0 bzw. nahe 1.0 mit der Linearkombination der beiden anderen Einkommen. Almo eliminiert bei dieser Eingabe das HaushaltEinkommen. Die Ergebnisse, die Almo bei dieser Analyse liefert sind korrekt. Sie sind identisch mit einer Analyse, bei der der Benutzer nur Eigen- und PartnerEinkommen als Kovariate eingesetzt hat. Als Regressionskoeffizienten gibt Almo aus:

	Regress. koeff.
-----	
V5 EigenEinkommen	0.0104
V6 PartnerEinkommen	-0.2432
V21 HaushaltEinkomm	eliminiert

Almo überprüft die Korrelationsmatrix (allgemein: die Streuungsmatrix) der unabhängigen Variablen von links nach rechts. Sobald es auf eine lineare Abhängigkeit trifft, wird die betreffende Variable eliminiert. Bei folgender Variablen-Eingabe .....



.....würde von Almo dann das PartnerEinkommen eliminiert werden und nicht das Haushalts-Einkommen. Als Regressionskoeffizienten gibt Almo aus:

	Regress. koeff.
V21 HaushaltsEinkomm	-0.2432
V5 EigenEinkommen	0.2537
V6 PartnerEinkommen	eliminiert

Der Benutzer sollte also überlegen, wo eine Linearkombination in seinen Variable versteckt sein könnte. Und dann die Variablen entsprechen hintereinander reihen. Noch besser ist es natürlich, er schließt die verursachende Variable - in unserem Beispiel: das Haushalts-einkommen - aus der Analyse aus.

Wir haben für dieses Beispiel eine Kovarianzanalyse gerechnet, bei der wir zusätzlich die nominalen Variablen Geschlecht und Beruf hinzugefügt haben. Für die 1. Reihenfolge der Kovariaten (mit Haushaltseinkommen als 3. Kovariate) kann der Benutzer das Programm "LinAbh\_Kovar.Alm" nachrechnen. Um die 2. Reihenfolge (mit Haushaltseinkommen als 1. Kovariate) zu rechnen muss der Benutzer nur in der Eingabebox "Analyse-Variable: Unabhängige Variable" die Reihenfolge ändern in: HaushaltsEinkomm, EigenEinkomm, PartnerEinkomm

Unter dem Knopf "alle Progs" findet man auch das sehr elaborierte Simulationsprogramm "LIN\_ABH.ALM" mit dem lineare Abhängigkeiten generiert werden können.

Almo liefert für die beiden Analysen folgende Ergebnisse (stark gekürzt):

	Reihenfolge der Kovariaten in					
	1. Analyse			2. Analyse		
	EigenEinkomm	PartnerEinkomm	HaushaltsEinkomm	HaushaltsEinkomm	EigenEinkomm	PartnerEinkomm
	-----			-----		
Streuungsquelle	Streuung	Regress. koeff.	Korrel koeff.	Streuung	Regress. koeff.	Korrel koeff.
-----	-----	-----	-----	-----	-----	-----
Gesamtstreuung	376.6885			376.6885		
Fehlerstreuung	323.1622			323.1622		
gesamte erklärte Streuung	53.5263		0.3770	53.5263		0.3770
V5 EigenEinkomm	0.0196	0.0104	0.0078	6.6642	0.2537	0.1421
V6 PartnerEinkomm	16.0181	-0.2432	-0.2173	0.0000	-	-
V21 HaushaltsEinkomm	0.0000	-	-	16.0181	-0.2432	-0.2173
V1 Geschlecht	6.7677		0.1432	6.7677		0.1432
V3 Beruf	19.5715		0.2390	19.5715		0.2390
V1*V3	10.7408		0.1794	10.7408		0.1794

Prognosewerte fuer  
die ersten 10 Datensätze

	tatsächlich	Prognose	tatsächlich	Prognose
1	2.0000	2.60527	2.0000	2.60527
2	1.0000	2.85895	1.0000	2.85895
3	3.0000	3.09175	3.0000	3.09175
4	4.0000	3.31412	4.0000	3.31412
5	1.0000	3.02484	1.0000	3.02484
6	1.0000	2.53837	1.0000	2.53837
7	3.0000	3.24721	3.0000	3.24721
8	4.0000	2.78160	4.0000	2.78160
9	2.0000	5.27940	2.0000	5.27940
10	2.0000	4.32732	2.0000	4.32732
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.

Alle Koeffizienten aus den beiden Analysen - ausser denen für die 3 Kovariaten - stimmen voll überein. Auch die Prognosewerte der Probanden stimmen überein. Die Variable "EigenEinkomm" korreliert in der 1. Analyse mit der Zielvariablen mit 0.0078 und in der 2. Analyse mit 0.1421. Um das tatsächliche Gewicht zu entdecken, mit dem eine Variable - im Vergleich zu den anderen - die Zielvariable determiniert, ist es entscheidend wichtig, den eigentlichen Verursacher der linearen Abhängigkeit zu identifizieren. Das kann nur der Forscher leisten.

Das Programm kann auch mit SPSS nachgerechnet werden. Die Ergebnisse sind dieselben. Man beachte allerdings die unterschiedliche Terminologie von SPSS und Almo. Verwenden Sie dabei die Datei "LinAbh\_Kovar.sav" im Almo-Ordner "Testdat". Der SPSS-Syntax-File ist im gleichen Ordner unter dem Namen "LinAbh\_Kovar.sps" enthalten. Die durch die 3 Kovariaten erklärte Streuung wird in SPSS nicht ausgegeben, dafür jedoch die *quadrierte* Korrelation Eta. Almo gibt die *nicht-quadrierte* Korrelation Eta aus

### ***P20.7.7.2 Almo eliminiert Haupt-Dummies wegen linearer Abhängigkeit***

Werden Dummies einer nominalen Variablen eliminiert, dann bedeutet dies, (a) dass eine oder mehrere Ausprägungen der nominalen Variablen nicht besetzt sind bzw. im Vergleich zu den anderen Ausprägungen nur sehr minimal besetzt sind oder (b) dass zwei Ausprägungen zwar verschieden benannt sind, aber dasselbe beinhalten. In diesem Falle fassen Sie diese nicht (oder schwach) besetzte oder inhaltlich gleichen Ausprägung mit einer anderen Ausprägung zusammen oder setzen Sie auf "KeinWert".

Beispiel:

Die Variable "Beruf" besitzt 6 Ausprägungen. Die Ausprägung 3 ist nicht (oder sehr schwach) besetzte. Dann schreiben Sie folgende Umkodierung

```
Beruf (2,3=2; 4=3; 5=4; 6=5)      # 2 und 3 werden zusammengefasst      #
                                   # und die nachfolgenden Ausprägungen #
                                   # werden um 1 herunter gesetzt      #
```

oder

```
Beruf (3=KeinWert; 4=3; 5=4; 6=5) # 3 wird KeinWert gesetzt            #
                                   # und die nachfolgenden Ausprägungen #
                                   # werden um 1 herunter gesetzt      #
```

### ***P20.7.7.3 Almo eliminiert Interaktions-Dummies wegen linearer Abhängigkeit.***

#### ***Das Problem der "leeren Zellen"***

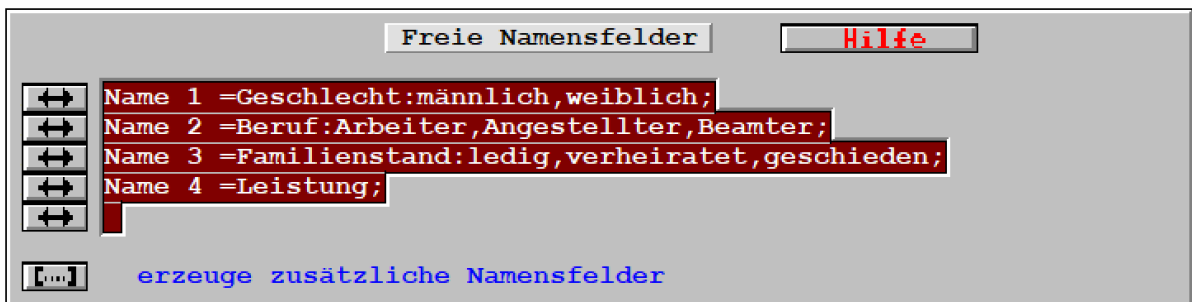
Die eigentliche Ursache der linearen Abhängigkeit bei Varianz-, Kovarianzanalysen ist fast immer, dass die Datentabelle "leere Zellen" enthält. Es sei ausdrücklich darauf hingewiesen,

dass unsere folgenden Ausführungen in entsprechend gleicher Weise für die Kovarianzanalyse gelten, also für den Fall, dass die unabhängigen nominalen Variablen an Kovariate angepasst sind.

Wir rechnen eine Varianzanalyse mit der Zielvariable **Leistung** (in irgend einem Test) und den 3 unabhängigen nominalen Variablen

- Geschlecht:** (1)männlich, (2)weiblich
- Beruf:** (1)Arbeiter, (2)Angestellter, (3)Beamter
- Familienstand:** (1)ledig, (2)verheiratet, (3) geschieden

Die Daten sind in Almo enthalten in ".\Testdat\Leerzel2.fre". Es sind "erfundene" Daten. Das Programm dazu findet man unter dem Namen "Leerzel2a.Alm" durch Klick auf den Knopf "alle Progs" am Oberrand des Almo-Fensters. Die Box für die Variablennamen ist in folgender Weise ausgefüllt:



Die 53 Probanden verteilen sich in folgender Weise auf die Merkmalskombination der 3 nominalen Variablen

			Leistung		
			Mittelwert	Häufigkeit	
Geschlecht	Beruf	Familienstand			
männlich	Arbeiter	ledig	5.6667	3	
		verheira	1.6000	5	
		geschied	0	0	<--Leerzelle
	Angestel	ledig	4.2500	8	
		verheira	4.6250	8	
		geschied	5.0000	2	
	Beamter	ledig	0	0	<--Leerzelle
		verheira	0	0	<--Leerzelle
		geschied	0	0	<--Leerzelle
weiblich	Arbeiter	ledig	5.0000	2	
		verheira	5.0000	3	
		geschied	3.3333	3	
	Angestel	ledig	3.6000	5	
		verheira	4.5000	4	
		geschied	3.5000	2	
	Beamter	ledig	7.0000	3	
		verheira	4.0000	4	
		geschied	1.0000	1	
Gesamt			4.1887	53	

Almo bildet folgende 17 Dummies. Die redundanten Dummies A2, B3, C3 und ihre Interaktionen werden für den Kalkül nicht gebraucht. Almo überprüft, ob zwischen diesen 17 Dummies lineare Abhängigkeiten bestehen. Das geschieht dadurch dass die Choleskymatrix der Abweichungs-Quadratsummen-Matrix berechnet wird. Die Diagonalglieder werden von Almo ausgegeben

Diagonalglieder der Choleskymatrix  
zur Ermittlung und zum Ausschluss  
linearer Abhängigkeiten

```
-----
A1          52.981132
B1          21.538462
B2          23.923810
C1          24.918790
C2          25.500312
A1 B1       9.720683
A1 B2       0.000000 <---
A1 C1      22.080092
A1 C2      22.659283
B1 C1       9.141764
B1 C2      10.273456
B2 C1      10.011070
B2 C2       9.984757
A1 B1 C1    1.935484
A1 B1 C2    0.000000 <---
A1 B2 C1    0.000000 <---
A1 B2 C2    0.000000 <---
```

Almo eliminiert eine Variable  $i$ , wenn ihr Diagonalglied aus der Cholesky-Matrix kleiner ist als  $0.0001 * SS(i)$ .  $SS(i)$  ist das Diagonalglied  $ii$  der Variablen  $i$  aus der Streuungsmatrix. Würde anstelle der Abweichungs-Quadratsummen-Matrix mit der Korrelationsmatrix gerechnet, dann wäre  $SS(i) = 1.0$ .

In unserem Beispiel bringt Almo eine Fehlermeldung

```
Folgende Variable werden eliminiert
A1 B2
A1 B1 C2
A1 B2 C1
A1 B2 C2
```

**\*\*\*\*\* FEHLER**

```
Muessen Dummies (insbesondere Interaktions-Dummies)
eliminiert werden, dann koennen falsche Ergebnisse entstehen
Rechnen Sie eine Analyse ohne Interaktionen
oder mit nur einigen ausgewählten Interaktionen
Almo bricht den Kalkuel nicht ab
```

Almo hat festgestellt, dass zwischen den 17 Dummies lineare Abhängigkeiten bestehen, für welche die oben angegebenen 4 Interaktions-Dummies verantwortlich sind. Diese müssen eliminiert werden, damit überhaupt gerechnet werden kann.

Es sind die 4 leeren Zellen, welche die linearen Abhängigkeiten in einer schwer zu durchschauenden Weise verursacht haben. Betrachten wir die eliminierte Interaktions-Dummy A1B2. Was wir sagen können ist folgendes:

A1B2 korreliert mit

(1) einer der vor ihr stehenden Interaktionsdummies also A1 A2 . . . . A1B1 mit (nahe) 1.0.  
oder

(2) mit einer Linearkombination der vor ihr stehenden Interaktionsdummies mit (nahe) 1.0

Aber welcher der beiden Fälle zutrifft oder gar wie die Linearkombination gestaltet ist, kann Almo nicht erkennen.

Allgemein gilt:

Hängen mehrere Variable in irgend einer linearen Weise zusammen, dann wird von Almo die Variable eliminiert, die in der Reihenfolge der Variablen die letzte ist.

Das ist durch den Kalkül der Cholesky-Matrix bedingt, der in Almo verwendet wird. Das ist aber genau so beim Gauss-Jordan-Verfahren, das in einigen anderen Statistikprogrammen verwendet wird.

Für Leser, die sich für diese Thematik besonders interessieren, empfehlen wir, mit dem Simulationsprogramm "Lin\_Abh.Alm" zu experimentieren. Dort wird mit der Gleichung

$$x_2 = \text{Wurzel}(1.0 - H_2 * H_2) * x_2 + H_2 * x_1;$$

eine lineare Abhängigkeit von  $x_1$  und  $x_2$  erzeugt. Anstelle von  $x_1$  könnte nun eine Linearkombination von z.B.  $(x_3 + x_4) / 2$  eingesetzt werden. Das Simulationsprogramm findet man durch Klick auf den Knopf "alle Progs" am Oberrand des Almo-Fensters.

Nach der oben abgebildeten Fehlermeldung (dass Variable eliminiert werden) rechnet Almo weiter. *Der Benutzer sollte aber den Ergebnissen misstrauen*, insbesondere den erklärten Streuungen und Korrelationen je unabhängige Variable.

Wir haben die Analyse ein 2. Mal gerechnet, dabei wurde die Variable Beruf (in der Umkodierungsbox) so umkodiert:

**Beruf(1=2; 2=1)**

Das bedeutet, dass Arbeiter in der 2. Analyse mit 2 kodiert sind und Angestellte mit 1. Werden den Ausprägungen einer unabhängigen nominalen Variablen andere Ziffern zugewiesen, dann dürfen dadurch die erklärte Streuung und die Korrelation nicht verändert werden. Das ist aber offensichtlich nicht der Fall, wenn Interaktionsdummies eliminiert werden.

Streuungsquelle	1. Analyse		2. Analyse	
	Streuung	Korrel Koeff.	Streuung	Korrel Koeff.
Gesamtstreuung	360.1132		360.1132	
Fehlerstreuung	274.6083		274.6083	
alle unabh. Var. zusammen	85.5049	0.4873	85.5049	0.4873
V1 Geschlecht	3.0445	0.1047	9.0348	0.1785
V2 Beruf	9.7101	0.1848	11.3189	0.1990
V3 Familienstand	11.6820	0.2020	27.5580	0.3020
V1*V2	13.6796	0.2178	13.6796	0.2178
V1*V3	1.4126	0.0715	13.3787	0.2155
V2*V3	28.0110	0.3042	42.4551	0.3659
V1*V2*V3	6.0694	0.1471	6.0694	0.1471

**Prognosewerte**

Datensatz	tatsaechlicher Wert Leistung	prognostizierter Wert Leistung	Residuen (Differenz) Leistung
1	0.00000	1.60000	-1.6000
2	6.00000	1.60000	4.40000
3	1.00000	1.60000	-0.6000
4	2.00000	5.66667	-3.6667
5	7.00000	4.62500	2.37500
6	9.00000	4.62500	4.37500
7	3.00000	4.25000	-1.2500
8	1.00000	4.62500	-3.6250
9	7.00000	5.66667	1.33333
10	0.00000	1.60000	-1.6000
.	.	.	.
.	.	.	.
.	.	.	.

Beim Vergleich der Ergebnisse stellen wir fest:

- die gesamte Streuung, die Fehlerstreuung und die gesamte erklärte Streuung sind in beiden Ergebnissen gleich
- die Prognosewerte sind gleich. Für einen Probanden wird korrekt der Zellenmittelwert prognostiziert derjenigen Zelle, in der er sich befindet. Der 1. Proband aus der 1. Analyse befindet sich in der 2. Zelle. Deren Zellenmittelwert ist 1,6. Siehe Tabelle oben. In der 2. Analyse befindet er sich (bedingt durch die Umkodierung) in der 5. Zelle. Auch deren Zellenmittelwert ist 1,6.
- die erklärten Streuungen der Variablen und ihrer Korrelationen (mit der grundsätzlichen Ausnahme der höchsten Interaktion) sind zwischen den 2 Analysen verschieden.

Wir haben also ein widersprüchliches Ergebnis. Beide Analysen reproduzieren die Prognosewerte für die Probanden gleich und korrekt. Die erklärten Streuungen und Korrelationen der unabhängigen Variablen sind jedoch verschieden.

In unserem Beispiel wurden die Codeziffern der beiden Ausprägungen umgedreht. Dadurch werden die Hauptdummies und die aus ihnen gebildeten Interaktionsdummies in ihrer Reihenfolge in der Streuungsmatrix auch verdreht. Wir wollen diesen Sachverhalt detaillierter betrachten. Die Streuungsmatrix aus der 1. Analyse ist folgende. Wir zeigen nur einen Ausschnitt:

**Matrix 1: Abweichungs-Quadratsummen-Matrix aus 1. Analyse bei der die Variable Beruf wie im Programm kodiert ist**

	A1	B1	B2	C1	C2	A1B1	A1B2	
A1	52.9811	8.1509	15.3962	5.2453	6.3019	8.1509	21.2830	...
B1	8.1509	22.7925	4.8302	-1.9623	-0.4151	-9.2075	-10.2642	...
B2	15.3962	4.8302	28.6792	1.8491	-1.3396	-11.1698	-6.9434	...
C1	5.2453	-1.9623	1.8491	25.8113	4.0755	4.0377	1.3208	...
C2	6.3019	-0.4151	-1.3396	4.0755	27.1698	5.5849	2.4717	...
A1 B1	8.1509	-9.2075	-11.1698	4.0377	5.5849	22.7925	5.7358	...
A1 B2	21.2830	-10.2642	-6.9434	1.3208	2.4717	5.7358	32.7547	...
.	.	.	.	.	.	.	.	...

Die Zeile/Spalte der Interaktionsdummy A1B2, von der wir wissen, dass sie im Verlauf des Kalküls eliminiert wird, ist farblich hervorgehoben. Die Zeile/Spalte enthält die Streuungen

und Co-Streuungen von A1B2 mit den anderen Variablen. Wir rechnen nun die Inverse von Matrix 1. Sie wird gebraucht um die Effekte, erklärten Streuungen usw. zu ermitteln. Wir zeigen nur einen Ausschnitt.

**Matrix 2: Inverse mit Identifizierung von Kollinearitäten**  
gerechnet nach Clarke (1982) und Ridout/Cobby (1988)

	A1	B1	B2	C1	C2	A1B1	A1B2	
A1	0.0472	-0.0137	-0.0352	-0.0137	-0.0123	-0.0313	0.0000	...
B1	-0.0137	0.2141	0.0136	0.0492	-0.0668	0.1762	0.0000	...
B2	-0.0352	0.0136	0.0809	0.0074	0.0148	0.0463	0.0000	...
C1	-0.0137	0.0492	0.0074	0.1008	-0.0198	0.0359	0.0000	...
C2	-0.0123	-0.0668	0.0148	-0.0198	0.0921	-0.0718	0.0000	...
A1 B1	-0.0313	0.1762	0.0463	0.0359	-0.0718	0.2271	0.0000	...
A1 B2	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	...
.	.	.	.	.	.	.	.	...

Wir rechnen das Gauss-Jordan-Verfahren mit dem Algorithmus AS 178 (in der C-Version) nach Clarke/Ridout/Cobby von der 1. Zeile/Spalte nach rechts bis das Verfahren eine lineare Abhängigkeit entdeckt. Das ist erwartungsgemäß bei Zeile/Spalte A1B2 der Fall. Diese Zeile/Spalte wird auf 0.0 gesetzt. Offensichtlich korreliert die Zeile/Spalte A1B2 mit einer Linearkombination von Variablen, die in der Menge A1, B1, ... bis A1B1 enthalten sind. Vermutlich werden dabei die Zeilen/Spalten C1 und C2 keine Rolle spielen. Der Algorithmus AS 178 ist als C-Prozedur im Almo-Ordner "Algorithmen\_in\_C" unter dem Namen "a\_gsweep" enthalten. Siehe dort den File "Info2".

Nun betrachten wir die 2. Analyse, bei der die Variable Beruf umkodiert wurde.

**Matrix 3: Abweichungs-Quadratsummen-Matrix** aus 2. Analyse  
bei der die Variable Beruf umkodiert wurde

	A1	B1	B2	C1	C2	A1B1	A1B2	
A1	52.9811	15.3962	8.1509	5.2453	6.3019	21.2830	8.1509	...
B1	15.3962	28.6792	4.8302	1.8491	-1.3396	-6.9434	-11.1698	...
B2	8.1509	4.8302	22.7925	-1.9623	-0.4151	-10.2642	-9.2075	...
C1	5.2453	1.8491	-1.9623	25.8113	4.0755	1.3208	4.0377	...
C2	6.3019	-1.3396	-0.4151	4.0755	27.1698	2.4717	5.5849	...
A1 B1	21.2830	-6.9434	-10.2642	1.3208	2.4717	32.7547	5.7358	...
A1 B2	8.1509	-11.1698	-9.2075	4.0377	5.5849	5.7358	22.7925	...
.	.	.	.	.	.	.	.	...

Man beachte, dass jetzt unter der Interaktionsdummy A1B2 die Zahlenspalte beginnend mit 8.1509 steht. Bei der 1. Analyse war dies 21.2830. Diese Zahlenspalte steht jetzt unter A1B1. Matrix 3 wird invertiert.

**Matrix 4: Inverse mit Identifizierung von Kollinearitäten**

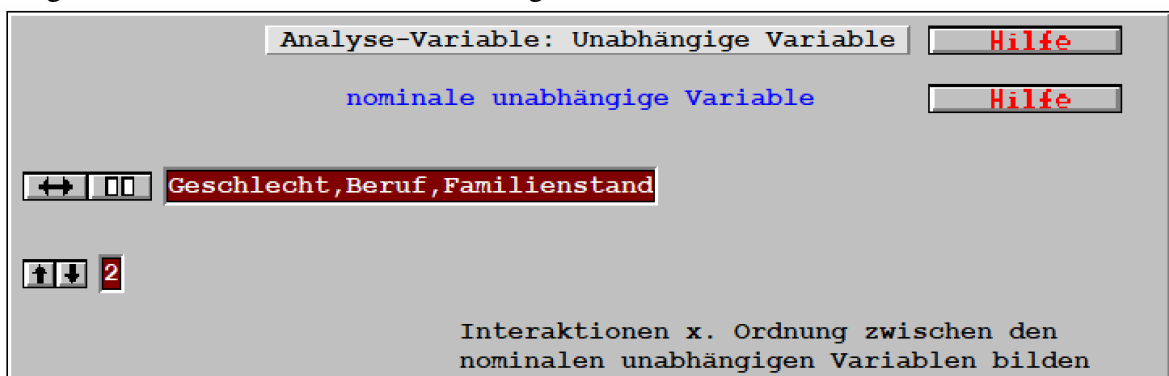
	A1	B1	B2	C1	C2	A1B1	A1B2	
A1	0.2118	-0.1847	-0.0333	-0.2007	-0.0840	-0.1958	0.0000	....
B1	-0.1847	0.2154	0.0182	0.2023	0.0866	0.1808	0.0000	....
B2	-0.0333	0.0182	0.0889	-0.0025	0.0049	0.0509	0.0000	....
C1	-0.2007	0.2023	-0.0025	0.3708	0.0958	0.1912	0.0000	....
C2	-0.0840	0.0866	0.0049	0.0958	0.0921	0.0718	0.0000	....
A1 B1	-0.1958	0.1808	0.0509	0.1912	0.0718	0.2271	0.0000	....
A1 B2	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	....
.	.	.	.	.	.	.	.	....

Der Unterschied zwischen der 1. und der 2. Analyse wird nun offenkundig. Bei der 1. Analyse wird die Zahlenreihe beginnend mit 21.283 eliminiert, bei der zweiten die Zahlenreihe

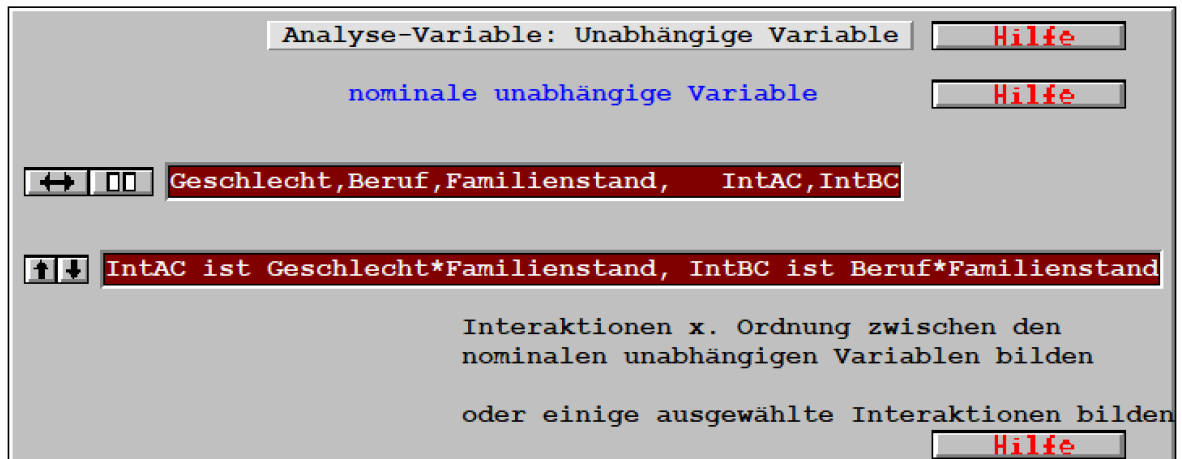
beginnend mit 8.1509. Die Inverse wird dadurch insgesamt verändert. Die Folge davon ist, dass andere Regressionskoeffizienten und Effekte für die Interaktionsdummies und andere erklärte Streuungen und Korrelation errechnet werden - allerdings in einer solchen Weise, dass die Probanden mit den gleichen Prognosewerten wie in der 1. Analyse reproduziert werden.

Wir würden dem Benutzer empfehlen, sich nicht mit dem Ergebnis abzufinden, das Almo im Fall einer linearen Abhängigkeit liefert. Für den Benutzer gibt es 5 Möglichkeiten zu reagieren:

1. Lineare Abhängigkeiten bei den Interaktionsvariablen entstehen fast immer durch leere Zellen. Diese sind sehr häufig dadurch zu verhindern, dass man Variable auf weniger Ausprägungen zusammenfasst. Dadurch entstehen dann auch weniger Ausprägungskombinationen (Zellen), auf die sich die Probanden verteilen,
2. Er rechnet ein Haupteffekte-Modell, verzichtet also auf Interaktionen.  
Für den Fall, dass Interaktionen inhaltlich nicht oder nur schwer interpretierbar sind, ist das die angemessene Lösung des Problems. Der Autor dieses Textes kann sich nicht erinnern, je eine Studie gelesen zu haben, in der eine 3-er Interaktion inhaltlich interpretiert werden konnte. Auch 2-er Interaktionen sind selten interpretierbar. Werden Interaktionen in eine Analyse mit eingeschlossen, dann wird dadurch die Fehlerstreuung reduziert, wodurch dann die Korrelationskoeffizienten für die Haupteffekte erhöht werden !!
3. Er rechnet mit Interaktionen, gibt aber als Interaktionsordnung nur die an, für die noch alle Interaktions-Dummies vorhanden sind.  
Würde in unserem Beispiel die 7. Zelle mit auch nur einem Probanden besetzt sein, dann müsste Almo die Interaktions-Dummy **A1B2** nicht eliminieren. Der Benutzer füge in der Datei ".\Testdat\Leerzel2.fre" noch einen Datensatz mit den Werten 1 3 1 5 an. In diesem Falle könnte als Interaktionsordnung "2" eingegeben werden. In die Eingabebox im Programm "Leerzel2a.Alm" würde dann geschrieben werden

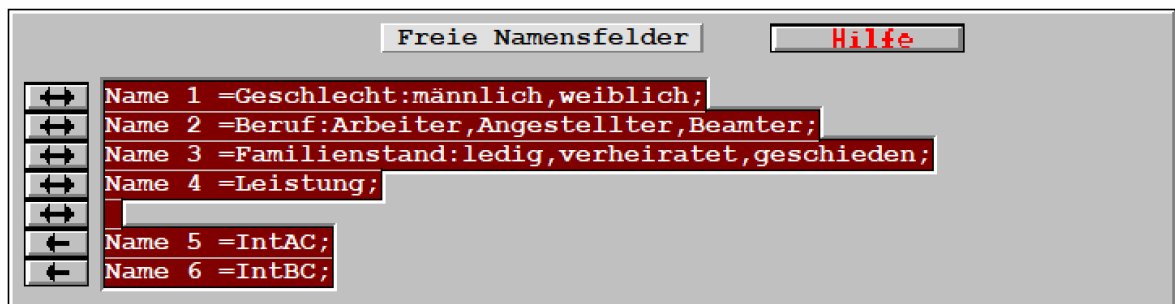


4. Er rechnet mit *selbst gebildeten Interaktionsvariablen*. Dabei werden nur die vom Benutzer *selbst gebildeten Interaktionsvariable* in die Analyse aufgenommen, von denen alle Dummies vorhanden sind - anders formuliert: ...von denen keine Dummies durch Almo eliminiert werden. In unserem Beispiel wären das die Interaktionen AC und BC. Die Interaktion AB entfällt weil die Interaktions-Dummy **A1B2** von den 2 vorhandenen von Almo eliminiert wird.  
In die Eingabebox wird eingesetzt:



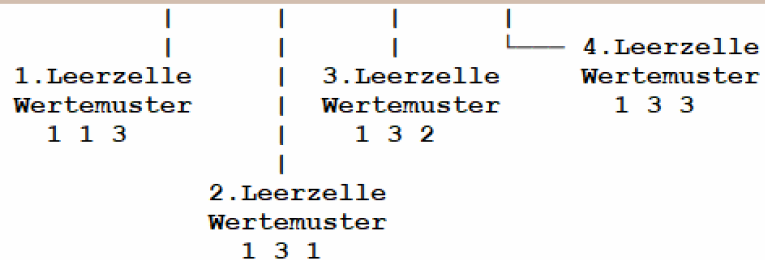
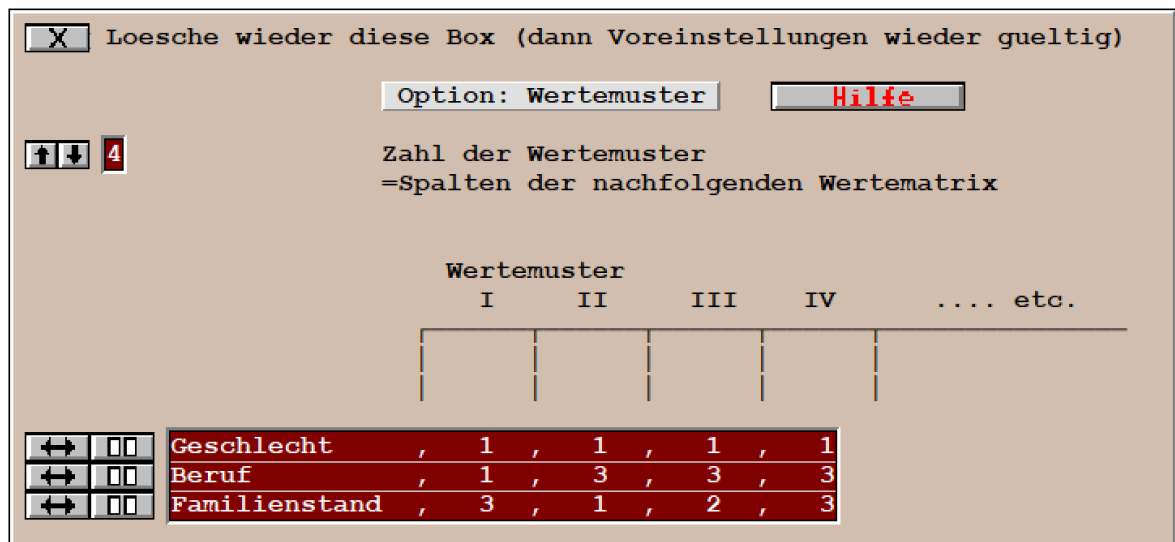
Das Programm findet man unter dem Namen "Leerzel2b.Alm" nach Klick auf den Knopf "alle Progs" am Oberrand des Almo-Fensters. Dies ist unseres Erachtens, die beste Lösung des Problems der leeren Zellen bzw. der linearen Abhängigkeiten zwischen den Interaktionsdummies.

Die Namen der Interaktionsvariable **IntAC** und **IntBC** müssen noch in der Eingabebox für die Variablennamen mit einer freien Variablennummer versehen werden.



Wie die *selbst gebildeten Interaktionsvariablen* zu bilden sind, ist ausführlich beschrieben in Abschnitt P20.14. im Almo-Dokument 13a "Allgemeines Lineares Modell II". Das Programm kann auch in dieser speziellen Weise mit SPSS gerechnet werden. Der Syntax-File ist im Almo-Ordner TESTDAT unter dem Namen Leerzell2b.sps enthalten.

- Er setzt in die leeren Zellen Ersatzwerte ein. Im einfachsten Falle kann das der Gesamt-Mittelwert sein. Aufwendiger sind dann elaborierte Daten-Imputationsverfahren. Siehe dazu das Almo-Dokument Nr. 12 "Daten-Imputation". Naheliegend ist folgende Vorgehensweise: Man rechnet mit dem in Punkt 2 dargestellten Verfahren eine Analyse, bei der die Optionsbox "Wertemuster" geöffnet wird und für die leeren Zellen in folgender Weise ausgefüllt wird:



Für die 1. leere Zelle errechnet Also einen Prognosewert von 3.03568.

Der künstliche Proband für .....

die 1. leere Zelle besitzt also folgenden Datensatz	1 1 3 3.03568
für die 2. leere Zelle	1 3 1 5.10998
für die 3. leere Zelle	1 3 2 4.2986
für die 4. leere Zelle	1 3 3 3.85712

Die 4 Datensätze müssen am Ende der Datei angefügt werden. Der Benutzer sollte diese vergrößerte Datei unter einem anderen Namen abspeichern.

Diese Methode, das Problem leerer Zellen bzw. linearer Abhängigkeiten zu lösen, ist sicherlich akzeptabel, vielleicht sogar vorzuziehen, wenn die Zahl leerer Zellen, die durch einen künstlichen Probanden gefüllt werden müssen, im Verhältnis zu Gesamtzahl der Probanden, sehr gering ist.

Wie unterscheiden sich nun die Ergebnisse, die diese 5 Methoden erbringen. Wir zeigen nur einen Ausschnitt

	1. Methode		2. Methode		4. Methode		5. Methode	
	Dummy eliminiert		Haupteffekte		Interkt. var.		Daten-Imputation	
Streuungsquelle	Streuung	Korrel Koeff.	Streuung	Korrel Koeff.	Streuung	Korrel Koeff.	Streuung	Korrel Koeff.
-----	-----	-----	-----	-----	-----	-----	-----	-----
Gesamtstreuung	360.1132		360.1132		360.1132		362.4098	
Fehlerstreuung	274.6083		342.5838		288.2882		274.6083	
unabh. Variab. zusammen	85.5049	0.4873	17.5294	0.2206	71.8250	0.4466	87.8015	0.4922
V1 Geschlecht	3.0445	0.1047	0.4185	0.0349	0.4674	0.0402	0.0289	0.0103
V2 Beruf	9.7101	0.1848	3.6043	0.1020	0.2493	0.0294	0.8136	0.0543
V3 Fam.stand	11.6820	0.2020	11.6232	0.1811	30.6912	0.3102	19.2498	0.2559

V1*V2	13.6796	0.2178	-	-	6.5910	0.1531
V1*V3	1.4126	0.0715	11.8112	0.1984	7.6267	0.1644
V2*V3	28.0110	0.3042	45.4011	0.3689	25.0103	0.2889
V1*V2*V3	6.0694	0.1471	-	-	12.7889	0.2109

Bei allen fünf Methoden hat keine der Variablen einen signifikanten Einfluss auf die abhängige Variable. In unseren "erfundenen" Testdaten determinieren die drei nominalen Variablen nicht die Zielvariable "Leistung".

### P20.7.8 Weitere Eigenschaften von Programm 20

1. Die Fehlerstreuung wird als Differenz der Gesamtstreuung der abhängigen Variablen und der erklärten Streuung bestimmt, die durch alle in das Modell eingeführten unabhängigen Variablen erklärt wird.

Das bedeutet, dass bei 2 gleichen Analysen, bei der im einen Fall die Interaktionen zwischen den nominalen Variablen als unabhängige Variable eingeführt wurden, im anderen Fall jedoch nicht, die Fehlerstreuung verschieden ist. Befinden sich unabhängige nominale Variable in der Analyse und werden alle Interaktionen, die sie bilden können, miteinbezogen, dann ist die Fehlerstreuung gleich der Summe der Zellenstreuungen (engl. "within-group-error"). Bei Versuchsplänen mit "geschachtelter Variabler" hat der Benutzer die Möglichkeit, die Fehlerstreuung anders zu definieren. Siehe Abschnitt P20.17.2 in Teil II.

2. Analysen mit ungleichen Zellenhäufigkeiten sind möglich. ALMO errechnet dann - je nach Wahl des Benutzers eines der folgenden Verfahren: weighted squares of means (empfohlen), sequentielles Verfahren, fitting constants I oder II. Außerdem kann der Benutzer die Variablen- bzw. Gruppenhierarchie (durch Schrägstriche) selbst bestimmen. Zusammen mit der "Partial"-Anweisung wird dem Benutzer hier die Möglichkeit eröffnet, ein Verfahren selbst zu programmieren. Für alle Verfahren außer fitting\_constants\_II errechnet ALMO das volle Repertoire an Koeffizienten, also Effekte, Regressionskoeffizienten, erklärte Streuungen etc.
3. Analysen mit leeren Zellen, d.h. mit unbesetzten Merkmalskombinationen der unabhängigen nominalen Variablen, sind möglich. Siehe oben P20.7.7.3.
4. Analysen mit linearen Abhängigkeiten zwischen den Variablen sind möglich. Variable, die diese linearen Abhängigkeiten verursachen, werden eliminiert.
5. Unvollständige experimentelle Pläne können analysiert werden.
6. "Geschachtelte" Versuchspläne können analysiert werden.
7. Es können Analysen durchgeführt werden mit nur einer Untersuchungseinheit je Zelle.
8. Analysen mit Messwiederholungen können durchgeführt werden.
9. Analysen mit **vielen** unabhängigen nominalen Variablen (mit vielen Ausprägungen) - ohne Interaktionen - sind möglich.

## P20.8 Die Eingabe in Programm 20

### Unabhängige und abhängige Variable

Wir haben 6 Typen von Variablen

unabhängige nominale Variable (=Faktoren)  
 unabhängige ordinale Variable  
 unabhängige quantitative Variable (=Kovariate)

abhängige nominale Variable  
 abhängige ordinale Variable  
 abhängige quantitative Variable

1. **Regel:** Diese 6 Variablentypen können beliebig kombiniert werden. Selbstverständlich muss mindestens **eine** unabhängige und mindestens **eine** abhängige Variable vorhanden sein.
2. **Regel:** Ist die abhängige Variable nominal, dann darf nur **eine** abhängige nominale Variable (mit allerdings beliebig vielen Ausprägungen) vorhanden sein und es dürfen auch keine abhängigen Variablen der beiden anderen Typen vorhanden sein. Neben der einen abhängigen nominalen Variablen dürfen also keine weiteren abhängigen Variablen der anderen Typen angegeben werden.

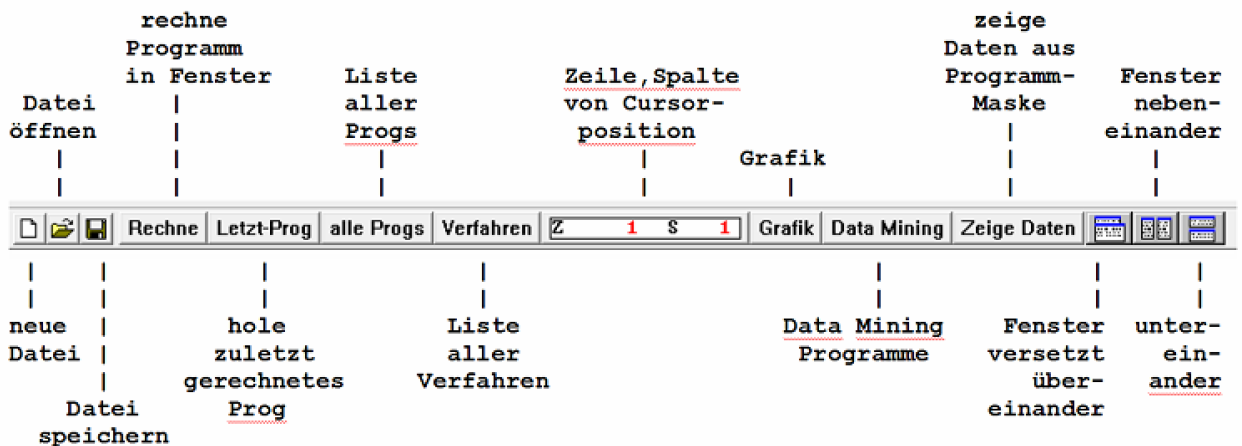
Wir werden nachfolgend zwei Maskenprogramme darstellen und erläutern

### P20.8.0 Eingabe in Maskenprogramm Prog20mx

#### *Wie sie die Maskenprogramme in Almo finden!*

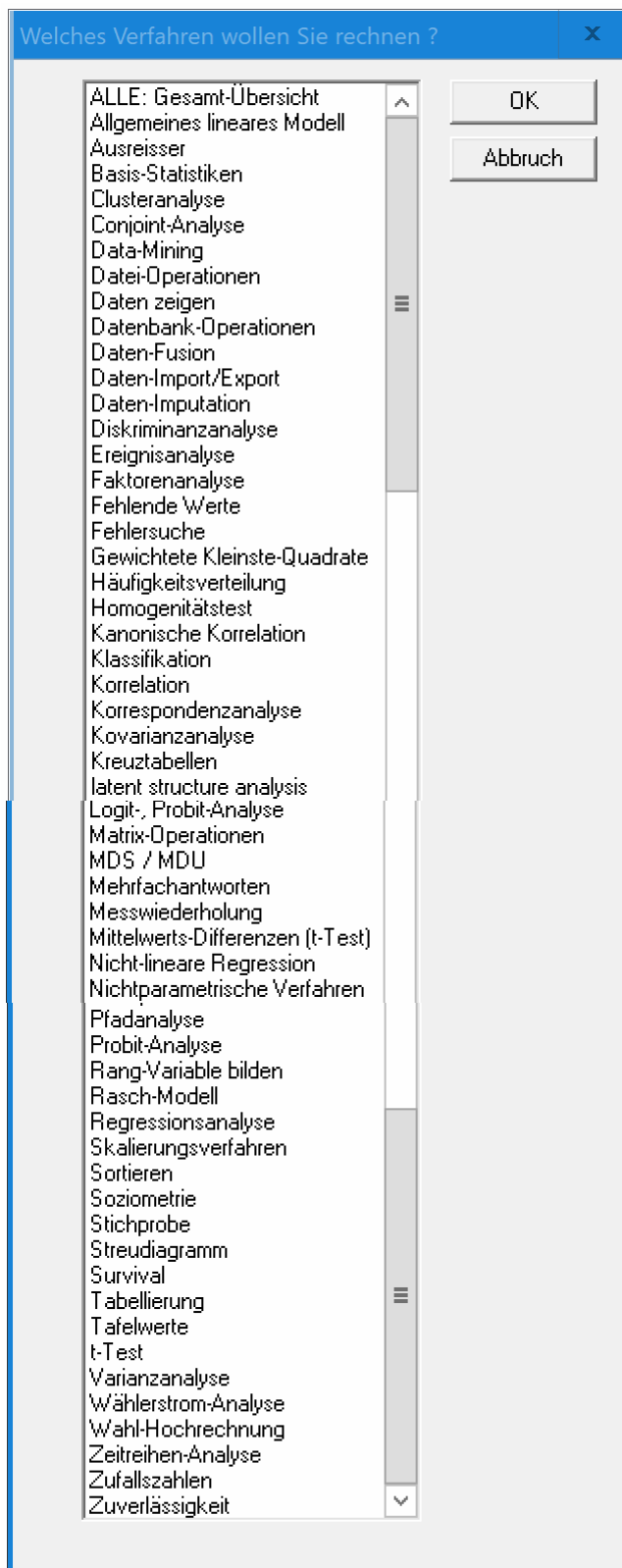
Gehen Sie folgendermaßen vor:

- a. Nach dem Start von Almo finden Sie unterhalb der Menüleiste nachfolgend abgebildete Knopfleiste. Klicken Sie in ihr auf "Verfahren".



Almo präsentiert Ihnen dann die nachfolgend abgebildete Übersicht über die vorhandenen Verfahren.

Sie werden gefragt: "Welches Verfahren wollen Sie rechnen ?"



b. Selektieren Sie als Verfahren „Allgemeines lineares Modell“. Also zeigt Ihnen dann eine Liste der zu diesem Verfahren vorhandener Maskenprogramme.

Für das gewählte Verfahren sind folgende Programme vorhanden  
 Klicken Sie in die Zeile, die mit 'Prog \_\_ ' beginnt

- Prog20mx: Allgemeines Lineares Modell \*
- Prog20mo: Allgemeines Lineares Modell mit vielen Optionen \*

---

- Prog20m6: Allgemeines Lineares Modell \*  
mit Eingabe einer fertigen Streuungsmatrix

---

- Prog20mk: Gewichtetes Allgemeines Lineares Modell für eine abhängige \*  
nominal-dichotome Variable

---

- Prog20m9: Gewichtetes Allgemeines Lineares Modell für eine abhängige \*  
nominal-polytome Variable

---

- Prog20m8: Allgemeines Lineares Modell für eine abhängige Rangvariable \*  
Prog ist äquivalent zu Prog08M1 (H- bzw. U-Test)

---

- Prog20am: Allgemeines Lineares Modell — wie Prog20mo \*  
zusätzlich mit (unabhängigen und abhängigen) Rangvariablen

---

- Prog20m1: Prognose mit dem Allgemeinen Linearen Modell

Sonderprogramme für Verfahren fitting constants I

- ProgFI\_2: Varianz-, Kovarianzanalyse mit 2 nominalen Faktoren
- ProgFI\_3: Varianz-, Kovarianzanalyse mit 3 nominalen Faktoren
- ProgFI\_4: Varianz-, Kovarianzanalyse mit 4 nominalen Faktoren
- ProgFI\_5: Varianz-, Kovarianzanalyse mit 5 nominalen Faktoren

Sonderprogramme für Verfahren fitting constants II (=SS-Typ II )

- ProgFII2: Varianz-, Kovarianzanalyse mit 2 nominalen Faktoren
- ProgFII3: Varianz-, Kovarianzanalyse mit 3 nominalen Faktoren
- ProgFII4: Varianz-, Kovarianzanalyse mit 4 nominalen Faktoren
- ProgFII5: Varianz-, Kovarianzanalyse mit 5 nominalen Faktoren

Sonderprogramme für Verfahren sequentiell (=SS-Typ I )

- ProgSq\_2: Varianz-, Kovarianzanalyse mit 2 nominalen Faktoren
- ProgSq\_3: Varianz-, Kovarianzanalyse mit 3 nominalen Faktoren
- ProgSq\_4: Varianz-, Kovarianzanalyse mit 4 nominalen Faktoren

Abbruch

Lade selektiertes Programm

Gebe Info

Die beiden ersten Programme, Prog20mx und Prog20mo sind die Standard-Masken-Programme. Mit diesen wird der Benutzer überwiegend rechnen.

Prog20mx ist bezüglich der Eingabe durch den Benutzer und der Ausgabe der Ergebnisse, stark vereinfacht. Prog20mo bietet viele Optionen, die der Benutzer aber negieren kann.

Prog20mk und Prog20m9 werden verwendet, wenn die abhängige Variable nominal ist und die modellbedingte Varianzheterogenität durch eine besondere Gewichtung beseitigt werden soll.

Prog20m8 und Prog20m9 sind zwei Programme, die es erlauben eine Rangvariable als abhängige zu verwenden.

Die Sonderprogramme sollten verwendet werden, wenn nicht das übliche "weighted squares of means" als Schätzverfahren eingesetzt wird, sondern "fitting constants I" oder II oder das sequentielle Verfahren. Die beiden letzteren werden bei SPSS "SS-Typ II" und "SS-Typ I" genannt. Zu den Sonderprogrammen siehe Abschnitt P20.7. zur Wahl eines der Sonderprogramme Abschnitt P20.7.2.3.

- c. Selektieren Sie nun durch Einmal-Klick das Programm **Prog20mx**. Wenn Sie auf den Knopf "Gebe Info" klicken, dann erhalten Sie eine kurze Übersicht darüber, was das Programm leistet.

Wenn Sie auf den Knopf „Lade selektiertes Programm“ klicken, dann öffnet Almo das Maskenprogramm. Es setzt sich aus hintereinander stehenden Dialogboxen zusammen.

### **Programm-Maske Prog20mx**

Prog20mx.Msk  
Allgemeines lineares Modell

Abhängige Variable : quantitativ/ordinal oder nominal  
Unabhängige Variable: nominal und quantitativ und ordinal

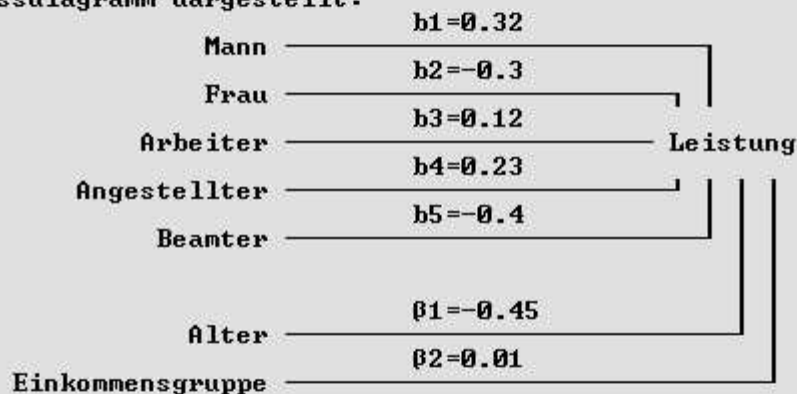
Beispiel: Der Einfluss der nominalen Variablen Geschlecht (U1) und Beruf (U3) sowie der quantitativen Variablen Alter (U6) und Kinderzahl(U8) sowie der ordinalen Variablen Einkommensgruppe (U7) auf die Leistung in einem Test (abhängige quantitative Variable) soll ermittelt werden.

Sie rechnen also folgendes Modell:

$$\begin{aligned} \text{Leistung} = & b_1 * \text{Mann} & + & b_2 * \text{Frau} & + & \\ & b_3 * \text{Arbeiter} & + & b_4 * \text{Angestellter} & + & b_5 * \text{Beamter} & + \\ & \beta_1 * \text{Alter} & + & \beta_2 * \text{Einkommensgruppe} & + & \\ & \text{Konstante} & & & & & \end{aligned}$$

Möglich ist es auch Interaktionen als unabh. Variable einzuführen also z.B. Mann/Arbeiter, Mann/Angestellter, ..... Frau/Beamter

Als Flussdiagramm dargestellt:



Die Variablen "Mann", "Frau" "Arbeiter" etc. sind 0-1 kodiert  
b1 bis b5 = das sind die Effekte der Ausprägungen der nominal. Var.  
 $\beta_1, \beta_2$  = das sind die Regressionskoeffizienten der quantitat. und der ordinalen Variablen  
Konstante = das ist die Konstante

Almo berechnet nicht nur diese Koeffizienten, sondern auch die durch die unabhängigen Variablen in der abhängigen Variablen erklärte Streuung, sowie eine Vielzahl weiterer Koeffizienten.

Als abhängige Variable sind erlaubt:

1. Eine oder mehrere quantitativen Variable  
oder eine oder mehrere ordinale Variable  
oder quantitative und ordinale Variable gemischt

oder (exklusiv)

2. Eine nominale Variable mit beliebig vielen Ausprägungen

Siehe Handbuch, Abschnitt P20.8.0

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

Hilfe

Hilfe

1 Speicher fuer x Variable   
Vereinbare Variable= 20 ;

2  Option: Weitere Vereinbarungen - nur wenn Almo dazu auffordert

3 Datei der Variablennamen   
  "C:\Almo7\TESTDAT\Uarnamen.nam"  
  zeige zeige = Namensdatei in Output zeigen  
leer = nicht

4 Freie Namensfelder   
    
 erzeuge zusätzliche Namensfelder

5 Datei aus der gelesen wird  bei Datei-Problemen  
 "C:\Almo7\Testdat\TESTDAT.FRE"  
 frei Format der Daten   
  U1:20 der Datensatz enthält diese Variablen  
Bei Format DIREKT schreiben Sie: alle\_U

6  Wenn Dateiformat FIX oder Nicht-Standard-FREI

7 Analyse-Variable: Abhängige Variable   
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  
  Leistung  
-----  
ordinale abhängige Zielvariable   
    
-----  
nominale abhängige Zielvariable

Analyse-Variable: Unabhängige Variable Hilfe

nominale unabhängige Variable Hilfe

**Geschlecht, Beruf**

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

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

---

quantitative unabhängige Variable Hilfe

**Alter, Kinderzahl**

---

ordinale unabhängige Variable Hilfe

**█**

8

**Option: Umkodierungen und Kein-Wert-Angaben**

9

**Ausgabe der Ergebnisse**

**1** 0= Ergebnisse in voller Länge ausgeben  
 1= Ergebnisse etwas verkürzt ausgeben  
 2= Ergebnisse stark verkürzt ausgeben

10

### *P20.8.0.1 Erläuterung zu den Eingabe-Boxen von Prog20mx.Msk*

#### **Eingabe-Box: Vereinbarungen**

Siehe P0.1.

#### **Eingabe-Box: Option: Weitere Vereinbarungen**

Siehe P0.2.

#### **Eingabe-Box: Datei der Variablennamen**

Siehe P0.3.

#### **Eingabe-Box: Freie Namensfelder**

Siehe P0.3.

#### **Eingabe-Box: Datei aus der gelesen wird**

Siehe P0.4.

#### **Eingabe-Box: Wenn Dateiformat FIX oder Nicht-Standard-FREI**

Siehe P0.4.

#### **Eingabe-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

← [ ] Leistung →

---

ordinale abhängige Zielvariable Hilfe

← [ ] █ →

---

nominale abhängige Zielvariable Hilfe

← [ ] █ →

Erlaubt sind:

1. a. Eine oder mehrere quantitative Variable  
b. oder eine oder mehrere ordinale Variable  
c. oder gemischt: beliebig viele quantitative und beliebig viele ordinale Variable

oder (exklusiv)

2. Eine nominale Variable mit beliebig vielen Ausprägungen

Nicht erlaubt ist es also, nominale Variable zu mischen mit quantitativen oder ordinalen Variablen.

Es ist zulässig mehrere abhängige Variable einzugeben. Also rechnet dann eine multivariate Analyse. Dies gilt nicht für nominale Variable. Es darf nur eine nominale Variable als abhängige eingeführt werden.

Dichotom-nominale Variable können als quantitative abhängige Variable angegeben werden. Dann ist es auch möglich, mehrere dichotom-nominale Variable als abhängige anzugeben.

Werden ordinale Variable als unabhängige oder abhängige Variable eingeführt, dann ermittelt Also die Streuungsmatrix nach einem modifizierten tau-b-Kalkül. Siehe dazu Abschnitt P20.6.9. Es muss darauf hingewiesen werden, dass die Berechnung parametrischer Tests (F-Test, t-Test), wie sie Also verwendet, dann sehr problematisch ist. Siehe auch die Regeln zur Einbeziehung ordinaler Variablen in P20.8.2.

### Eingabe-Box: Analyse Variable: Unabhängige Variable

Sie können 3 Typen von unabhängigen Variablen einführen:

- nominale
- quantitative
- ordinale

**Quantitative Variable** können beliebig mit Ganzzahlen oder Dezimalzahlen kodiert sein.

**Nominale Variable** werden normalerweise ganzzahlig mit Schrittweite 1 kodiert sein. Beispiel:

Beruf	Codeziffer
-----	-----
Arbeiter	1
Angestellter	2
Beamter	3
Bauer	4
.	.
.	.
.	.

Im vorliegenden Maskenprogramm ist es jedoch zulässig, dass sie auch Dezimalwerte besitzen. Beispiel

Beruf	Codeziffer
-----	-----
Arbeiter	10.5
Angestellter	12
Beamter	13.99
Bauer	8.1
.	.
.	.
.	.

Als Codeziffern sind hier etwa Qualifikationspunkte verwendet worden. Die Variable soll aber als *nominale* in der Analyse behandelt werden.

Almo kodiert diese Werte zwangsweise um, und zwar in folgender Weise

alte Codeziffer	neue Codeziffer
-----	-----
8.1	1
10.5	2
12	3
13.99	4
.	.
.	.
.	.

Es wird bei 1 begonnen und mit Schrittweite 1 aufsteigend umkodiert. Beachte: Bauer hat mit 8.1 die niedrigste Codeziffer. Almo weist ihm die neue niedrigste Codeziffer 1 zu.

Wenn der Benutzer diese zwangsweise Umkodierung vermeiden will, weil er anders als Almo umkodieren würde, dann muss er in der Eingabe-Box

"Kein-Wert-Angabe und Umkodierungen"

selbst umkodieren, etwa so:

```
Beruf(8.1=1; 10.5=2; 12, 13.9=3)
```

Hier hat der Benutzer Angestellte und Beamte in der Art ihrer Tätigkeit als sehr ähnlich erachtet und sie deswegen in einer Kategorie zusammengefasst.

**Ordinale Variable** werden normalerweise ganzzahlig mit Schrittweite 1 kodiert sein. Beispiel:

Schulbildung	Codeziffer
-----	-----
Volksschule	1
Hauptschule	2
Gymnasium	3
Fachschule	4
Universität	5

Im vorliegenden Maskenprogramm ist es jedoch zulässig, dass sie auch Dezimalwerte besitzen. Beispiel

Schulbildung	Codeziffer
-----	-----
Volksschule	8.2
Hauptschule	10.5
Gymnasium	13.1
Fachschule	13.9
Universität	19.22

Als Codeziffern sind etwa die durchschnittlichen Ausbildungsjahre verwendet worden

Almo kodiert diese Werte zwangsweise um, und zwar in folgender Weise

alte Codeziffer	neue Codeziffer
8.2	1
10.5	2
13.1	3
13.9	4
19.22	5

Es wird bei 1 begonnen und mit Schrittweite 1 aufsteigend umkodiert.

Wenn der Benutzer diese zwangsweise Umkodierung vermeiden will, weil er anders als Almo umkodieren würde, dann muss er in der Eingabe-Box

"Kein-Wert-Angabe und Umkodierungen"

selbst umkodieren, etwa so:

Schulbildung (8.2=1; 10.5=2; 13.1, 13.9=3; 19.22=4)

Hier hat der Benutzer Gymnasium und Fachschule für gleichrangig erachtet und in einer Kategorie zusammengefasst.

Werden ordinale Variable als unabhängige oder abhängige Variable eingeführt, dann ermittelt Almo die Streuungsmatrix nach einem modifizierten tau-b-Kalkül. Siehe dazu, Abschnitt P20.6.9. Es muss darauf hingewiesen werden, dass die Berechnung parametrischer Tests (F-Test, t-Test), wie sie Almo verwendet, dann sehr problematisch ist. Siehe auch die Regeln zur Einbeziehung ordinaler Variablen in P20.8.2.

### Variablenhierarchie und gleichrangige Variablengruppen

Es ist auch möglich, eine hierarchische Ordnung zwischen den unabhängigen Variablen einzuführen.

Wir würden aber raten, dies nur zu tun, wenn man einen guten Grund dafür hat.

Die Vorgehensweise ist folgende: Die unabhängigen quantitativen Variablen können in hierarchische und/oder gleichrangige Gruppen gruppiert werden. Das geschieht dadurch, dass zwischen die unabhängige quantitative Variablen Schrägstriche und/oder senkrechte Striche geschrieben werden.

Die unabhängigen nominalen Variablen können in hierarchische Gruppen gruppiert werden (jedoch nicht in gleichrangige).

Das geschieht dadurch, dass zwischen die unabhängige nominalen Variablen Schrägstriche (aber keine senkrechten Striche) geschrieben werden.

Beispiel:



Die Schrägstriche bezeichnen die Variablen-Hierarchie. Zuerst wird V6 eingeführt, dann V7 und an V6 angepasst, dann V8 und an V6,7 angepasst, dann V20 und an V6,7,8 angepasst.

Anstelle der Variablennummern könnten hier und in den folgenden Beispielen auch die Variablennamen verwendet werden.

Eine gruppenweise Hierarchie entsteht z.B. durch:

Zuerst werden V6,7 eingeführt und V6 und V7 gegenseitig "auspartielliert". Dann wird V8,14 eingeführt, an V6,7 angepasst und dann noch gegenseitig "auspartielliert" etc.

Durch einen senkrechten Strich, z. B.

werden 3 Variablengruppen V6,7 und V8,14 und V19,20 gebildet. Für jede Variablengruppe wird die erklärte Streuung, der F-Wert und all die anderen Koeffizienten ermittelt, die Almo standardmäßig errechnet. Dadurch ist es also auch möglich die so genannten "partiellen multiplen Korrelationen" der 3 Variablengruppen zu ermitteln.

Möglich ist z.B. auch

Dabei bestehen die gleichrangigen Gruppen innerhalb der hierarchischen Gruppe. In unserem Beispiel besteht die 2. hierarchische Gruppe aus

V8,14, 15,16

*Innerhalb* dieser existieren die zwei gleichrangigen Gruppen.

V8,14 und V15,16

Der Kalkül ist folgender: Zuerst wird die 1. hierarchische Gruppe V6,7 eingeführt. Dann wird die 2. hierarchische Gruppe V8,14,15,16 eingeführt und an die 1. Gruppe angepasst. Erst danach werden die erklärten Streuungen und partiellen multiplen Korrelationen der beiden gleichrangigen Gruppen berechnet. Zum Schluss wird die 3. hierarchische Gruppe an die 1. und 2. angepasst. Der Vorgang der Anpassung bzw. "Auspartiellierung" wird in Teil II, Abschnitt P20.12 sehr ausführlich erläutert.

Hierarchische Gruppen *innerhalb* von gleichrangigen Gruppen sind nicht möglich.

V6,7 | 8,14 / 15,16 | 19,20

Der Benutzer möchte mit dieser Eingabe folgende 3 gleichrangige Gruppen bilden:

V6,7 und V8,14,15,16 und V19,20

wobei die mittlere Gruppe aus den beiden hierarchisch gereihten Gruppen V8,14 und V15,16 besteht. *Das ist nicht möglich.* Almo würde diese Eingabe als die zwei hierarchisch geordneten Gruppen V6,7,8,14 und V15,16,19,20 interpretieren, wobei jede nochmals in 2 gleichrangige Gruppen unterteilt ist.

BEACHTE: Die unabhängigen nominalen Variablen können nur in hierarchische Gruppen gruppiert werden - jedoch nicht in gleichrangige. Zwischen sie dürfen also nur Schrägstriche (aber keine senkrechten Striche) geschrieben werden.

BEACHTE: Nicht nur unabhängige quantitative sondern auch unabhängige ordinale Variable können in eine hierarchische Reihenfolge gebracht werden, sowie hierarchisch und gleichrangig gruppiert werden. Das darf auch gemischt geschehen. Mit ordinalen Variablen geht das nur, wenn die Programm-Maske Prog20mo eingesetzt wird. Mit Prog20mx geht das nicht. Siehe die ausführliche Darstellung in Teil II, Abschnitt P20.12.

BEACHTE: Quantitative/ordinale unabhängige Variable und nominale unabhängige Variable können nicht in eine hierarchische Reihenfolge gebracht werden

BEACHTE: Wenn Sie bei den unabhängigen Variablen sowohl nominale als auch quantitative/ordinale haben, dann nimmt Almo zuerst eine gegenseitige Anpassung zwischen den nominalen und den quantitativen/ordinalen Variablen vor. Erst danach werden innerhalb der nominalen bzw. innerhalb der quantitativen/ordinalen Variablen die vom Benutzer gewünschten hierarchischen oder gleichrangigen Gruppen gebildet. Siehe Teil II, Abschnitt P20.10 ff.

BEACHTE: Sollen hierarchische Gruppen gebildet werden, dann muß als Verfahren "fitting\_constants\_I" eingestellt werden. Dies ist nur in der Programm-Maske Prog20mo möglich, die im nachfolgenden Abschnitt dargestellt wird. Im hier behandelten Programm Prog20mx geht das nicht. Almo schaltet in diesem Programm deswegen automatisch auf "fitting\_constants\_I" um, wenn der Benutzer hierarchische Gruppen bildet. Gleichrangige Gruppen können ohne Komplikationen in Prog20mx gebildet werden.

Für das oben abgebildete letzte Beispiel gibt Almo folgendes Ergebnis aus (hier etwas gekürzt). Der Benutzer muss in der untersten Eingabebox auf "volle Ausgabe" einstellen.

Koeffizienten hinsichtlich der abhaengigen Variablen V5 Leistung

Variable	standard.		erklaeerte Streuung	part. Korrel.	F-Wert	Signifikanz	
	Regr. koeff.	Regr. koeff.				p	(1-p)100
V19 CD	0.0171	0.0367	0.0488	0.017	0.014	0.905	9.49
V20 Bewertung	-0.1068	-0.0771	0.6539	-0.063	0.188	0.667	33.34
<b>hierarch. Gruppe</b> V19,20			0.7031	0.065	0.101	0.898	10.19
V8 Kinderzahl	-0.1201	-0.0907	2.8207	-0.130	0.809	0.373	62.72
V14 AB	0.2236	0.4564	8.9170	0.227	2.558	0.117	88.33
<b>zusammen</b> (gleichrangige Gruppe) V8,14			11.0628	0.251	1.587	0.214	78.61
V15 AC	0.1495	0.3025	4.0189	0.155	1.153	0.288	71.16
V16 AD	-0.0750	-0.1746	0.9730	-0.077	0.279	0.599	40.07
<b>zusammen</b> V15,16			4.1468	0.157	0.595	0.561	43.92
<b>hierarch. Gruppe</b> V8,14,15,16			15.4062	0.293	1.105	0.366	63.43
V6 Alter	0.0440	0.0380	0.3727	0.048	0.107	0.746	25.42
V7 Einkommen	0.0076	0.0058	0.0109	0.008	0.003	0.953	4.66
<b>hierarch. Gruppe</b>			0.3735	0.048	0.054	0.938	6.21

Mit dem Wort **zusammen** wird die jeweilige gleichrangige Variablen­gruppe bezeichnet.

Beispiel: Die hierarchische Gruppe V8,14,15,16 erklärt 15.4062 Streuungseinheiten. Dem entspricht ein partieller multipler Korrelationskoeffizient für diese Gruppe von 0.293. Innerhalb dieser hierarchischen Gruppe erklärt dann die erste Untergruppe V8,14 11.0628 Streuungseinheiten. Das entspricht einem partiellen multiplen Korrelationskoeffizienten von 0.251. Die zweite Untergruppe erklärt 4.1468 Streuungseinheiten und besitzt somit eine partielle multiple Korrelation von 0.157.

Die hier kurz behandelte hierarchische Reihung von Variable, sowie die hierarchische und gleichrangige Gruppierung von Variablen wird ausführlich in Teil II, Abschnitt P20.10 bis P20.12 dargestellt.

## 2. Eingabefeld: Interaktionen

Siehe dazu auch P20.6.5.2.

Geben sie 0 ein, wenn Sie keine Interaktionen zwischen den nominalen Variablen bilden wollen. 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

wird in das Eingabefeld 2 geschrieben, dann werden die Interaktionen höherer Ordnung, also 3. Ordnung, nicht gebildet.

**Empfehlung:** Interaktionen höherer Ordnung verursachen eine wesentlich längere Rechenzeit. Erfahrungsgemäß ist es auch schwierig, Dreier-Interaktionen und Interaktionen noch höherer Ordnung inhaltlich zu interpretieren. Sogar bei der Interpretation von Zweier-Interaktionen tut man sich schwer. Wir empfehlen deswegen, bei der Festlegung der Interaktionsordnung nicht zu übertreiben. Werden Interaktionen höherer Ordnung angegeben, dann kann auch das Problem der "leeren Zellen" auftreten. Dies ist für das Allgemeine Lineare Modell ein ernsthaftes Problem. In Abschnitt P20.7.7 haben wir dieses Problem ausführlich behandelt.

Es besteht auch die Möglichkeit nur einige ausgewählte Interaktionen zu bilden, z.B. die Interaktionen

AC BC

Die Vorgehensweise ist dann (etwas kompliziert) folgende

(Siehe dazu auch Abschnitt P20.14)

1. In der Eingabe-Box "Freie Namensfelder" schreiben Sie die 2 Namen

Name 21=IntAC;  
Name 22=IntBC;

Die Namen können Sie wählen, wie Sie wollen. Verwenden Sie dabei hinter "Name ..." Variablen-Nummern die frei sind, am besten Nummern, die höher sind als die Nummer der letzten eingelesenen Variablen.

2. In der Eingabe-Box "Analysevariable: Unabhängige Variable" geben Sie als unabhängige nominale Variable zusätzlich IntAC und IntBC an. Im Eingabefeld steht dann (in unserem Beispiel)

A, B, C, IntAC, IntBC

3. Im Eingabefeld für die Interaktionen schreiben Sie:

IntAC ist A mal C, IntBC ist B mal C

Sie müssen also angeben, aus welchen nominalen Variablen die Interaktionsvariablen gebildet werden sollen. Die Eingabe-Box sieht also folgendermaßen aus:



Anstelle des Wortes "mal" kann auch das Multiplikationszeichen verwendet werden. Die Eingabe würde dann so lauten:



Siehe dazu die ausführliche Darstellung im Abschnitt P20.14 im Almo-Dokument Nr. 13a "Allgemeines lineares Modell II". Dort wird auch auf die Problematik dieser Modelle mit unvollständigen Interaktionsvariablen eingegangen.

Zum Ein- oder Ausschluss von Interaktionen gilt prinzipiell:

Werden Interaktionen x-ter Ordnung oder "selbst gebildete" Interaktionsvariable eingeschlossen, dann wird dadurch die Fehlerstreuung reduziert. Dadurch werden die Korrelationskoeffizienten insbesondere für die Haupteffekte erhöht. Sind Interaktionen nicht interpretierbar oder sogar unsinnig, dann sollten sie ausgeschlossen bleiben und damit Teil der Fehlerstreuung werden.

### 3. Eingabefeld: Paarweise Vergleiche (Kontraste)

Beispiel: Die unabhängige nominale Variable A besitze 3 Ausprägungen. Almo berechnet dann die paarweisen Vergleiche (Kontraste) zwischen diesen 3 Ausprägungen und liefert folgendes Ergebnis (gekürzt):

Kontraste	t-Wert	Signifikanz
		p (1-p)100
A1 - A2	0.2876 0.4987	0.6197 38.03%
A1 - A3	1.2086 1.8518	0.0696 93.04%
A2 - A3	0.9210 1.5513	0.1269 87.31%

**Eingabe-Box: Option: Kein\_Wert-Angabe und Umkodierungen**  
 Siehe P0.5.

## P20.8.1 Eingabe in Maskenprogramm mit Optionen Prog20mo

Prog20mo.Msk  
Allgemeines lineares Modell mit Optionen  
mit Prognosewerten und Residuen  
bei nominaler abhängiger Variablen mit Prognoseerfolg

Unabhängige Variable: nominal und quantitativ und ordinal  
Abhängige Variable : nominal oder quantitativ/ordinal

Beispiel: Der Einfluss der nominalen Variablen Geschlecht (U1) und Beruf (U3) sowie der quantitativen Variablen Alter (U6) und Kinderzahl(U8) sowie der ordinalen Variablen Einkommensgruppe (U7) auf die Leistung in einem Test (abhängige quantitative Variable) soll ermittelt werden.

Sie rechnen also folgendes Modell:

$$\text{Leistung} = b_1 * \text{Mann} + b_2 * \text{Frau} + b_3 * \text{Arbeiter} + b_4 * \text{Angestellter} + b_5 * \text{Beamter} + \beta_1 * \text{Alter} + \beta_2 * \text{Einkommensgruppe} + a$$

Möglich ist es auch Interaktionen als unabh. Variable einzuführen also z.B. Mann/Arbeiter, Mann/Angestellter, ..... Frau/Beamter

Als Flussdiagramm dargestellt:

	b1=0.32	
Mann	-----	]
Frau	-----	
	b2=-0.3	
Arbeiter	-----	]
Angestellter	-----	
	b3=0.12	
Beamter	-----	]
	b4=0.23	
	b5=-0.4	
Alter	-----	]
Einkommensgruppe	-----	
	\beta_1=-0.45	
	\beta_2=0.01	

Die Variablen "Mann", "Frau" "Arbeiter" etc. sind 0-1 kodiert  
b1 bis b5 = das sind die Effekte der Ausprägungen der nominal. Var.  
\beta\_1, \beta\_2 = das sind die Regressionskoeffizienten der quantitat. und der ordinalen Variablen  
a = das ist die Konstante

Almo berechnet nicht nur diese Koeffizienten, sondern auch die durch die unabhängigen Variablen in der abhängigen Variablen erklärte Streuung, sowie eine Vielzahl weiterer Koeffizienten.

Als abhängige Variable sind erlaubt:

- Eine oder mehrere quantitativen Variable oder eine oder mehrere ordinale Variable oder quantitative und ordinale Variable gemischt oder (exklusiv)
- Eine nominale Variable mit beliebig vielen Ausprägungen

Siehe Handbuch, Abschnitt P20.8.1

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

1 Speicher fuer x Variable

Vereinbare Variable= **20** ;

2  Option: Weitere Vereinbarungen - nur wenn Almo dazu auffordert

3 Datei der Variablennamen

"C:\Almo7\TESTDAT\Uarnamen.nam"

**zeige**      zeige = Namensdatei in Output zeigen  
leer = nicht

4 Freie Namensfelder

**erzeuge zusätzliche Namensfelder**

5 Datei aus der gelesen wird  bei Datei-Problemen

"C:\Almo7\Testdat\TESTDAT.FRE"

**frei**      Format der Daten

**U1:20**      der Datensatz enthält diese Variablen  
Bei Format DIREKT schreiben Sie: alle\_U

6  Wenn Dateiformat FIX oder Nicht-Standard-FREI

7 Analyse-Variable: Abhängige Variable

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**

**Leistung**

---

**ordinale abhängige Zielvariable**

---

**nominale abhängige Zielvariable**

8

Analyse-Variablen: Unabhängige Variable Hilfe

    nominale unabhängige Variable Hilfe

    ↔   **Geschlecht, Beruf**

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

    ↔   **Geschlecht, Beruf**      paarweise Vergleiche (Kontraste) für die  
    nominalen unabhängigen Variablen rechnen

---

    quantitative unabhängige Variable Hilfe

    ↔   **Alter, Kinderzahl**

---

    ordinale unabhängige Variable Hilfe

    ↔

 Option: Ein- und Ausschliessen von Untersuchungseinheiten

 Option: Umkodierungen und Kein-Wert-Angaben

 Option: Spezielle Kein-Wert-Behandlung

 Option: Ausreisser vom Typ 1 identifizieren Hilfe

 Option: Untersuchungseinheiten gewichten

 Option: Streuungsmatrix












 Option: Verfahren

 Option: Nenner für Varianz und Kovarianz

 Option: Behandlung eventueller Multikollinearität

 Option: Spezielle Programm-Optionen

Hilfe

	Option: Programm-Optionen lt. Handbuch
	Option: Gewichtete Kleinste-Quadrate-Schätzung
	Option: Prognosewerte und Residuen
	Option: Wertemuster
	Option: "Aussehen" der auszugebenden Tabelle bzw. Matrix
	Grafik-Optionen
	Option: Die errechnete Streuungsmatrix in eine Datei geben
	Option: Die errechneten Koeffizienten in eine Datei speichern
<b>Ausgabe der Ergebnisse</b>	
 <input type="text" value="0"/>	0= Ergebnisse in voller Länge ausgeben 1= Ergebnisse etwas verkürzt ausgeben 2= Ergebnisse stark verkürzt ausgeben
 <input type="text" value="0"/>	1= Basisstatistiken ausgeben 2= Basisstatistiken und "diverse Werte" ausgeben 0= nicht
	Option: Auf einzelne Teile der Ausgabe verzichten
<b>Programmende</b>	

### ***P20.8.1.1 Erläuterungen zu den Eingabe-Boxen von Maskenprogramm Prog20mo.Msk***

Wir erläutern im Folgenden nur jene Boxen, die nicht bereits bei Prog20mx.Msk erklärt wurden.

**Eingabe-Box 1 bis Eingabe-Box 8:** Siehe Erläuterungen zu Prog20mx in P20.8.0.1


#### **Eingabe-Box: Ein- und Ausschließen von Untersuchungseinheiten**

Siehe Almo-Dokument Nr. 0 "Arbeiten mit Almo.PDF", Abschnitt P0.7.

#### **Eingabe-Box: Umkodierungen und Kein-Wert-Angaben**

Siehe P0.5.

#### **Eingabe-Box: Option: Spezielle Kein-Wert-Behandlung**

	Option: Spezielle Kein-Wert-Behandlung
---	--

Die Voreinstellung ist das paarweise Ausscheiden. D.h.: wird die Optionsbox nicht geöffnet, dann wird diese Methode eingesetzt.

Optionsbox geöffnet:

↓ Loesche wieder diese Box

Option: **Spezielle Kein-Wert-Behandlung**

↑ ↓ !

0= Kein-Wert-Fälle in Analyse-Variable nicht vorhanden

1= **Paarweises Ausscheiden I** <---- ist Voreinstellung

2= **Paarweises Ausscheiden II**

- a. Paarweises Ausscheiden bei quantitativen und ordinalen Variablen.
- b. Vollständiges Ausscheiden bei nominalen Variablen und deren Interaktionen, wenn auch nur eine der nominalen Analyse-Variablen den Wert "Kein\_Wert" besitzt

3= **Vollständiges Ausscheiden**  
Vollständiges Ausscheiden des gesamten Datensatzes, wenn auch nur eine der Analyse-Variable "Kein\_Wert" ist

4= **Mittelwert-Einsetzung I** Hilfe

Für Kein\_Wert wird eingesetzt:

- a. bei quantitativen Variablen der Mittelwert
- b. bei ordinalen Variablen der Median.  
Der zum Median nächst gelegene empirische Skalenwert wird dann eingesetzt
- c. bei nominalen Variablen der Erwartungswert

5= **Mittelwert-Einsetzung II** Hilfe

Für Kein\_Wert wird eingesetzt:

- a. bei quantitativen Variablen der zum Mittelwert naechste empirisch vorkommende Wert
- b. bei ordinalen der Median (wie bei 4)
- c. bei nominalen Variablen der Erwartungswert (wie bei 4)

6= **Mittelwert-Einsetzung III** Hilfe

Für Kein\_Wert wird eingesetzt:

- a. bei quantitativen Variablen der Mittelwert +/- einem normalverteilten Zufallswert mit Mittelwert=0 und Standardabweichung der Variablen
- b. bei ordinalen der Median (wie bei 4)  
Ist die Variable mit gleicher Schrittweite kodiert, dann wird ein Wert X errechnet, der sich ergibt aus Median +/- einem normalverteilten Zufallswert mit Mittelwert=0 und Standardabweichung in der Größe des halben Quartilsabstands der Variablen. Der zu X nächst gelegene empirische Skalenwert wird dann eingesetzt
- c. bei nominalen der wahrscheinlichste Ausprägungswert

7= **Mittelwert-Einsetzung IU**

a. bei quantitativen Variablen zunächst wir bei 6  
Der nächst gelegene empirische Skalenwert  
wird dann eingesetzt

b. bei ordinalen der Median (wie bei 6)

c. bei nominalen Variablen (wie bei 6)

---

1 = nur relevant für Allgemeines Lineares Modell (ALM) !!  
1 = wenn abhängige Variable Kein-Wert besitzt, dann  
Datensatz aus Analyse vollständig ausschliessen  
unabhängig davon welche Kein-Wert-Behandlung  
oben im ersten Eingabefeld dieser Box gewählt wurde

0 = gewählte Kein-Wert-Behandlung gilt auch für  
abhängige Variable

---

Startwert für Zufallsgenerator   
fuer Kein-Wert-Behandlung 6 und 7

---

als "gemeinsame" Fallzahl für Signifikanztest  
wird verwendet - wenn Kein-Wert-Behandlung =1 oder =2  
und wenn Kein-Wert-Fälle auftreten:

0 = die kleinste Fallzahl, aus der die Co-Streuungen zwischen  
je 2 Variablen i und k errechnet wurden

1 = das harmonisches Mittel aus den Fallzahlen, aus denen  
die Co-Streuungen zwischen den Variablen errechnet wurden

2 = die Zahl der Fälle, die in allen  
Analysevariablen valide Werte besitzen

3 = die Zahl der eingelesenen Fälle

Es ist nahezu normal, dass manche Untersuchungseinheiten in manchen Variablen keine Werte besitzen. Bei Befragungen wird beispielsweise von einem relativ hohen Prozentsatz der Befragten die Frage nach dem Einkommen nicht beantwortet. In diesem Falle wird man dann etwa -1 als Einkommenshöhe kodieren. Anders formuliert:  
-1 ist der "Kein-Wert-Code" für die Einkommensvariable

In der Eingabe-Box "Kein-Wert-Angabe und Umkodierungen" muss dann stehen

Einkommen (-1 = Kein\_Wert)

Almo überführt dann den Wert -1 in einen internen Almo-Code.

Der interne Almo-Code für "Kein\_Wert" ist eine riesige negative Zahl. Der Vorgang ist also folgender: Die Zahl -1 wird in diese riesige negative Zahl umkodiert. Diese Zahl wird von allen Almo-Programmen als "Kein-Wert-Code" begriffen. Wenn Almo auf diese Zahl stößt, dann weiß es, dass es diese auf eine besondere Weise behandeln muss.

Wenn Almo im Programm auf diesen internen Kein\_Wert-Code stößt, dann wird beim Errechnen der Streuungsmatrix (z.B. der Korrelationsmatrix) standardmäßig das "paarweise Ausscheiden" durchgeführt. Wenn Sie das akzeptieren, dann brauchen Sie diese Optionen-Box nicht zu aktivieren. Wenn nicht, dann stehen Ihnen folgende 7 Vorgehensweisen zur Verfügung:

### **Eingabe: 0 Kein-Wert-Fälle nicht vorhanden**

Der Benutzer ist sich sicher, dass keine Kein-Wert-Fälle vorliegen. Almo kann die Streuungsmatrix dann schneller und speicherplatzsparend errechnen. Der Zeitgewinn ist bei kleinerer Variablenzahl und kleineren Datenmengen kaum spürbar.

### **Eingabe: 1 Paarweises Ausscheiden**

Almo führt das "paarweise Ausscheiden" durch. Betrachten wir ein Beispiel: Es sollen die 3 Variablen x, y, z korreliert werden bzw. deren Kovarianzen sollen ermittelt werden. Bei der 12. Untersuchungseinheit besitzt x keinen Wert. Dann werden die Kreuzprodukte etc. zwischen x einerseits und y sowie z andererseits für diese Untersuchungseinheit nicht berechnet. Für die 12. Untersuchungseinheit liegen jedoch valide Werte für y und z vor, so dass die Kreuzprodukte etc. zwischen y und z berechnet werden können. Almo merkt sich dann, dass die Korrelation xy und xz auf einer um 1 verringerten Zahl von Untersuchungseinheiten beruht. Die Zahl der Untersuchungseinheiten, auf der die jeweilige Korrelation zwischen dem Variablenpaar ik beruht, wird in der Ergebnis-Ausgabe mitgeteilt. Das "paarweise Ausscheiden" wird auch durchgeführt, wenn die Variable x, für die ein Wert fehlt eine nominale ist. Da die nominalen Variablen in Dummies aufgelöst werden, werden in diesem Falle die Kreuzprodukte etc. aller Dummies von x mit y und z nicht berechnet. Siehe die ausführliche Diskussion der Probleme des „paarweisen Ausscheidens“ im Almo-Dokument Nr. 24 „Statistische Datenanalyse I“, Abschnitt P45.12.4.

### **Eingabe: 2 Paarweises Ausscheiden II**

Liegt für eine Untersuchungseinheit auch nur für eine nominale Variable A kein Wert vor, dann werden die Werte aller nominaler Variablen nicht berücksichtigt, auch wenn für die anderen nominalen Variablen B, C, ... Werte vorhanden sind. Zwischen den quantitativen / ordinalen Variablen hingegen wird das "paarweise Ausscheiden" durchgeführt.

### **Eingabe: 3 Vollständiges Ausscheiden**

Liegt für eine Untersuchungseinheit auch nur in einer Variablen kein Wert vor, dann wird die gesamte Untersuchungseinheit aus der Analyse ausgeschlossen. Diese Vorgehensweise wird gelegentlich "listenweises Ausscheiden" genannt.

### **Eingabe: 4 Mittelwert-Einsetzung I**

Almo ermittelt zuerst Mittelwerte (für quantitative Variable), Median (für ordinale Variable) und den Erwartungswert (für nominale Variable).

Almo gibt diese Werte aus.

Für Kein\_Wert wird eingesetzt:

- a. bei quantitativen Variablen der Mittelwert
- b. bei ordinalen Variablen der Median (=der mittlere Wert) liegt der Median nicht auf einem empirischen Wert, sondern zwischen 2 empirischen Werten, dann wird der nächst gelegene Nachbarwert als KW-Einsetzungswert verwendet.
- c. bei nominalen Variablen die zum Erwartungswert nächste empirisch vorkommende Codeziffer

Die Berechnung des Erwartungswerts soll an einem Beispiel gezeigt werden. Die nominale Variable sei der Beruf mit den 3 Ausprägungen Arbeiter, Angestellte, Sonstige. Dabei wurden folgende Häufigkeiten ermittelt.

	Code	Häufigkeit	Anteil	Code*Anteil
Arbeiter	1	250	0.25	0.25
Angestellte	2	400	0.40	0.80
Sonstige	3	350	0.35	1.05
Summe			1.00	2.10

Der Erwartungswert ist 2.1

Die nächste empirisch vorkommende Codeziffer ist 2. Der KW-Einsetzungswert ist also 2.

### **Eingabe: 5 Mittelwert-Einsetzung II**

Für Kein\_Wert wird eingesetzt:

- a. bei quantitativen Variablen der zum Mittelwert nächste empirisch vorkommende Wert
- b. bei ordinalen Variablen der Median wie bei Kein-Wert-Behandlung 4
- c. bei nominalen Variablen der Erwartungswert wie bei Kein-Wert-Behandlung 4

### **Eingabe: 6 Mittelwert-Einsetzung III**

Für Kein\_Wert wird eingesetzt:

- a. bei quantitativen Variablen der Mittelwert +/- einem normalverteilten Zufallswert mit Mittelwert=0 und Standardabweichung der Variablen.  
Wir könnten auch formulieren: Es wird ein normalverteilter Zufallswert mit Mittelwert und Standardabweichung der Variablen eingesetzt.
- b. bei ordinalen Variablen der Median.

Ist die Variable (was eher ungewöhnlich ist) mit ungleichen Schrittweiten kodiert (z.B. 1, 2, 5, 6, 23), dann wird der Median eingesetzt.

Liegt dieser zwischen zwei empirisch vorkommenden Werten, dann wird der zum Median nächst gelegene empirische Wert verwendet.

Ist die Variable mit gleicher Schrittweite kodiert, dann wird ein Wert X errechnet, der sich ergibt aus Median +/- einem normalverteilten Zufallswert mit Mittelwert=0 und Standardabweichung in der Größe des halben Quartilsabstands der Variablen. Der zu X nächst gelegene empirische Skalenwert wird dann eingesetzt.

Bei quantitativen und bei ordinalen Variablen wird also eine normalverteilte Zufallszahl mit Mittelwert=0 generiert.

Als Standardabweichung wird bei quantitativen Variablen die der jeweiligen Variablen verwendet. Bei ordinalen Variablen wird der halbe Quartilsabstand verwendet.

Betrachten wir ein Beispiel: Die quantitative Variable sei das Lebensalter. Also errechnet für sie einen Mittelwert von 40 und eine Standardabweichung von 20. Dann wird eine normalverteilte Zufallszahl mit Mittelwert=0 und Standardabweichung=20 erzeugt. Nehmen wir an es entsteht der Zufallswert -15.25. Für den fehlenden Wert wird dann eingesetzt  $X = 40 - 15.25 = 24.75$ .

Bei einer ordinalen Variablen wird entsprechend verfahren. Als Standardabweichung für die Generierung der Zufallszahl wird der halbe Quartilsabstand verwendet. Der ermittelte X-Wert wird bei der ordinalen Variablen aber noch nicht als KW-Einsetzungswert verwendet. Es wird nach dem empirisch vorkommenden Wert gesucht, der am dichtesten bei X liegt. Dieser wird als KW-Einsetzungswert verwendet. So wird verhindert, dass KW-Einsetzungswerte entstehen, die empirisch nicht vorkommen.

- c. Bei nominalen Variablen wird der wahrscheinlichste Ausprägungswert eingesetzt. Die Vorgehensweise soll an einem Beispiel gezeigt werden. Die nominale Variable sei der Beruf mit den 3 Ausprägungen Arbeiter, Angestellte, Sonstige. Dabei wurden folgende Häufigkeiten ermittelt.

	Code	Häufigkeit	in %	in % kummuliert
Arbeiter	1	250	25	25
Angestellte	2	400	40	65
Sonstige	3	350	35	100

Dann wird eine gleichverteilte Zufallszahl zwischen 0 und 100 erzeugt.

Liegt sie zwischen  
0 und 25, dann wird für den fehlenden Wert 1 eingesetzt  
25 65 2  
65 100 3

### **Eingabe: 7 Mittelwert-Einsetzung IV**

Für Kein\_Wert wird eingesetzt:

a. bei quantitativen Variablen:

Es wird zunächst ein Wert X errechnet, der sich ergibt aus dem Mittelwert +/- einem normalverteilten Zufallswert mit Mittelwert=0 und der Standardabweichung der Variablen. Dann wird der zu X nächst gelegener empirische Skalenwert für Kein\_Wert eingesetzt. So wird verhindert, dass KW-Einsetzungswerte entstehen, die empirisch nicht vorkommen.

b. bei ordinalen Variablen wie bei Kein-Wert-Behandlung 6

c. bei nominalen Variablen wie bei Kein-Wert-Behandlung 6

Kein-Wert-Behandlung 4 und 5 unterscheiden sich von 6 und 7 dadurch, dass bei 6 und 7 eine Zufallsvariation dem Mittelwert bzw. Median bzw. Erwartungswert hinzugefügt wird.

Die Kein-Wert-Behandlung 4 unterscheidet sich von 5 nur dadurch dass für die quantitativen Variablen ein Mal der Mittelwert und das andere Mal der zum Mittelwert nächste empirisch vorkommende Wert als KW-Einsetzungswert verwendet wird.

### **Warum Zufallswert hinzufügen?**

Es muss noch folgende Frage beantwortet werden: Warum wird der Mittelwert bzw. der Median bei Kein-Wert-Behandlung 6 und 7 durch einen Zufallswert überlagert?

Wird als KW-Einsetzungswert nur der Mittelwert (bzw. der Median) verwendet, dann wird die Varianz der Variablen verringert, weil für Kein-Wert immer derselbe Wert eingesetzt wird.

Werden mit den so erzeugten „vollständigen“ Daten beispielsweise Korrelationen errechnet, dann werden die Signifikanzen dieser Korrelationen überschätzt. Siehe dazu etwa R. J. A. Little/D. B. Rubin (1990, S. 381).

Die Überlagerung durch einen normalverteilten Zufallswert mit der Standardabweichung der Variablen bezweckt also, dass die Varianz der Variablen (fast) unverändert bleibt. Gleiches gilt auch für nominale Variable. Der Erwartungswert der Variablen ist immer derselbe. Dadurch wird die Varianz verringert. Durch den "wahrscheinlichsten" Wert bleibt die Streuung (fast) unverändert.

### Eingabefeld 2: Wenn abhängige Variable Kein-Wert besitzt

1 = wenn die abhängige Variable Kein-Wert besitzt, dann wird der gesamte Datensatz aus der Analyse vollständig ausgeschlossen unabhängig davon welche Kein-Wert-Behandlung oben im ersten Eingabefeld dieser Eingabe-Box gewählt wurde.

0 = die im 1. Eingabefeld gewählte Kein-Wert-Behandlung gilt auch für die abhängige Variable

### Eingabefeld: Startwert für Zufallsgenerator



Die Zufallswerte, die in den oben beschriebenen Kein-Wert-Behandlungen 6 und 7 benötigt werden, erzeugt Almo mit einem "Zufallsgenerator". Wenn die Startzahl nicht verändert wird, dann werden bei einem 2. und jedem weiteren Lauf des Programms Prog45mo immer dieselben Zufallszahlen und damit dieselben KW-Einsetzungswerte erzeugt. Ist dies jedoch nicht erwünscht, dann muss der Benutzer die Startzahl ändern. Verwenden Sie eine 6-stellige ungerade Zahl.

### Eingabefeld: Gemeinsame Fallzahl



Wird als Kein-Wert-Behandlung=1 oder =2, das "paarweise Ausscheiden" erwählt, dann liefert Almo folgende Tabelle - sofern Kein-Wert-Fälle auch tatsächlich auftreten:

Beispiel:

Zahl der Einheiten, die in die Analyse eingegangen sind  
je Zelle der Streuungsmatrix bei "paarweisem Ausscheiden"

	x1	x2	x3
x1	90	81	73
x2	81	90	75
x3	73	75	80

Für das Variablenpaar x1 x2 werden nur die Einheiten, die in beiden Variablen valide Daten besitzen, ausgewertet. Das sind in unserem Beispiel 81.

Für das Variablenpaar x1 x3 werden nur 73 Einheiten ausgewertet und für das Variablenpaar x2 x3 nur 75.

Almo benötigt für den weiteren Rechengang eine einzige  
"gemeinsame Fallzahl"

Diese wird für die Ermittlung der Signifikanzen der Korrelationen gebraucht.

Almo bietet für die "gemeinsame Fallzahl" 4 Alternativen an:

- 0 = die kleinste Fallzahl, aus der die Korrelationen zwischen je 2 Variablen i und k errechnet wurden
- 1 = das harmonisches Mittel aus den Fallzahlen, aus denen die Korrelationen bzw. Kovarianzen zwischen den Variablen errechnet wurden
- 2 = die Zahl der Fälle, die in allen Analysevariablen valide Werte besitzen
- 3 = die Zahl der eingelesenen Fälle

Die vorsichtigste Alternative ist =2.

Für die Signifikanztests im Korrelationsprogramm werden nur so viele Fälle verwendet, wie sie beim "vollständigen Ausscheiden" vorhanden wären.

Die optimistischste Alternative ist =3.

Hier werden alle eingelesenen Fälle für die Signifikanztests verwendet. 1 und 2 liegen dazwischen.

### Vergleich

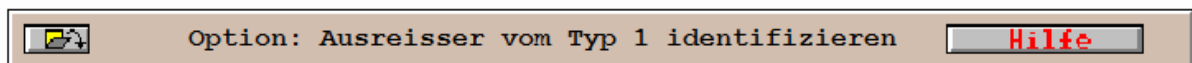
Beim "paarweisen Ausscheiden" werden die Daten optimal "ausgenützt". Der Nachteil dieses Verfahrens ist jedoch, dass die Korrelationen zwischen den Variablen auf unterschiedlichen Häufigkeiten beruhen. Sind die Unterschiede nicht zu groß, dann kann man diesen Nachteil ignorieren.

Das "listenweise Ausscheiden" ist korrekt, die Menge an Daten, die nicht benutzt werden, kann jedoch beträchtlich sein.

### Schätzwerte für fehlende Werte

Almo ermöglicht es mit Prog45mm und Prog45mz Schätzwerte für fehlende Werte einzusetzen. Dabei wird das ALM bei quantitativen und die Logitanalyse bei nominalen Variablen eingesetzt. Siehe die ausführliche Darstellung im Almo-Dokument Nr. 24 „Statistische Datenanalyse I“, Abschnitt P45.7.

### Eingabe-Box: Option: Ausreisser vom Typ 1 identifizieren



Siehe hierzu die ausführliche Darstellung im Almo-Dokument Nr. 23 "Ausreisser entdecken"

Was sind Ausreisser ?

Ausreisser sind Werte, die ausserhalb "valider Grenzen" liegen. Die "validen Grenzen" definiert der Forscher. Anders formuliert: Es gibt keine "objektive", eindeutige Definition, was ein Ausreisser ist. Der Forscher legt fest, was für ihn ein Ausreisser ist. Werden Ausreisser vom Forscher aus der Analyse ausgeschlossen, dann tut er dies, weil er unterstellt, dass diese Daten - obwohl empirisch gewonnen - falsch sind oder er tut dies, weil sie ihm einen Variablen-Zusammenhang seiner Meinung nach verfälschen.

In Almo werden 2 Typen von Ausreissern unterschieden:

Ausreisser vom Typ 1:

- Ein Variablenwert liegt ausserhalb des "validen Wertebereichs" der Variablen. Hier können nochmals 2 Untertypen unterschieden werden
  - a. Schreibfehler
  - b. Extremwerte

Ausreisser vom Typ 2:

Ein Variablenwert liegt ausserhalb der "validen Punktwolke"  
eines mehrdimensionalen Variablen-Zusammenhangs.

#### Zu Typ 1a: **Schreibfehler oder Messfehler als Ausreisser**

Ausreisser entstehen sehr oft dadurch, dass beim Schreiben der Daten Fehler gemacht werden. Beispiel: Anstelle 9 wird versehentlich der Wert 99 geschrieben. Diese Art der Ausreisser sollte der Benutzer durch das Programm Prog03m versuchen zu entdecken. Dieses Programm untersucht, ob Variablenwerte auftreten, die ausserhalb der zulässigen Unter- und Obergrenzen liegen. Diese muss der Benutzer in das Programm eingeben. Wird beispielweise Geschlecht (männlich, weiblich) mit 1 und 2 kodiert, dann liegt der Wert 3 ausserhalb der zulässigen Unter- und Obergrenzen und beruht auf einem Schreibfehler. Prog03m findet man durch Klick auf den Knopf "Verfahren", dann Eintrag "Fehlersuche". Diese Schreibfehler sollte man, bevor man mit der Datenanalyse überhaupt beginnt, bereinigen, d.h. in den Daten selbst korrigieren.

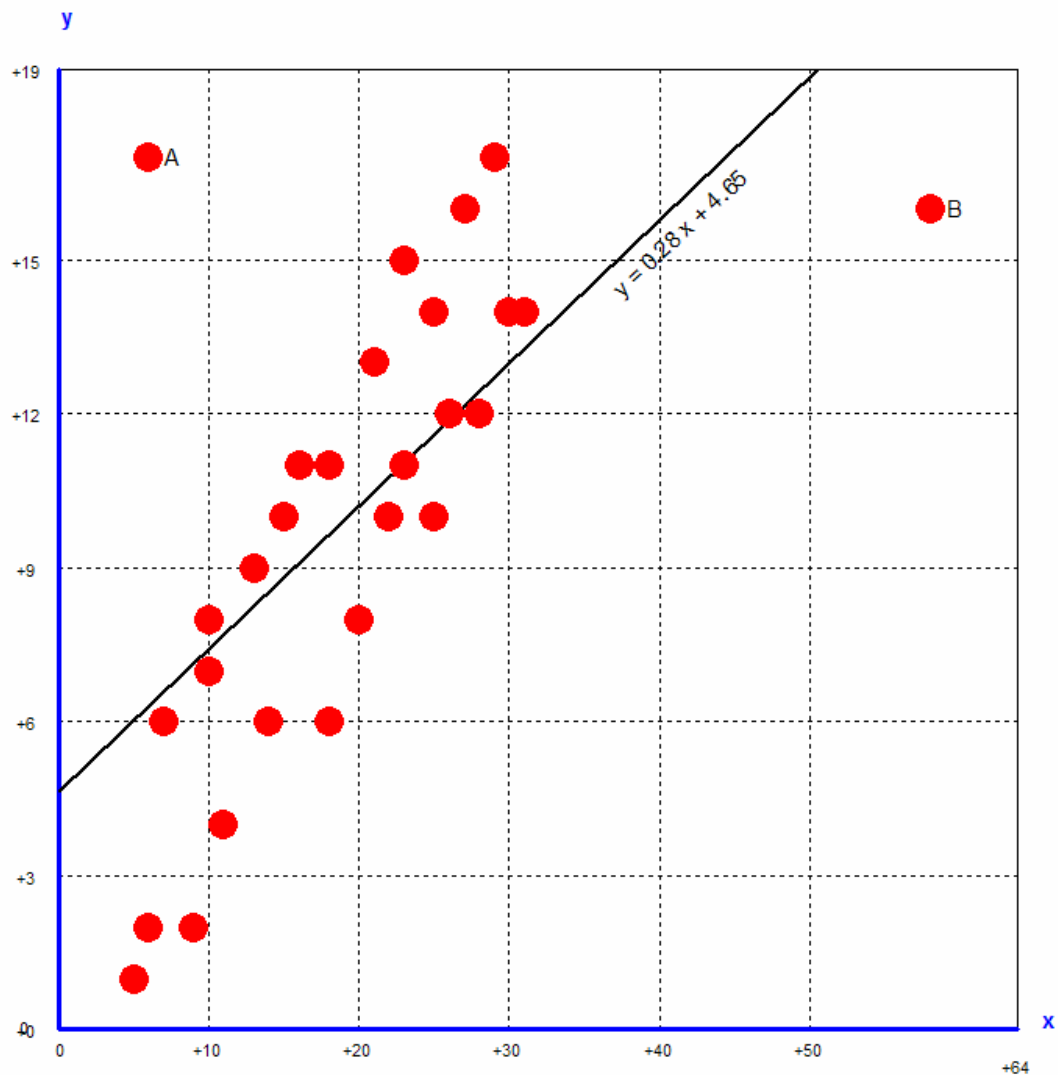
Messfehler können, müssen aber nicht, ausserhalb der zulässigen Wertegrenzen liegen. Liegen Sie innerhalb der Wertegrenzen werden sie in der Regel nicht entdeckt. Am ehesten gelingt es noch, sie als Ausreisser vom Typ 2 zu entdecken, d.h. innerhalb eines mehrdimensionalen Variablen-Zusammenhangs.

#### Zu Typ 1b: **Extremwert als Ausreisser**

Natürlich gibt es auch "echte" Ausreisser, die nicht durch Schreibfehler entstanden sind. Beispiel: Für eine Stichprobe von 1000 Personen wird das Einkommen erhoben. Dabei sind einige wenige Milliardäre in die Stichprobe gelangt. Deren Einkommen liegt ausserhalb des "validen Wertebereichs". Wird nun die Korrelation zwischen Einkommen und beispiesweise Schulbildung ermittelt, so kann der Korrelationskoeffizient durch die Milliardäre dramatisch verändert werden. Hier ist es sinnvoll, die Milliardäre als Ausreisser zu identifizieren und aus der Analyse auszuschliessen. Für diesen Zweck ist diese Optionsbox gedacht, die in vielen Almo-Maskenprogrammen angeboten wird. Natürlich kann Sie auch verwendet werden, um Ausreisser vom Typ 1a, also Schreibfehler und Messfehler, zu finden. Der Benutzer kann auch **Prog05m3** einsetzen. Dies ist ein spezielles Programm, dessen Zweck folgender ist: Die Ausreisser vom Typ 1 sollen identifiziert werden und es soll eine neue "Ausreisserbereinigte" Datei erstellt werden. Diese kann dann für weitere Analysen verwendet werden. Man findet dieses Programm durch Klick auf den Knopf "Verfahren", dann "Ausreisser".

#### Zu Typ 2: **Ausreisser liegt ausserhalb der "validen Punktwolke"**

Betrachten wir ein Beispiel:



Der Zusammenhang zwischen den Variablen x und y wird durch ein Streudiagramm grafisch dargestellt. Die kleinen roten Punkte sind Messpunkte. Die durchgezogene Linie ist die Regressionsgerade.

Der Messpunkt B ist ein Ausreisser vom Typ I. Sein x-Wert liegt weit ausserhalb des validen Wertebereichs von x. Ausreisser vom Typ 1 sind also in der Regel auch Ausreisser vom Typ 2. Anders formuliert, mit der in Almo angebotenen Methode zur Identifizierung von Ausreissern vom Typ 1 (mit der Optionsbox "Ausreisser vom Typ 1 identifizieren") wird auch ein Teil der Ausreisser vom Typ 2 ermittelt - aber eben nur ein Teil.

Der Messpunkt A ist ein Ausreisser vom Typ II. Sein x-Wert und sein y-Wert liegt zwar innerhalb des validen Wertebereichs von x und y. In Bezug auf den Zusammenhang von x und y ist er jedoch ein Ausreisser. Er liegt ausserhalb der "validen Punktwolke xy".

Um Ausreisser vom Typ 2 zu identifizieren muss der Benutzer das Programm Prog20bm verwenden. Man findet dieses Programm durch Klick auf den Knopf "Verfahren", dann "Ausreisser".

Wird die Optionsbox geöffnet, dann sieht man folgendes.

X Loesche wieder diese Box (dann Voreinstellungen wieder gueltig)

**Ausreisser vom Typ 1 identifizieren**  
für ordinale und quantitative Analysevariable

Was sind Ausreisser ? ---> [Hilfe](#)  
Ausreisser vom Typ 1 ---> [Hilfe](#)  
Ausreisser vom Typ 2 ---> [Hilfe](#)

↑↓ 1  
1 = Ausreisser über Quartile  
und Quartilsdifferenz identifizieren  
2 = Ausreisser über Mittelwert  
und Standardabweichung identifizieren  
0 = nicht

↔ 2.5  
Multiplikator für Quartilsdifferenz bzw. Stand.abweichung  
Bestimmt Breite des validen Bereichs [Hilfe](#)

↑↓ 1  
Reaktion [Hilfe](#)  
0 = nichts tun  
1 = Ausreisser in Ergebnisliste nur melden  
2 = melden u. auf KeinWert setzen  
3 = melden u. auf validen Grenzwert setzen  
4 = melden u. auf bereinigte Ober- bzw. Untergrenze setzen  
5 = melden u. ganzen Datensatz ausschliessen

↑↓ 1 [Hilfe](#)  
ordinale Variable in Ausreisser-Suche einbeziehen  
1 = einbeziehen  
0 = ausschliessen  
für ordinale Variable nicht nach Ausreissern auch

Im 1. Eingabefeld der Ausreisser-Optionsbox bietet Almo 2 Methoden an, einen Variablenwert als Ausreisser vom Typ 1 zu identifizieren.

### Methode 1: Ausreisser-Identifizierung über Quartile und Quartilsabstand

Für jede quantitative und ordinale Analysevariable wird das 1. und 3. Quartil und der Quartilsabstand ermittelt.

Wenn ein Variablenwert ausserhalb des "validen Bereichs" von

- 1.  $\text{Quartil} - x \cdot \text{Quartilsabstand}$
- und
- 3.  $\text{Quartil} + x \cdot \text{Quartilsabstand}$

liegt, dann betrachtet Almo ihn als Ausreisser.

Der Benutzer kann  $x$ , den Multiplikator für den Quartilsabstand im 2. Eingabefeld beliebig verändern. Er bestimmt damit die Breite des "validen Bereichs", in dem ein Variablenwert als "valide", also nicht als Ausreisser betrachtet wird.

Der Multiplikator wird im 2. Eingabefeld eingetragen.

Für x sollte ein Wert zwischen ca. 1.5 und 3.5 eingesetzt werden.

Generell gilt (auch für die folgende Methode 2), dass man mit verschiedenen x-Werten experimentieren sollte.

Methode 1 kann nur verwendet werden, wenn die Zahl der diversen Werte, die eine Analysevariable annimmt, kleiner 500 ist.

Anmerkung: Werden Residuen auf Ausreisser untersucht (wie im speziellen Prog20bm), dann sollte Methode 1 nur verwendet werden, wenn die Zahl der Datensätze kleiner 500 ist, da die Werte der Residuen alle voneinander verschieden sein können (so dass die Zahl der diversen Werte gleich der Zahl der Datensätze sein kann).

### **Methode 2: Ausreisser-Identifizierung über Mittelwert und Standardabweichung**

Für jede quantitative und ordinale Analysevariable wird das arithmetische Mittel und die Standardabweichung ermittelt. Beachte: Auch für die ordinalen Variablen wird für die Ausreisser-Identifizierung das arithmetische Mittel und die Standardabweichung berechnet. Die ordinalen Variablen werden also so behandelt, wie wenn sie quantitativ wären.

Wenn ein Variablenwert ausserhalb des "validen Bereichs" von

Mittelwert-x\*Standardabweichung  
und  
Mittelwert+x\*Standardabweichung

liegt, dann betrachtet Almo ihn als Ausreisser.

Der Benutzer kann x, den Multiplikator der Standardabweichung im 2. *Eingabefeld* beliebig verändern. Er bestimmt damit die Breite des "validen Bereichs", in dem ein Variablenwert als "valide", also nicht als Ausreisser betrachtet wird.

Für x sollte ein Wert zwischen ca. 2.5 und 5 eingesetzt werden.

### **Grubbs-Test**

Wird Methode 2 gewählt, dann errechnet Almo den Test nach Grubbs. Dieser setzt voraus, dass die Daten normalverteilt sind - was zuvor durch einen Test auf Normalverteilung überprüft werden muss (z.B. mit Prog04m2 durch einen Chi-Quadrat-Test oder den Kolmogorov-Smirnov-Test).

Beim Grubbs-Test wird zuerst der maximale Datenwert  $X_{max}$  ermittelt. Dieser wird dann mit den von Grubbs entwickelten Formeln daraufhin untersucht, ob er sich in die normalverteilte Datenmenge  $X_1 \dots X_n$  einfügt oder ob er ein Ausreisser ist. Der entdeckte Ausreisser wird eliminiert und dann das Verfahren mit dem nächsten Maximum wiederholt usw. Dieses Verfahren wurde durch einen von Rossner entwickelten Test ergänzt, der versucht, die Zahl der signifikanten Ausreisser zu ermitteln. Siehe dazu in der Literatur-Angabe "NIST-Agency", Abschnitt 1.3.5.17.3.

In Almo werden in einem Datendurchlauf alle ausserhalb des (vom Benutzers definierten) validen Bereichs liegenden Werte durch den Grubbs-Test überprüft. Es wird in folgender Weise verfahren:

Der ausserhalb des validen Bereichs liegender Wert wird standardisiert nach der Formel

$$(1) G_x = (X-M)/s$$

G<sub>x</sub>=standardisierter Wert

X =Rohwert

M =Mittelwert

s =Standardabweichung

X ist ein Ausreisser, wenn G<sub>x</sub> grösser ist als G<sub>z</sub>, wobei

$$(2) G_z = [(n-1)/\sqrt{n}] * [\sqrt{t_q / (n-2+t_q)}]$$

sqrt = Wurzel aus(...)

n = Zahl der Untersuchungsobjekte

t<sub>q</sub> = ist der quadrierte t-Wert für die Signifikanz alpha/2n  
mit n-2 Freiheitsgraden

alpha= ist die vom Benutzer festgelegt Signifikanz

Almo unterstellt, dass der Benutzer mit dem von ihm eingegebenen Multiplikator auch das Signifikanzniveau alpha für den Grubbs-Test festgelegt hat. Hat der Benutzer den Multiplikator z.B. auf 2 (genauer 1.96) gesetzt, dann ist alpha=0.05. Bei einem Multiplikator von 2.5 (genauer 2.58) ist alpha=0.01. Almo ermittelt den (quadrierten) t-Wert für alpha/2n und errechnet gemäß (2) den G<sub>z</sub>-Wert.

Liegt G<sub>x</sub> über G<sub>z</sub>, dann ist der Wert ein Ausreisser aus der normalverteilten Datenmenge X<sub>1</sub>...X<sub>n</sub>

## Reaktion

Im 3. Eingabefeld kann der Benutzer festlegen, wie Almo auf identifizierte Ausreisser reagieren soll. Folgende Möglichkeiten gibt es dabei:

Wird in der Optionsbox im 3. Eingabefeld eine 0 eingesetzt, dann reagiert Almo nicht. Es erfolgt auch keine Warnung. Der Ausreisser wird wie ein valider Wert behandelt.

Wird eine 1 eingesetzt,  
dann bringt Almo die Warnung, dass dieser Wert ein Ausreisser ist.

Wird eine 2 eingesetzt,  
dann bringt Almo die Warnung, dass dieser Wert ein Ausreisser ist und setzt den Variablenwert auf KeinWert. Die Folge davon ist in der Regel, dass dieser Wert für die Analyse nicht berücksichtigt wird.

Wird eine 3 eingesetzt,  
dann bringt Almo die Warnung, dass dieser Wert ein Ausreisser ist und setzt den Variablenwert auf den nächst liegenden validen Grenzwert. Beispiel: Die Grenzwerte des validen Bereichs sind -3 und +10. Hat der Ausreisser z.B. den Wert 15, dann wird er auf +10 gesetzt. Hat der Ausreisser z.B. den Wert -5, dann wird er auf -3 gesetzt. Siehe nachfolgende Erläuterung zum Begriff "valider Grenzwert". Diese Reaktion ist nicht möglich, wenn eine Residuen-Variable auf Ausreisser untersucht wird (wie in Prog20bm).

Wird eine 4 eingesetzt,

dann bringt Almo die Warnung, dass dieser Wert ein Ausreisser ist und setzt den Variablenwert auf die "Ausreisser-bereinigte Unter- bzw. Obergrenze" Beispiel: Die Ausreisser-bereinigte Untergrenze ist 0 und die Ausreisser-bereinigte Obergrenze ist 13.

Hat der Ausreisser z.B. den Wert 15 und liegt damit oberhalb des "oberen validen Grenzwerts", dann wird er auf die "bereinigte Obergrenze", also auf 13 gesetzt. Hat der Ausreisser z.B. den Wert -5 und liegt damit unterhalb des, "unteren validen Grenzwerts", dann wird er auf die "bereinigte Untergrenze", also auf 0 gesetzt. Siehe nachfolgende Erläuterung zum Begriff "Ausreisser-bereinigte Unter- bzw. Obergrenze" Diese Reaktion ist nicht möglich, wenn Residuen auf Ausreisser untersucht werden (wie in Prog20bm).

Wird eine 5 eingesetzt,

dann bringt Almo die Warnung, dass dieser Wert ein Ausreisser ist und schliesst den gesamten Datensatz aus der Analyse aus.

Empfehlung: Wir empfehlen, zunächst eine 1 einzusetzen, sich also die Ausreisser zunächst nur melden zu lassen. Dabei kann der Benutzer verschiedene Werte für den Multiplikator ausprobieren. Erst dann sollte er eine 2 oder 3 oder 4 oder 5 einsetzen. Die klarste Lösung des Ausreisser-Problems entsteht sicherlich durch Reaktion 5 "gesamten Datensatz überspringen"

### Literatur zu Ausreisser

Grubbs, Frank E.: Sample criteria for testing outlying observations,

The Annals of Mathematical Statistics 21(1), 1950, S.27-58

NIST-Agency: Engineering statistics (e-Handbook of Statistical Methods)

Das Handbuch im htm-Format ist im Internet zu finden unter

<http://www.itl.nist.gov/div898/handbook/index.htm>

Das Kapitel über Ausreisser unter

<http://www.itl.nist.gov/div898/handbook/eda/section3/eda35h.htm>

### Eingabe-Box: Untersuchungseinheiten gewichten

Siehe P0.8.

### Eingabe-Box: Option: Streuungsmatrix



Die Voreinstellung ist "Quadratsumme". Gemeint ist "Abweichungsquadratsumme". D.h. wenn der Benutzer die Optionsbox ungeöffnet lässt, dann wendet Almo den Kalkül des Allgemeinen Linearen Modells auf die Matrix der Abweichungsquadrate an.

Wird die Optionsbox geöffnet, dann sieht man folgendes:



Der Kalkül des allgemeinen linearen Modells kann auf unterschiedliche Streuungsmatrizen angewendet werden.

Folgende Streuungsmatrizen können analysiert werden	die analysierten Streuungen sind	Dabei entsteht folgender Regress.koeff./Effekt
Korrelation	Varianzen/Kovarianzen *	standardisiert
Quasi_Korrelation !	Varianzen/Kovarianzen *	standardisiert
Kovarianz	Varianzen/Kovarianzen	nicht standardisiert
Quadratsumme	Abweichungsquadrate	nicht standardisiert
Kreuzprodukt	Abweichungsquadrate **	nicht standardisiert
d_Kreuzprodukt	Produkte/Kreuzprodukte ***	nicht standardisiert

! siehe dazu auch das Almo-Dokument Nr. 25 "Statistische Datenanalyse II", Abschnitt P45.12.4.2 und P20.8.5.3.1

\* standardisierter Variabler

\*\* und teilweise Produkte/Kreuzprodukte (siehe unten)

\*\*\* durch n dividiert

Vorzugsweise sollten Sie "Quadratsumme" verwenden. Nur ausnahmsweise verwenden sollten Sie "Kreuzprodukt" und "d\_Kreuzprodukt".

"d\_Kreuzprodukt" ist die durchschnittliche Kreuzprodukte-Matrix. Sie entsteht dadurch, dass die Kreuzprodukte-Matrix mit n (=der Zahl der Fälle) dividiert wird.

**BEACHTE:** Bei der Verwendung von "Kreuzprodukt" und "d\_Kreuzprodukt" gibt es noch ein weiteres Problem: Die von Almo ermittelte (1) Gesamtstreuung, (2) die durch alle unabhängigen Variablen erklärte Streuung, (3) die durch die quantitativen/ordinalen Variablen insgesamt erklärte Streuung sind nicht Abweichungsquadratsummen sondern Summen quadrierter Rohwerte bzw. deren Kreuzprodukte. Die Korrelationskoeffizienten, F-Werte und Signifikanzen dieser 3 Streuungen sind falsch. Almo teilt dies dem Benutzer auch mit.

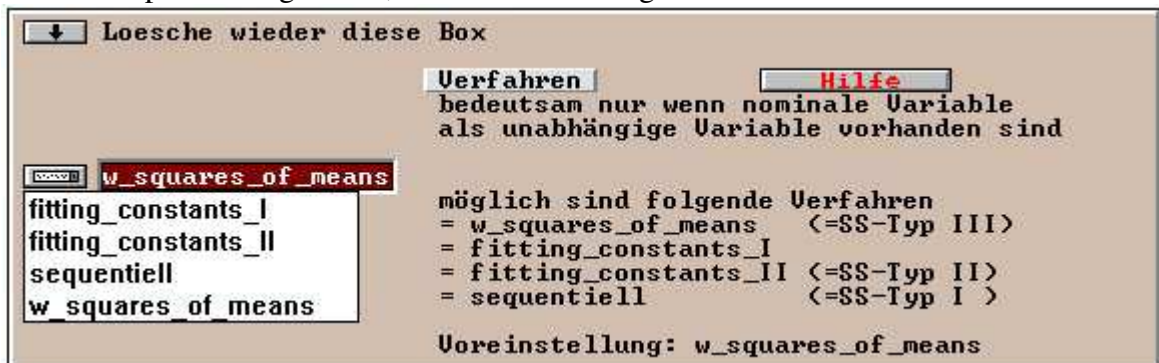
"Kreuzprodukt" und "d\_Kreuzprodukt" sollten nur als *zusätzliche 2. Analyse* eingesetzt werden, wenn für die Konstante nicht nur deren Wert gebraucht wird, sondern auch noch deren erklärte Streuung, Standardfehler und Signifikanz p.

### Eingabe-Box: Option: Verfahren



Die Voreinstellung ist "weighted squares of means". D.h. wenn der Benutzer die Optionsbox ungeöffnet lässt, dann rechnet Almo diese Verfahren.

Wird die Optionsbox geöffnet, dann sieht man folgendes:



4 Verfahren stehen zur Auswahl (siehe die ausführliche Darstellung in P20.7):

### ***w\_squares\_of\_means***

Bezeichnung in SAS und SPSS: SS Typ III

Siehe Abschnitt P20.7.3. Alle Effekte (Haupteffekte und Interaktionseffekte) werden gegenseitig auspartiiert. Unabhängige quantitative Variable (=Kovariate) können in beliebiger Zahl in die Analyse eingeschlossen werden. Sie werden aus der Gesamtmenge der nominalen Variablen auspartiiert.

Ist nur eine einzige unabhängige nominale Variable vorhanden (und eventuell noch Kovariate), dann sollte der Benutzer das Verfahren der fitting constants I wählen. Almo schaltet nicht selbsttätig auf dieses Verfahren um, bringt jedoch eine Warnung, in der diese Umschaltung empfohlen wird.

In allen anderen Fällen ist *w\_squares\_of\_means* das empfehlenswerte Verfahren !!

### ***sequentiell***

Bezeichnung in SAS und SPSS: SS Typ I

Siehe Abschnitt P20.7.2. Die Variable werden hierarchisch angeordnet und auspartiiert. Beispiel bei 3 unabhängigen nominalen Variablen A, B, C ist die Reihenfolge:

A B C AB AC BC ABC

Jede Variable besteht aus einer Gruppe von Dummies. Diese Dummy-Gruppen werden dann hierarchisch auspartiiert. Unabhängige quantitative Variable (=Kovariate) können in beliebiger Zahl in die Analyse eingeschlossen werden. Sie werden aus der Gesamtmenge der nominalen Variablen auspartiiert.

Das Verfahren sollte nur angewendet werden, wenn eine kausale Reihenfolge unter den unabhängigen Variablen unterstellt wird.

Sonderprogramme für sequentielles Verfahren: Für 2, 3 und 4 nominale Variable (und beliebig viele Kovariate) sind noch die drei Programm-Masken ProgSq\_2 bis ProgSq\_4 in Almo enthalten. Siehe dazu die ausführlichen Erläuterungen in Abschnitt P20.7.2 und P20.7.2.3. Man findet die Sonderprogramme nach Klick auf den Knopf "Verfahren/Allgemeines lineares Programm" oder „alle Progs“ am Oberrand des Almo-Fensters.

### ***fitting\_constants\_I***

Kein Äquivalent in SAS und SPSS

Siehe Abschnitt P20.7.1. Die Variable werden zuerst zu Gruppen zusammengefasst.

Beispiel: Bei 3 unabhängigen nominalen Variablen A, B, C werden folgende Gruppen gebildet

Gruppe 1:	A B C	(die nominalen Variablen)
Gruppe 2:	AB AC BC	(die 2-er Interaktionen)
Gruppe 3:	ABC	(die 3-er Interaktionen)

Jede Gruppe besteht aus den Dummies der betreffenden Variablen. Die Gruppen werden dann zuerst hierarchisch auspartiiert und dann innerhalb der Gruppe gegenseitig auspartiiert. Unabhängige quantitative Variable (=Kovariate) können in beliebiger Zahl in die Analyse eingeschlossen werden. Sie werden aus der Gesamtmenge der nominalen Variablen und deren Interaktionen auspartiiert.

Sind (bei ungleichen Zellenbesetzungen) mehr als zwei unabhängige nominale Variable A, B und ihre Interaktionen vorhanden, dann sind die Interaktionseffekte nicht eindeutig ermittelbar. Für die Interaktionseffekte treten dann "wechselnde Werte" auf (siehe Abschnitt P20.6.5.1)

Bei ungleichen Zellenbesetzungen sollte das Verfahren nur für Analysen ohne Interaktionen verwendet werden. Kovariate dürfen in beliebiger Zahl vorhanden sein.

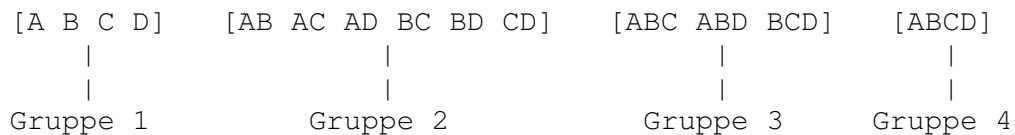
Sonderprogramme für fitting constants I: Für 2, 3, 4 und 5 nominale Variable (und beliebig viele Kovariate) sind noch die vier Programm-Masken ProgFI\_2 bis ProgFI\_5 in Almo enthalten. Siehe dazu die ausführlichen Erläuterungen in Abschnitt P20.7.1 und P20.7.2.3. Man findet die Sonderprogramme nach Klick auf den Knopf "Verfahren/ Allgemeines lineares Programm" oder „alle Progs“ am Oberrand des Almo-Fensters.

### *fitting\_constants\_II*

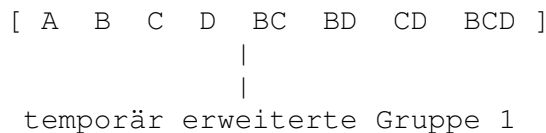
Bezeichnung in SAS und SPSS: SS Typ II

Siehe Abschnitt P20.7.1.1.

Betrachten wir ein Beispiel mit 4 unabhängigen nominalen Variablen A, B, C, D. Die Variable werden zuerst wie bei fitting constants I zu Gruppen zusammengefasst.



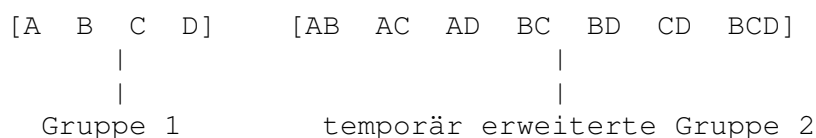
Um die durch die Variable i (die zur Gruppe G gehört) erklärte Streuung zu finden werden aus den Gruppen, die hinter der Gruppe G stehen zu der i gehört noch jene Variable in G eingefügt, an denen i nicht beteiligt ist. Um beispielsweise die durch A erklärte Streuung zu finden, wird die Gruppe 1 temporär in folgender Weise erweitert



A ist nicht beteiligt an BC BD CD BCD aus den nachfolgenden Gruppen 2,3,4. Also werden diese in die Gruppe 1 temporär aufgenommen. Die 4. Gruppe enthält nur die Interaktion ABCD, die sich aus allen einzelnen Variablen zusammensetzt.

Innerhalb dieser temporär erweiterten Gruppe 1 wird nun A an die anderen Variablen "angepasst". Deutlicher formuliert: Die anderen Variablen in Gruppe 1 werden aus A auspartiielliert.

Um beispielsweise die durch die Interaktion AB erklärte Streuung zu finden wird folgende temporär erweiterte Gruppe 2 gebildet



An der 3-er Interaktionen BCD aus den nachfolgenden Gruppen ist AB nicht beteiligt. Sie wird in die Gruppe 2 aufgenommen.

Die Variablen der hierarchisch übergeordneten Gruppe 1 plus die anderen Variablen der temporär erweiterten Gruppe 2 werden aus A auspartiiert. Für die so entstandenen Partial-Dummies von A werden dann die Streuungen errechnet, die A in der abhängigen Variablen Y erklärt.

Unabhängige quantitative Variable (=Kovariate) können in beliebiger Zahl in die Analyse eingeschlossen werden. Sie werden jeweils aus der Gruppe, mit der gerade gerechnet wird, auspartiiert. Wenn z.B. die erklärte Streuung von A ermittelt wird, dann werden die Kovariaten aus der Gruppe [ A B C D BC BD CD BCD ] auspartiiert.

Sonderprogramme für fitting constants II: Für 2, 3, 4 und 5 nominale Variable (und beliebig viele Kovariate) sind noch die vier Programm-Masken ProgFII2 bis ProgFII5 in Almo enthalten. Siehe dazu die ausführlichen Erläuterungen in Abschnitt P20.7.1.1 und P20.7.2.3. Man findet die Sonderprogramme nach Klick auf den Knopf "Verfahren/ Allgemeines lineares Programm" oder „alle Progs“ am Oberrand des Almo-Fensters. Die Programme verwenden spezielle Interaktions-variable und die „Partial“-Anweisung aus der Almo-Programmiersprache. Siehe dazu die ausführliche Darstellung im 2. Teil dieses Handbuchs, Abschnitt P20.14 und P20.15.

Almo ermittelt mit dem Standard-Maskenprogramm Prog20mo bei fitting constants II nur die erklärten Streuungen und ihre Signifikanzen, aber keine Effekte. Werden die Sonderprogramme ProgFII2 bis ProgFII5 eingesetzt, dann erhält man zusätzlich die Haupteffekte und die von diesen abgeleiteten Koeffizienten. Siehe nachfolgend.

fitting constants I ist identisch mit fitting constants II,

- wenn keine Interaktionen in das Modell aufgenommen werden
- wenn nur die beiden Faktoren A und B und ihre Interaktion AB, aber keine Kovariaten sich im Modell befinden. Wird fitting constants mit den Sonderprogrammen ProgFI\_2 bis ProgFI\_5 gerechnet, dann sind die beiden auch in diesem Falle gleich. Siehe die ausführliche Erläuterung dazu in Abschnitt P20.7.1 und P20.7.1.1.

In diesen beiden Fällen schaltet Almo automatisch um auf fitting constants I, wenn der Benutzer fitting constants II angegeben hat.

#### BEACHTE:

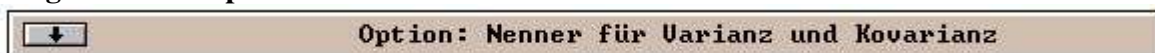
Bei gleichen Zellenhäufigkeiten erbringen alle 4 Verfahren das gleiche Ergebnis. Geben Sie in diesem Falle am besten "w\_squares\_of\_means" als Verfahren an.

Der SS Typ IV aus SAS und SPSS ist in Almo nicht enthalten. Siehe dazu unsere "Anmerkungen zu SS Typ IV" in Abschnitt P20.7.4.1

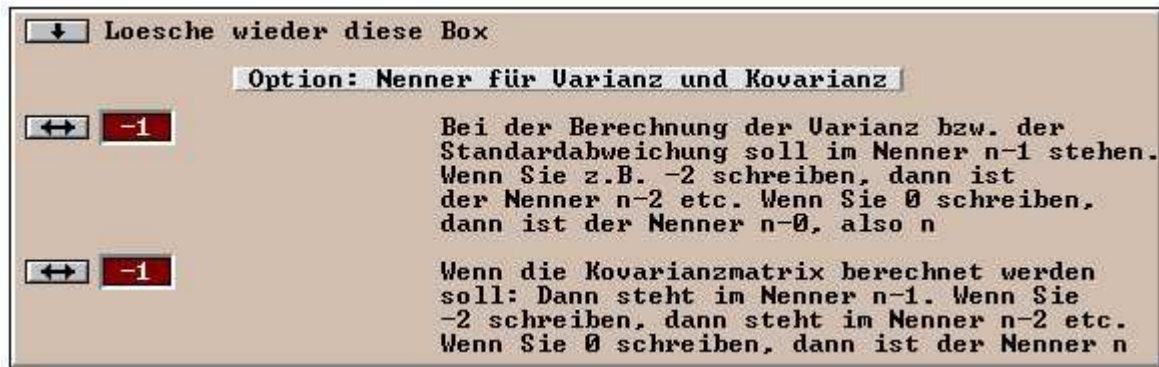
#### EMPFEHLUNG:

Rechnen Sie immer mit „weighted squares of means“; wollen Sie aber doch mit „fitting constants I“ rechnen, dann nur ohne Interaktionen oder mit Interaktionen, wenn ein 2-faktorielles Modell vom Typ [A B AB] vorliegt und keine Kovariaten vorhanden sind.

#### Eingabe-Box: Option: Nenner für Varianz und Kovarianz



Optionsbox geöffnet:



Almo gibt die Standardabweichung und die Kovarianzmatrix als deskriptive Statistik aus und setzt dabei im Nenner  $n$  (die Zahl der Fälle) ein. Wenn Sie wollen, dann können Sie von  $n$  eine Zahl abziehen. Dies hat keinerlei Einfluss auf die Ergebnisse des allgemeinen linearen Modells. Wenn die Kovarianzmatrix als Streuungsmatrix eingesetzt wurde, dann ändern sich die Werte der Kovarianzmatrix, der Fehlerstreuung und der erklärten Streuungen – jedoch in einer solchen Relation, dass dieselben F-Werte, p-Werte, Korrelationskoeffizienten, Effekte etc. entstehen.

### Eingabe-Box: Option: Behandlung eventueller Multikollinearität



Optionsbox geöffnet:



Almo überprüft die zu analysierende Streuungsmatrix auf Multikollinearität. Dabei eliminiert es eine Variable  $i$ , wenn ihr Diagonalglied aus der Cholesky-Matrix kleiner ist als  $0.0001 \cdot SS(i)$ .

$SS(i)$  ist das Diagonalglied  $ii$  der Variablen  $i$  aus der Streuungsmatrix

Wurde mit der Korrelationsmatrix gerechnet, dann ist  $SS(i) = 1.0$ . Der Wert  $0.0001$  ist voreingestellt. Dieser Schwellenwert ist sehr klein. Er wird eine vollständige lineare Abhängigkeit in den meisten Fällen identifizieren können. Er ist jedoch zu klein um eine "Beinahe"-Multikollinearität zu entdecken.

Der Schwellenwert kann in der Eingabe-Box des Maskenprogramms verändert werden.

Wenn Sie einen größeren Wert, z.B.  $0.005$  oder sogar  $0.01$  einsetzen, dann wird eine "Beinahe"-Multikollinearität früher entdeckt, die Variable  $i$  also früher eliminiert

Almo gibt eine Warnung, wenn das Cholesky-Diagonalglied kleiner ist als  $0.09 \cdot SS(i)$ . Der Wert  $0.09$  ist voreingestellt. Er kann in der Eingabe-Box verändert werden.

## Eingabe-Box: Option: Spezielle Programm-Optionen

↓
Option: Spezielle Programm-Optionen

Optionsbox geöffnet:

↓
Option: Spezielle Programm-Optionen

↑↓ 1

Regressionskoeffizienten der Dummies  
der unabhang. nominalen Variablen ausgeben  
(zusatzlich zu den Effekten)  
1 = ja  
0 = nein  
\*\* nur wenn nominale Variable als unabhangige  
\*\* Variable vorhanden sind und volle Ausgabe

---

↑↓ 1

1 = die Matrix der in den abhangigen  
Variablen erklarten Streuungen ausgeben  
und  
zusatzlich Hotelling-Lawleys Spur rechnen  
0 = nicht  
\*\* nur bei multivariater Analyse:

### 1. Eingabefeld:

Almo errechnet zunachst Regressionskoeffizienten fur die Dummies der unabhangigen nominalen Variablen. Danach werden diese in „Effekte“ umgerechnet. Siehe dazu Abschnitt P20.6.5.2 und insbesondere P20.7.5. Wollen Sie die eher schlecht interpretierbaren Regressionskoeffizienten ausgegeben haben, dann setzen Sie 1 in das

### 2. Eingabefeld: Siehe auch Abschnitt P20.7.5.

Wenn Sie zwei oder mehrere abhangige Variable haben oder eine nominale abhangige Variable – also im Fall der multivariaten Analyse – dann gibt Almo standardmaig als Teststatistik das Wilk’sche Lambda aus.

Wird das 2. Eingabefeld auf 1 gesetzt, dann wird zusatzlich ausgegeben:

1. die zwei Matrizen der in den abhangigen Variablen verbleibenden Fehlerstreuung und der erklarten Streuung
2. Pillais Spur
3. Hotelling-Lawley Spur

Die Almo-Ausgabe ist folgende:

```
Matrix der in den abhaengigen Variablen
verbleibenden Fehlerstreuung
          V5      V7      V20
          Leistung Einkommen Bewertung
-----|-----
V5 | 176.3919   -3.4768   -5.2957
V7 |  -3.4768   35.5135   -0.5360
V20 | -5.2957   -0.5360  118.6119
```

```
Matrix der in den abhaengigen Variablen
erklarten Streuung
          V5      V7      V20
          Leistung Einkommen Bewertung
-----|-----
V5 | 45.8048   -2.5068  -46.2945
V7 | -2.5068    3.4046   1.4868
V20 | -46.2945   1.4868  307.1586
```

```

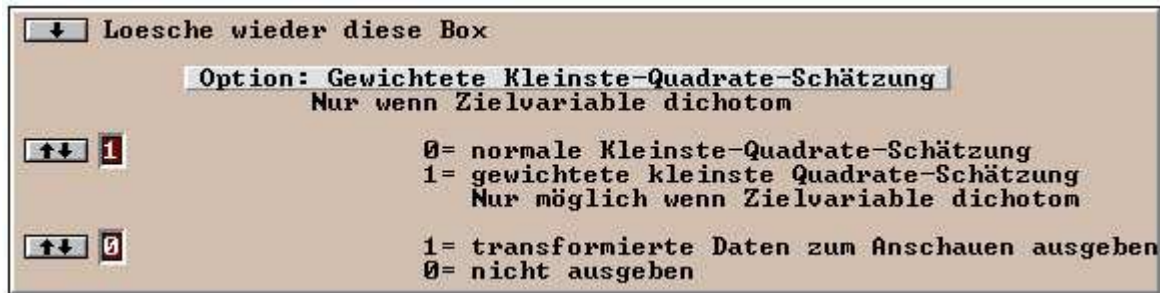
-----
Pillais Spur                                0.987686
F-Wert f. erklarte Streuung                 3.716216
Freiheitsgrade Nenner = 21
      Zaehler= 159
Signifikanz: p                              0.000017
Signifikanz: (1-p)*100                     99.998283 %
Teststaerke von F                          0.999971
-----
Hotelling-Lawley Spur                       2.923832
F-Wert f. erklarte Streuung                 6.915095
Freiheitsgrade Nenner = 21
      Zaehler= 149
Signifikanz: p                              0.000005
Signifikanz: (1-p)*100                     99.999497 %
Teststaerke von F                          1.000000

```

### Eingabe-Box: Option: Gewichtete Kleinste-Quadrate-Schätzung



Optionsbox geöffnet:



Ist die abhängige Variable nominal-dichotom, dann besteht modellbedingte Varianzheterogenität. Diese kann durch die Methode der "gewichteten Kleinste-Quadrate" beseitigt werden. Geben Sie zu diesem Zweck im 1. Eingabefeld „1“ ein.

Bei den gewichteten Kleinsten-Quadraten werden die Daten in bestimmter Weise transformiert. Die transformierten Daten können ausgegeben werden, wenn im 2. Eingabefeld eine „1“ eingesetzt wird.

Das Allgemeine Lineare Modell (ALM) kann also auch auf den Fall angewendet werden, dass die Zielvariable dichotom-nominal oder sogar polytom-nominal ist. Seine Anwendung auf dichotome und polytome Zielvariable schafft jedoch einige Probleme. Diese sind:

3. Das Modell kann Wahrscheinlichkeiten prognostizieren, die außerhalb des Bereichs 0 bis 1 liegen. Es kann beispielsweise prognostizieren, dass die Wahrscheinlichkeit für den Kauf eines Produkts (mit den beiden Ausprägungen „ja“ und „nein“)  $p=1.08$  (also 108%) ist.
4. Es besteht modellbedingte Varianz-Heteroskedastizität mit der Folge, dass die Schätzer für die Parameter der ursächlichen Variablen zwar unverzerrt und konsistent, aber nicht mehr effizient sind. Das bedeutet, dass die Standardfehler der Effekte und Regressionskoeffizienten der ursächlichen Variablen nicht minimal sind, mit der Folge, dass die Signifikanzüberprüfung mit t- und F-Test nicht korrekt ist. Siehe dazu die ausführliche Darstellung bei Aldrich/Nelson (1984, S. 12ff) und Urban (1993, S. 17ff), sowie Urban (1982, Abschnitt 3.1 und 3.1.1).

Auf das 1. Problem werden wir im Almo-Dokument Nr. 25 „Statistische Datenanalyse II“, Abschnitt P45.15.1.7 ausführlich eingehen. Wir wollen hier aber schon vorwegnehmen, dass die Reproduzierungs- bzw. Prognosefähigkeit des Modells dadurch nicht beeinträchtigt wird.

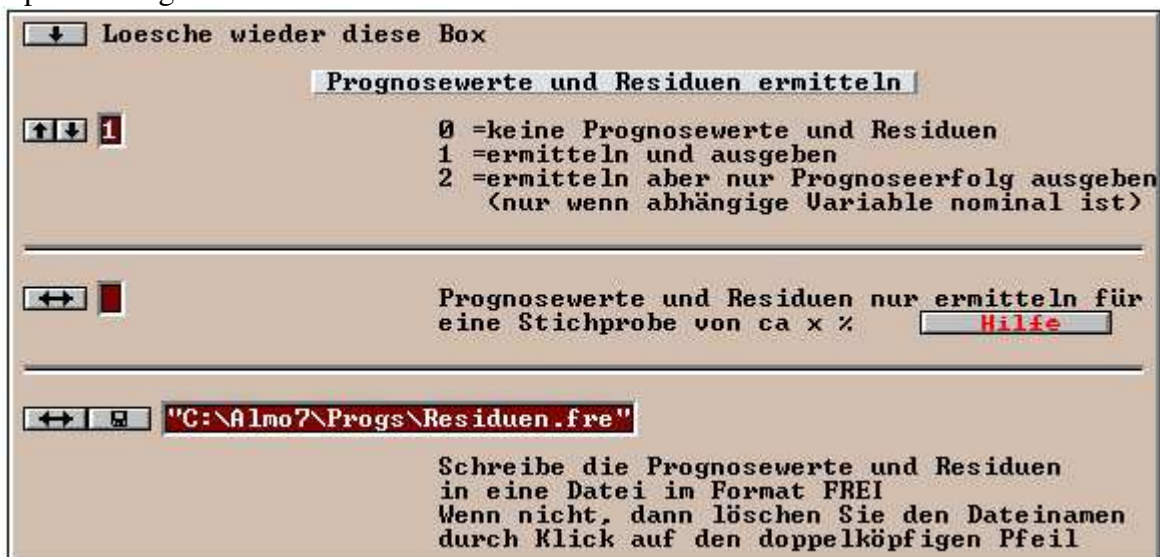
Das 2. Problem kann im Rahmen des ALM durch die „gewichtete Kleinste-Quadrate-Schätzung“ gelöst werden. Dieses Verfahren geht auf Goldberger (1964) zurück. Man nimmt dabei allerdings in Kauf, dass die Reproduzierungs-fähigkeit dieser Modellvarianten schlechter ist. D.h. die Fähigkeit des Modells, die Untersuchungseinheiten aus der Stichprobe der ersten oder der zweiten Ausprägung der dichotomen Zielvariablen richtig zuzuweisen, ist schlechter als bei der normalen Kleinste-Quadrate-Lösung. Wir bieten im Programm Prog20mo und Prog45mf die gewichtete Kleinste-Quadrate-Schätzung für nominal-dichotome Zielvariable als Option an. In Prog45gw bieten wir ein Programm an, das eine gewichtete Kleinste-Quadrate-Schätzung für nominal-polytome Zielvariable leistet. Siehe dazu die Darstellung im Almo-Dokument Nr. 25 „Statistische Datenanalyse II“, Abschnitt P45.15.2.2.

Als beste Alternative zum ALM für dichotome und polytome Zielvariable wird das Logit-Modell empfohlen. Siehe Prog22m und Prog45m9. Es ist in Almo-Dokument Nr. 9 „Logitanalyse“ und im Almo-Dokument Nr. 25 „Statistische Datenanalyse II“, Abschnitt P45.7.6 dargestellt. Das Logit-Modell leidet zwar nicht unter diesen beiden Problemen, seine Ergebnisse sind aber nicht so einsichtig interpretierbar wie die des ALM. Dies gilt insbesondere für polytome Zielvariable.

### Eingabe-Box: Option: Prognosewerte und Residuen



Optionsbox geöffnet:



*Eingabefeld 1:* Wird hier „1“ eingegeben, dann werden für alle Untersuchungseinheiten die „Prognosewerte“ und „Residuen“ ermittelt. Residuen sind die Differenz zwischen Prognosewert und wirklichem Wert.

Ist die abhängige Variable dichotom dann wird auch ausgezählt, wie oft die Prognose richtig bzw. falsch war. Wird „2“ eingegeben, dann wird nur der Prognoseerfolg ausgezählt. Diese Eingabe ist empfehlenswert, wenn die Datei sehr viele Datensätze umfasst.

*Eingabefeld 2:* Wenn die Datei sehr groß ist, dann ist es empfehlenswert, nur eine

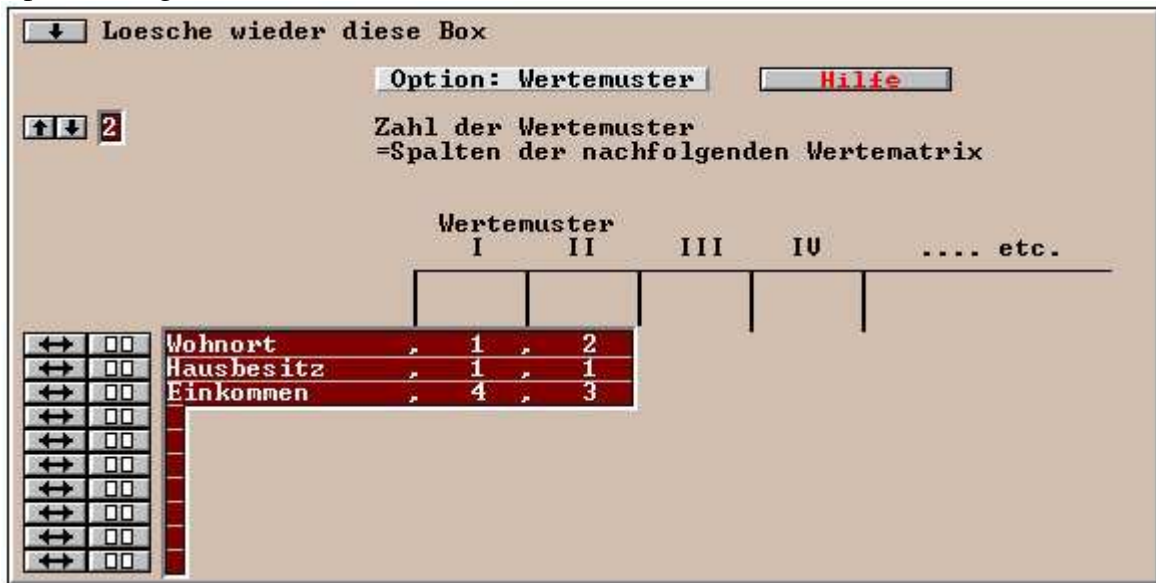
Zufallsstichprobe von beispielsweise 25% einzubeziehen. In diesem Fall wird in das Eingabefeld 25 geschrieben. Sollen alle Datensätze einbezogen werden, dann lässt man das Eingabefeld leer.

*Eingabefeld 3:* Prognosewerte und Residuen können in eine Datei gespeichert werden. Geben Sie einen Dateinamen an.

**Eingabe-Box: Wertemuster**



Optionsbox geöffnet:



Der Benutzer kann sich von Almo prognostizieren lassen, welchen Wert in der abhängigen Variablen (=der Zielvariablen) eine Untersuchungseinheit besitzen wird, die bestimmte Werte in einer oder mehreren oder allen unabhängigen Variablen hat.

Wir greifen das Beispiel aus dem Almo-Dokument Nr. 25 „Statistische Datenanalyse II“ auf. Siehe dort Abschnitt P45.15.12, Erläuterung zu Eingabe-Box 16. Die Zielvariable ist die Rückzahlung eines Kredits mit den Ausprägungen „Nein“ und „Ja“. Die Zielvariable ist also nominal-dichotom. Wertemuster können aber auch für quantitative oder ordinale Zielvariable ermittelt werden. Die unabhängigen Variablen sind Wohnort (Stadt, Land), Hausbesitz (kein Haus, hat Haus), Einkommen sowie weitere quantitative unabhängige Variable.

Ist die Zielvariable die "Rückzahlung: Nein,Ja" dann kann sich der Benutzer beispielsweise von Almo berechnen lassen, welche Wahrscheinlichkeit der "Rückzahlung:Nein" bzw. der "Rückzahlung:Ja" eine Person hat, die ein Haus besitzt und ein Einkommen von 4 Einheiten bezieht.

Wir sprechen hier vom "Wertemuster" einer Person. In unserem Beispiel haben wir 2 Wertemuster. D.h. wir haben 2 Personen, von denen wir die Werte für einige ursächliche Variable angeben und dann von Almo die Wahrscheinlichkeit der "Rückzahlung:Nein" bzw. der "Rückzahlung:Ja" geliefert haben wollen.

Betrachten wir unser Beispiel genauer:

Die abhängige Variable ist:

Rückzahlung eines Kredits: nein, ja

Die unabhängigen nominalen Variablen sind:

Wohnort: Stadt (=1) Land (=2)  
Hausbesitz: kein Haus (=1) hat Haus (=2)  
Produkt: Kleidung (=1) Möbel (=2) Technik (=3)

Die unabhängigen quantitativen Variablen sind:

Einkommen  
Rückrate  
Laufzeit

Wir wollen nun die Wahrscheinlichkeit der Rückzahlung prognostizieren für

1. Städter, die kein Haus besitzen und ein Einkommen von 4 Einheiten besitzen
2. Landbewohner, die kein Haus besitzen und ein Einkommen von 3 besitzen

Wir geben als Zahl der Wertemuster = 2 an und schreiben in die Eingabefelder der Wertemustermatrix

↓ Loesche wieder diese Box

Option: Wertemuster Hilfe

↑↓ 2

Zahl der Wertemuster  
=Spalten der nachfolgenden Wertematrix

	Wertemuster				
	I	II	III	IV	.... etc.
Geschlecht	1	2	Alter	48	58
Einkommen	7200	3500	Bildung	5	3

Zuerst wird also der Variablenname (oder -nummer) geschrieben, dann der Wert des 1. Wertemusters, dann der des 2. Es können beliebig viele Wertemuster angefordert werden. Die Schreibweise muss nicht so schön formatiert sein, wie in obiger Grafik. Wichtig ist, dass die Beistriche als Trennzeichen nicht vergessen werden. Am Zeilenende kein Beistrich!

**WICHTIG:**

Als Trennzeichen innerhalb eines Eingabefeldes muss ein Beistrich geschrieben werden, auch hinter dem Variablennamen (bzw. Variablennummer). Am Zeilenende wird kein Beistrich geschrieben.

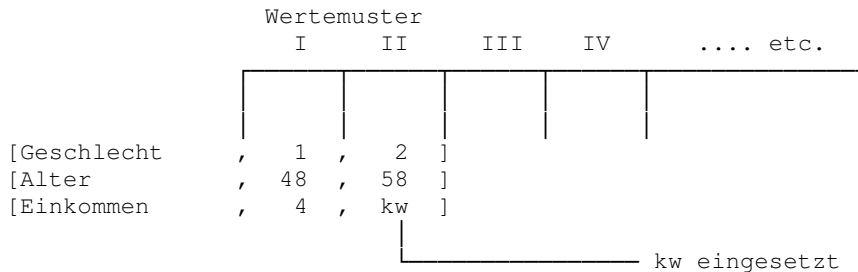
**BEACHTE:**

Almo setzt automatisch für die anderen unabhängigen Variablen, die der Benutzer nicht für die Wertemuster verwendet, deren Mittelwerte ein.

Das gilt auch für die nicht verwendeten nominalen Variablen. In unserem Beispiel wird die

nominale Variable "Produkt" nicht verwendet. Also löst intern diese Variable in Dummies auf und setzt für diese Dummies deren Mittelwert ein. Der Mittelwert einer Dummy-Variablen ist gleich dem Anteilswert der Probanden, die sich in der betreffenden Ausprägung befinden.

Möglich ist auch folgende Eingabe:



Sie wollen beim 1. Wertemuster das Einkommen mit einer Höhe von 4 einbeziehen - beim 2. Wertemuster jedoch nicht. Dann schreiben Sie beim 2. Wertemuster

KeinWert oder kurz: kw

Also setzt dann beim 2. Wertemuster für das Einkommen dessen Mittelwert ein.

Hinweis:

Wenn sie mehr Variable in das Wertemuster einbeziehen wollen als Zeilen vorhanden sind, dann gibt es folgende Möglichkeit, die wir an einem Beispiel illustrieren wollen.



Sie schreiben in ein Eingabefeld 2 oder sogar mehrere Variable mit ihren Werten. Der vorgegebene Rahmen braucht nicht eingehalten zu werden.

**BEACHTTE:** Alle Zahlenwerte und Variablennamen werden durch Beistrich getrennt. Am Schluss des Eingabefeldes wird kein Beistrich geschrieben. Die Rahmen und die Überschrift darüber dienen nur der "Schönheit". Sie haben keine Bedeutung für Almo.

**Eingabe-Box: Option: "Aussehen" der auszugebenden Tabelle bzw. Matrix**  
 Siehe P0.9.

## Eingabe-Box: Grafik-Optionen



Optionsbox geöffnet:

Loesche wieder diese Box (dann Voreinstellungen wieder gueltig)

**Grafik-Optionen**

Almo zeichnet

- Liniendiagramme
- Balkendiagramme
- Flussdiagramme
- lineare Funktionen

---

**Almo**      Almo      = Almo-Grafik ausgeben  
 0            = keine Grafik

---

Gruppierungsvariable   
 für lineare Funktionen

0 = für jede Ausprägung der Grupp.variablen  
 eine eigene Grafik zeichnen  
 1 = alle Ausprägung der Grupp.variablen  
 in einer gemeinsamen Grafik zeichnen

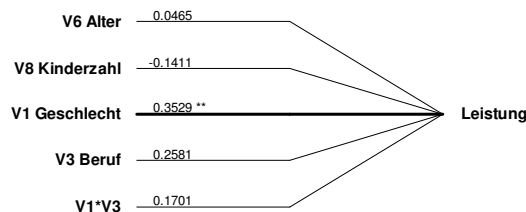
---

1 = Almo-Grafiken in Ergebnisliste einsetzen  
 0 = nicht

Almo zeichnet standardmäßig, auch ohne dass diese Optionsbox geöffnet wird Flussdiagramme und lineare Funktionen.

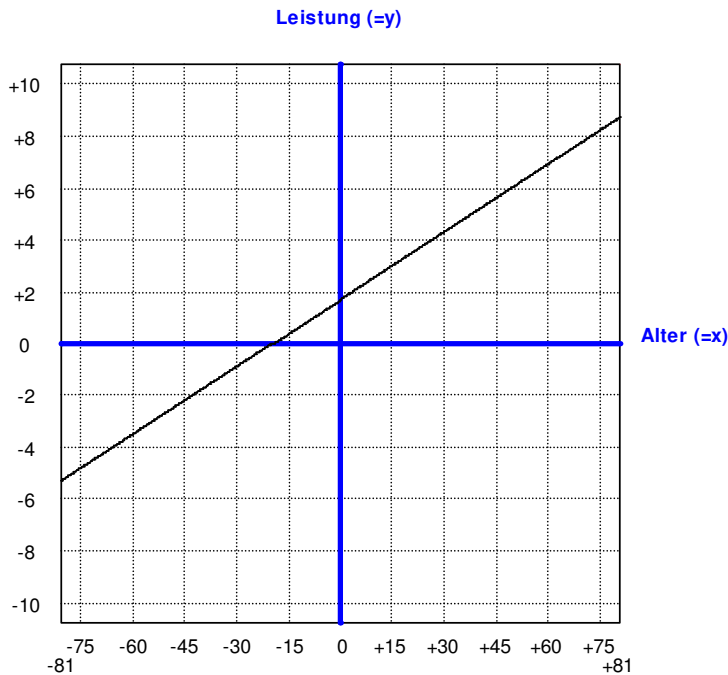
Beispielsweise wird folgendes Flussdiagramm der partiellen Korrelationskoeffizienten der unabhängigen hinsichtlich der abhängigen Variablen gezeichnet

Partielle Korrelationskoeffizienten



Es wird dann noch folgende lineare Funktion für  
 abhaengige Variable: V5 Leistung  
 unabhengige Variable: V6 Alter  
 gezeichnet:

Lineare Funktion  
 $Y = 0.087093 * X + 1.7071$



Betrachten wir ein anderes Beispiel:  
Die abhängige Variable sei "Preis für ein gekauftes Auto"  
(kurz: Autopreis)

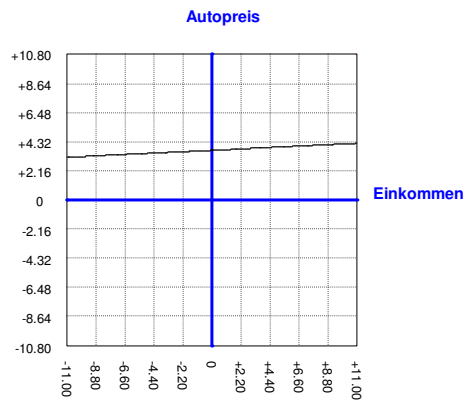
unabhängige nominale Variable  
Geschlecht: männlich, weiblich  
Beruf: Angestellter, Beamter

unabhängige quantitative Variable  
Einkommen  
Schulden

Almo zeichnet nun je eine lineare Funktion für die 2 unabhängigen quantitativen Variablen. Dabei wird die unabhängige quantitative Variable an die x-Achse geschrieben und die abhängige Variable "Autopreis" an die y-Achse.

Zuerst wird die lineare Funktion für "Einkommen" (x-Achse) und "Autopreis" (y-Achse) gezeichnet.

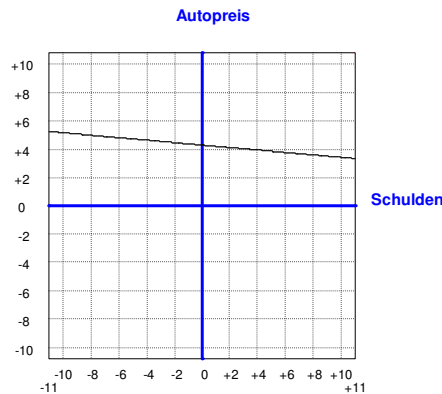
Lineare Funktion  
 $Y = 0.047093 * X + 3.7071$



Man erkennt, dass der Autopreis steigt, wenn das Einkommen größer wird.

Dann wird die lineare Funktion für "Schulden" (x-Achse) und "Autopreis" (y-Achse) gezeichnet.

Lineare Funktion  
 $Y = -0.086619 * X + 4.3042$



Je größer die Schulden, umso kleiner der Preis, den man für das Auto bezahlen will.

Die jeweils anderen unabhängigen quantitativen Variablen werden dabei auf ihren Mittelwert gesetzt. Auch die Dummies der unabhängigen nominalen Variablen werden auf ihre Mittelwert gesetzt. Dieser entspricht dem Anteilswert der Ausprägungen. In der Überschrift zur Grafik werden diese anderen Variablen in der Konstanten (oben z.B. 4.3042) zusammengefasst.

Die Gleichung für unser Beispiel ist nachfolgend in (1) angegeben, die Gleichung, die Almo zeichnet, in (2)

$$(1) A = \beta_1 * E + \beta_2 * S + \beta_3 * G_m + \beta_4 * G_w + \beta_5 * B_a + \beta_6 * B_b + const$$

$$(2) A = \beta_1 * E + \beta_2 * MS + \beta_3 * MG_m + \beta_4 * MG_w + \beta_5 * MB_a + \beta_6 * MB_b + const$$

A = Autopreis  
 E = Einkommen  
 S = Schulden

MS = Mittelwert aus Schulden  
 MG<sub>m</sub>, MG<sub>w</sub> = Anteilswert für Geschlecht: männlich bzw. weiblich  
 MB<sub>a</sub>, MB<sub>b</sub> = Anteilswert für Beruf: Angestellter bzw. Beamter

$\beta_1$  =Regressionskoeffizient für Einkommen  
 $\beta_2$  =Regressionskoeffizient für Schulden

$\beta_3$  =Effekt für Geschlecht: männlich  
 $\beta_4$  =Effekt für Geschlecht: weiblich  
 $\beta_5$  =Effekt für Beruf: Angestellter  
 $\beta_6$  =Effekt für Beruf: Beamter

const =Konstante

Für die unabhängige quantitative Variable "Schulden" ist in (2) deren Mittelwert eingesetzt worden. Ebenso für die Dummies der unabhängigen nominalen Variablen. Das entspricht der Einsetzung einer "Durchschnittsperson" in der betreffenden Variablen.

Wir können also etwas verkürzt formulieren:

In der Almo-Grafik wird für die "Durchschnittsperson" der lineare Zusammenhang zwischen Einkommen und "Autopreis" gezeichnet.

Im Titel der Almo-Graphik wird Gleichung (2) angegeben. Dabei wird der Gleichungsteil

$$\beta_2 * MS + \beta_3 * MGm + \beta_4 * MGw + \beta_5 * MBa + \beta_6 * MBb + const$$

aus obiger Gleichung in einem Zahlenwert zusammengefasst

Dabei ist

$$\beta_3 * MGm + \beta_4 * MGw = 0$$

$$\beta_5 * MBa + \beta_6 * MBb = 0$$

Die Summe von "Effekt mal Anteilswert" der Dummies einer unabhängigen nominalen Variablen ist =0.

Dies gilt unabhängig davon, ob mit dem Verfahren der "weighted squares of means" gerechnet wurde oder dem Verfahren der "fitting constants I".

### **Gruppierungsvariable**

Nun besteht die Möglichkeit eine oder mehrere Gruppierungsvariable anzugeben.

*Beachte:* Als Gruppierungsvariable können nur Variable verwendet werden, die als unabhängige nominale Variable angegeben wurden.

Es wird beispielsweise das "Geschlecht" als Gruppierungsvariable angegeben. Almo zeichnet dann die lineare Funktionen (so wie oben beschrieben) für die beiden Ausprägungen des Geschlechts. Es werden also folgende Kurven gezeichnet:

1. Einkommen (x-Achse) mit "Autopreis" (y-Achse) für die Männer.
2. Einkommen (x-Achse) mit "Autopreis" (y-Achse) für die Frauen.
3. Schulden (x-Achse) mit "Autopreis" (y-Achse) für die Männer.
4. Schulden (x-Achse) mit "Autopreis" (y-Achse) für die Frauen.

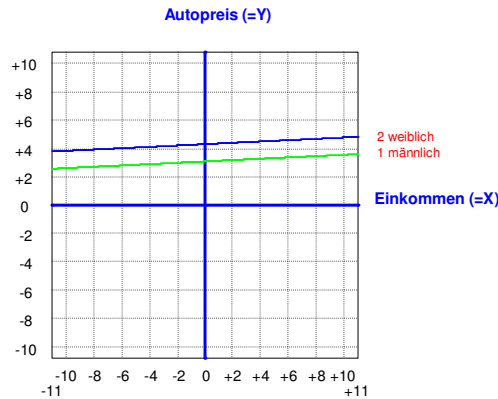
Die jeweils anderen ursächlichen Variablen sind dabei auf ihren Mittelwert gesetzt. Bei (1.)

wird also

Schulden    Beruf:Angestellter    Beruf:Arbeiter  
auf den Mittelwert bzw. Anteilswert gesetzt.

Almo zeichnet bei (1.) folgende Funktion, wobei der Benutzer wählen kann, ob er Männer und Frauen in einer gemeinsamen Grafik oder getrennt in 2 Grafiken darstellen will.

Lineare Funktion  
1: grüne Linie: männlich  $Y = 0.047093 * X + 3.1009$   
2: blaue Linie: weiblich  $Y = 0.047093 * X + 4.3132$



Die Gleichung für diese beiden Funktionen lautet:

$$(3) A = \beta_1 * E + \beta_2 * MS + \beta_3 * 1 + \beta_4 * 0 + \beta_5 * MBa + \beta_6 * MBb + const$$

Für die Männer wird der Effekt  $\beta_3$  ("männlich") mit 1 und  $\beta_4$  ("weiblich") mit 0 multipliziert. Für die Frauen umgekehrt. In der Grafik werden die Mittelwerte der anderen Variablen in der Konstanten zusammengefasst, z.B. für die Männer mit 3.1009.

Bei (2.)

$$(4) A = \beta_1 * E + \beta_2 * MS + \beta_3 * 0 + \beta_4 * 1 + \beta_5 * MBa + \beta_6 * MBb + const$$

Der Effekt "männlich" ist 0 und "weiblich" 1

A    =Autopreis  
E    =Einkommen  
S    =Schulden

MS    =Mittelwert aus Schulden  
MBa,MBb =Anteilswert für Beruf: Angestellter bzw. Beamter

$\beta_1$     =Regressionskoeffizient für Einkommen  
 $\beta_2$     =Regressionskoeffizient für Schulden

$\beta_3$     =Effekt für Geschlecht: männlich  
 $\beta_4$     =Effekt für Geschlecht: weiblich  
 $\beta_5$     =Effekt für Beruf: Angestellter  
 $\beta_6$     =Effekt für Beruf: Beamter

const =Konstante

Dabei gilt (siehe oben):  $\beta_5 * MBa + \beta_6 * MBb = 0$   
Dieser Ausdruck kann also aus der Gleichung gestrichen werden.

Würde sich nicht die unabhängige Variable "Schulden" im Modell befinden, dann würden die beiden Funktionen sehr einfach lauten:

$$(3a) A = \beta_1 * E + \beta_3 * 1 + \beta_4 * 0 + \text{const}$$

$$(4a) A = \beta_1 * E + \beta_3 * 0 + \beta_4 * 1 + \text{const}$$

Die Steigung der Geraden ist gleich  $\beta_1$ , dem Regressionskoeffizienten des Einkommens. Zur Konstanten const kommt dann noch der Effekt des jeweiligen Geschlechts hinzu.

### **Kombinierte Gruppierungsvariable**

Zwei oder mehrere Gruppierungsvariable können auch kombiniert werden. Dazu wird die MIT-Anweisung aus der Almo-Programmiersprache verwendet. Siehe dazu Handbuch Teil 2 „Almo-Programmiersprache“. Betrachten wir ein Beispiel:

In die Eingabe-Box "Grafik-Optionen" schreiben Sie in das Eingabefeld für die Gruppierungsvariable

```
Geschlecht MIT Beruf
```

**BEACHTEN:** Es können maximal 4 Variable durch MIT kombiniert werden.

Almo erzeugt dann folgende Kombinationen in folgender Reihenfolge

```
männlich mit Angestellter
männlich mit Beamter
weiblich mit Angestellter
weiblich mit Beamter
```

Die jeweils hintere Variable "läuft" über ihre Ausprägungen. Für jede Kombination wird eine Funktionsgrafik gezeichnet.

### **Mehrere Gruppierungsvariable**

Es können mehrere Gruppierungsvariable (durch Beistrich getrennt) angegeben werden. Beispiel:

```
Geschlecht, Beruf
```

Almo zeichnet dann beispielsweise für das Einkommen 4 Kurven, eine für die Männer, eine für die Frauen, eine für die Angestellten und eine für die Beamten. Ebenso werden 4 Kurven für die Variable "Schulden" gezeichnet.

Mehrere einzelne Gruppierungsvariable und mehrere durch MIT kombinierte Gruppierungsvariable können angegeben werden. Beispiel:

```
Geschlecht, Beruf, Geschlecht MIT Beruf
```

### **Eingabe-Box: Option: Die errechnete Streuungsmatrix in eine Datei geben**



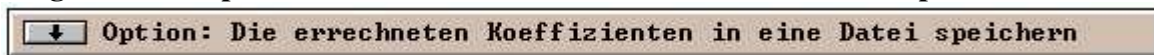
Optionsbox geöffnet:



Wenn Sie einen Dateinamen angeben, dann speichert Almo die Streuungsmatrix. Wird die Eingabe-Box "Streuungsmatrix" nicht geöffnet, dann wird standardmäßig die Matrix der Abweichungs-Quadratsummen gespeichert. Im anderen Falle wird die vom Benutzer selektierte Streuungsmatrix (z.B. die Korrelationsmatrix) gespeichert.

Almo speichert die Matrix in der Form, wie in der Erläuterung zu Prog20m6, oder in Handbuch Teil 2 „Almo-Programmiersprache“, Abschnitt 43 dargestellt. Die gespeicherte Matrix kann dann z.B. in Prog20m6 eingelesen werden.

#### Eingabe-Box: Option: Die errechneten Koeffizienten in eine Datei speichern

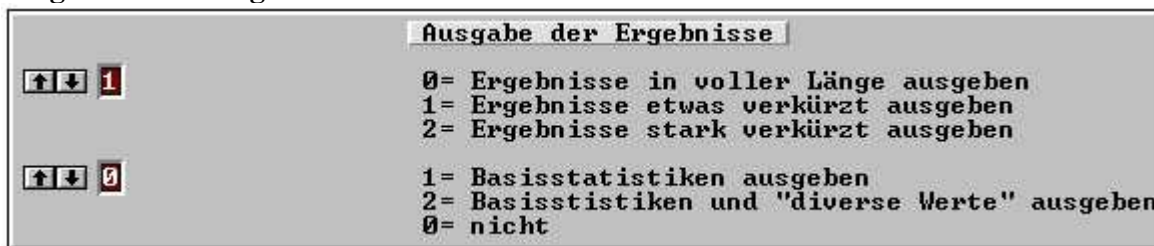


Optionsbox geöffnet:



Wenn Sie einen Dateinamen angeben, dann speichert Almo die Regressionskoeffizienten der unabhängigen quantitativen/ordinalen Variablen und die Effekte (siehe P20.6.5) der unabhängigen nominalen Variablen hinsichtlich der abhängigen Variablen in eine Datei.

#### Eingabe-Box: Ausgabe

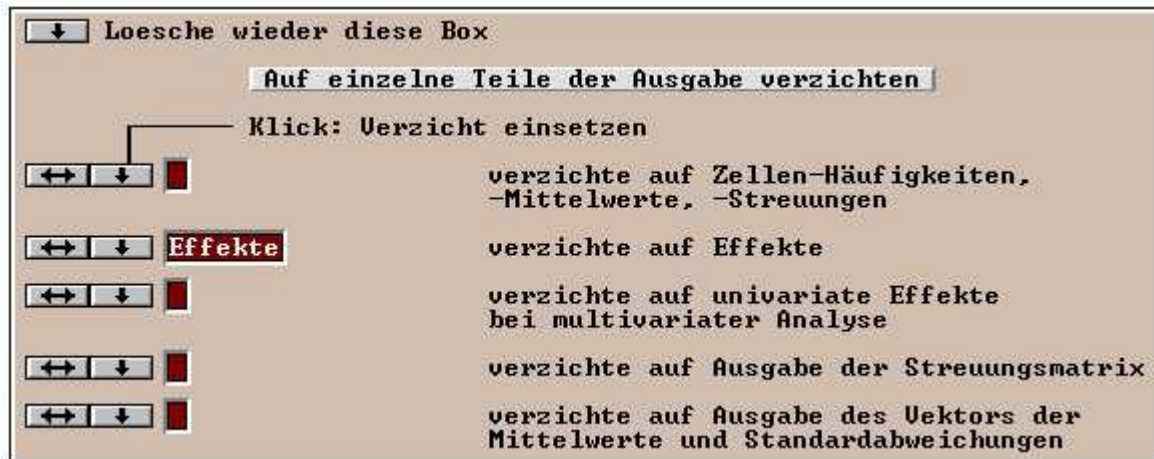


Die Ausgabe der Ergebnisse kann sehr umfangreich sein. Der Benutzer kann hier einschränken.

#### Eingabe-Box: Option: Auf einzelne Teile der Ausgabe verzichten



Optionsbox geöffnet:



Der Benutzer kann gezielt auf einzelne Teile der Ergebnis-Ausgabe verzichten. Das geschieht dadurch, dass der Benutzer auf den Knopf mit dem nach unten gerichteten Pfeil klickt. Siehe die ausführliche Darstellung in P20.8.6.7.

## P20.9 Ausgabe der Ergebnisse

Für die folgenden Beispiele verwenden wir eine Datenmatrix die unter Namen "C:\Almo\Testdat\Testdat.fre" in Almo enthalten ist.

### P20.9.1 Ausgabe bei Varianzanalyse

Wir wollen ein Beispiel betrachten, für das wir obige Datenmatrix verwenden.

Die Leistung in einem Test soll durch das Geschlecht und den Beruf der Versuchsperson erklärt werden. Die abhängige Variable ist V5 (= Testleistung). Als unabhängige nominale Variable verwenden wir aus unserer Datenmatrix die Variable 1 (= Geschlecht) mit 2 Ausprägungen und die Variable 3 (= Beruf) mit 3 Ausprägungen.

V5 bezeichnen wir in den folgenden Gleichungen auch mit  $y$ , V1 mit A bzw. a und V3 mit B bzw. b.

Wir verwenden alle 61 Datensätze und rechnen mit ungleichen Zellenhäufigkeiten. Wir wollen mit Rohwerten rechnen und dabei von der Abweichungs-Quadratsummen-Matrix ausgehen.

In den Maskenprogrammen Prog20mx (Abschnitt P20.8.0) und Prog20mo (Abschnitt P20.8.1) sehen die Eingabe-Boxen für die abhängige Variable und die unabhängigen Variablen so aus:

**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

**Leistung**

---

ordinale abhängige Zielvariable **Hilfe**

---

nominale abhängige Zielvariable **Hilfe**

**Analyse-Variable: Unabhängige Variable** **Hilfe**

nominale unabhängige Variable **Hilfe**

**Geschlecht, Beruf**

**2**

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

∅ =keine Interaktionen bilden

**Geschlecht, Beruf**

paarweise Vergleiche (Kontraste) für die nominalen unabhängige Variablen rechnen

---

quantitative unabhängige Variable **Hilfe**

---

ordinale unabhängige Variable **Hilfe**

In der vorletzten Eingabe-Box „Ausgabe der Ergebnisse“ haben wir auf „Ergebnisse in voller Länge ausgeben“ eingestellt. Der Benutzer kann auch 2 eigens für die Varianzanalyse entwickelte Maskenprogramme verwenden. Klicken Sie auf Verfahren/Varianzanalyse und selektieren dann Prog20my oder Prog20md.

### Ergebnisse

Im Folgenden werden die Ergebnisse, wie ALMO sie ausgibt, abgedruckt. Wir wollen die wichtigsten Ergebnisse herausgreifen:

Die Gleichung für die Analyse, die wir rechnen, lautet:

$$y' = \tau^*t + \alpha_1^*a_1 + \alpha_2^*a_2 + \beta_1^*b_1 + \beta_2^*b_2 + \beta_3^*b_3 + \alpha\beta_{11}^*\alpha_1\beta_1 + \alpha\beta_{12}^*a_1b_2 + \alpha\beta_{13}^*a_1b_3 + \alpha\beta_{21}^*a_2b_1 + \alpha\beta_{22}^*a_2b_2 + \alpha\beta_{23}^*a_2b_3$$

vereinfacht geschrieben:

$$y' = \tau + \alpha_i + \beta_k + \alpha\beta_{ik}$$

$y'$  = der vom Modell prognostizierte Wert der abhängigen Variablen

$\tau^*t$  = Konstanteneffekt (Mittelwert von  $y$ )

$\alpha_1, \alpha_2$  = Haupteffekte von A

$\beta_1, \beta_2, \beta_3$  = Haupteffekte von B

$\alpha\beta_{11}, \alpha\beta_{12}, \dots, \alpha\beta_{23}$  = Interaktionseffekte von AB

$a_1, a_2, b_1, b_2, b_3$  = 0-1 kodierte Nominaldummies von A und B

$a_1b_1, a_1b_2, \dots, a_2b_3$  = 0-1 kodierte multiplikative Dummies von AB.

Besitzt eine Untersuchungseinheit beispielsweise die Merkmalskombination A2 B2, dann besitzt sie in der Dummy-Variablen  $a_2 = 1, b_2 = 1, a_2b_2 = 1$  und in allen anderen Dummies den Wert 0.

ALMO gibt für die Effekte folgende Werte aus (bei Verfahren = weighted squares of means bzw. SS-Typ III):

$\tau^*t$	=	6.1278
$\alpha_1$	=	0.6961
$\alpha_2$	=	-0.6961
$\beta_1$	=	-0.5029
$\beta_2$	=	-0.2441
$\beta_3$	=	0.7471
$\alpha\beta_{11}$	=	0.0539
$\alpha\beta_{12}$	=	-0.3577
$\alpha\beta_{13}$	=	0.3039
$\alpha\beta_{21}$	=	-0.0539
$\alpha\beta_{22}$	=	0.3577
$\alpha\beta_{23}$	=	-0.3039

Der prognostizierte Wert  $y'$  ist im Falle der Varianzanalyse gleich dem Zellenmittelwert. Für eine Person, die sich beispielsweise in der Zelle  $A_2B_2$  befindet, ergibt sich folgender Wert:

$$y' = 6.1278 - 0.6961 - 0.2441 + 0.3577 = 5.5455$$

Dies ist genau der Zellenmittelwert von V5 in der Zelle  $A_2B_2$ , wie man in der Ausgabe in der Tabelle "Zellenmittelwerte der abhängigen Variablen" überprüfen kann.

Unter Residuen versteht man die Differenz zwischen prognostiziertem Wert  $y'$  und tatsächlichem  $y$ -Wert. Die Berechnung dieser Residuen in ALMO zeigen wir in Abschnitt P20.9.3.1.

Die Gesamtstreuung der abhängigen Variablen ist 222.1967 (Quadratsumme). Hätten wir mit standardisierten Daten gerechnet, also mit der Option "MATRIX=KORRELATION", dann ist die Gesamtstreuung immer 1.0.

Die durch alle diese Effekte erklärte Streuung ist 41.8583.

Daraus ergibt sich ein multipler Korrelationskoeffizient von 0.434, der mit  $(1-p)*100 = 96.262\%$  signifikant ist.

Die nominale Variable V1 (= Geschlecht) erklärt eine Streuung von 26.986, was einen (partiellen) Korrelationskoeffizienten von 0.3608 ergibt, der mit  $(1-p)*100=99.433\%$  signifikant ist.

B erklärt eine Streuung von 14.476 und besitzt einen (partiellen) Korrelationskoeffizienten von 0.273, der mit 88.231% signifikant ist.

Die Interaktionsvariable V1\*V3 erklärt eine Streuung von 4.7430 und besitzt einen (partiellen) Korrelationskoeffizienten von 0.1601, der mit 50.596% signifikant ist.

## Ergebnisse aus ALMO

Haeufigkeiten je Auspraegung der nominalen Variablen

```
-----
```

V1 Geschlecht	
V1-1 männlich	34
V1-2 weiblich	27
V3 Beruf	
V3-1 Arbeiter	16
V3-2 Angestellter	29
V3-3 Selbständiger	16

Ergebnisse aus ALMO

-----

Fuer Analyse ausgewaehlte Variable

```
V1    Geschlecht  männlich weiblich
V3    Beruf      Arbeiter Angestellter Selbständiger
V5    Testleistung
```

```
V1    wird bezeichnet mit A
      die Auspraegungen (bzw.Dummies) mit A1 A2
```

```
V3    wird bezeichnet mit B
      die Auspraegungen (bzw.Dummies) mit B1 B2 B3
```

Zellenmittelwerte der abhaengigen Variablen

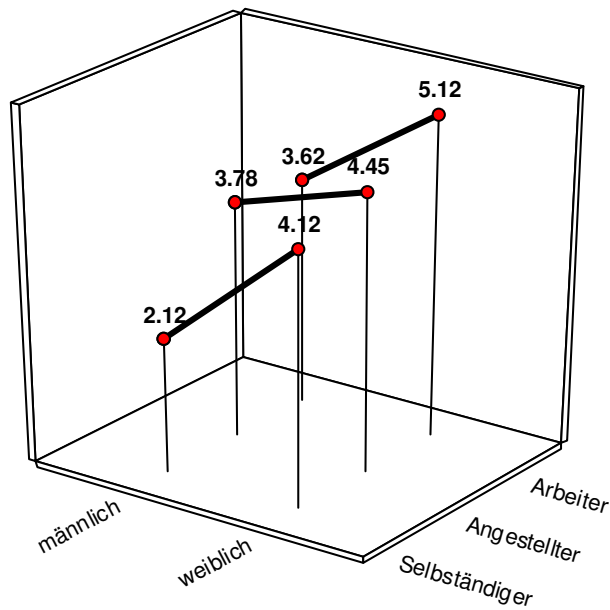
Geschlec Beruf	Leistung
männlich Arbeiter	3.6250
Angestel	3.7778
Selbstän	2.1250
weiblich Arbeiter	5.1250
Angestel	4.4545
Selbstän	4.1250
Gesamtmittel	0.8570

Mittelwert aus Zellenmittelwerten

3.8721

Almo zeichnet folgende Grafik der Zellenmittelwert:

Mittelwerte von Leistung



Erläuterung: Im Almo-Grafik-Editor kann diese Grafik auch 2-dimensional dargestellt werden und in vielfältiger Weise bearbeitet werden.

Streuung (Standardabweichung) der abhaengigen Variablen je Zelle  
 Standardabweichung ist mit n  
 nicht mit n-1 dividiert)

Geschlec Beruf		Leistung
männlich	Arbeiter	8
	Angestel	18
	Selbstän	8
weiblich	Arbeiter	8
	Angestel	11
	Selbstän	8
		-

=====  
 die Zellenmittelwerte und Streuungen der abhaengigen Variablen  
 beruhen auf folgenden Besetzungszahlen (Zellenhaeufigkeiten)

		Leistung
		v5
A1	B1	8
A1	B2	18
A1	B3	8
A2	B1	8
A2	B2	11
A2	B3	8

=====  
 Zahl der insgesamt eingelesenen Einheiten 61  
 Zahl der in die Analyse einbezogenen Einheiten 61  
 =====

\*\*\*\*\* MITTEILUNG

In einigen der nachfolgenden Matrizen ist die redundante letzte Dummy-Variable der (unabhaeng.) nominalen Variablen nicht enthalten

**↓ Zeige Ausgabe:** Zahl der Einheiten, die in die Analyse eingegangen sind je Zelle der Streuungsmatrix

\*\*\* **Erläuterung:** Wenn der Benutzer auf den Knopf klickt, dann zeigt Almo die nachfolgend abgebildete Tabelle.

Nachdem die Tabelle inkludiert wurde erscheint dann folgender Knopf:

**↓ Verberge Ausgabe:** Zahl der Einheiten, die in die Analyse eingegangen sind je Zelle der Streuungsmatrix

Durch Klick auf diesen Knopf löscht Almo die Tabelle wieder.

Soll die Tabelle jedoch, nachdem sie inkludiert wurde, fester Bestandteil der Ergebnisliste bleiben, dann speichern Sie die Ergebnisliste (durch Klick auf den Speichern-Knopf).

Zahl der Einheiten, die in die Analyse eingegangen sind je Zelle der Streuungsmatrix

	A1	B1	B2	A1B1	A1B2	Testleis V5
A1	61	61	61	61	61	61
B1	61	61	61	61	61	61
B2	61	61	61	61	61	61
A1 B1	61	61	61	61	61	61
A1 B2	61	61	61	61	61	61
Testleistung V5	61	61	61	61	61	61

\*\*\* **Erläuterung:** Wenn KEIN\_WERT-Fälle auftreten, können die Besetzungszahlen von Zelle zu Zelle verschieden sein. Der Benutzer kann auf diese Art und Weise feststellen, aus wievielen Untersuchungseinheiten die Streuung in der jeweiligen Zelle berechnet wurde. Siehe dazu P20.9.1.2.

Die Zahl der Einheiten, je Zelle der Streuungsmatrix ist gleich  
Es sind keine Kein-Wert-Faelle aufgetreten

als Fallzahl fuer Signifikanztest wird verwendet: 61

\*\*\* **Erläuterung:** Sind die Besetzungszahlen in obiger Matrix verschieden, dann wird aus ihnen das harmonische Mittel errechnet. Dieses wird für die verschiedenen Signifikanztests als N verwendet. Wenn dem Benutzer ein so gebildetes N suspekt ist, dann kann er eine andere Vorgehensweise wählen. Siehe die ausführliche Darstellung in P20.8.1.1, Eingabe-Box „spezielle Kein-Wert-Behandlung“, insbesondere den Abschnitt „gemeinsame Fallzahl“.

**↓ Zeige Ausgabe:** Standardabweichungen und Mittelwerte der Variablen (bei nominalen Variablen: der nicht-redundanten Dummies)

\*\*\* **Erläuterung:** Wenn der Benutzer auf den Knopf klickt, dann zeigt Almo die nachfolgend abgebildete Tabelle.

Nachdem die Tabelle inkludiert wurde erscheint dann folgender Knopf:

**Verberge Ausgabe: Standardabweichungen und Mittelwerte der Variablen**  
(bei nominalen Variablen: der nicht-redundanten Dummies)

Durch Klick auf diesen Knopf löscht Almo die Tabelle wieder.

Soll die Tabelle jedoch, nachdem sie inkludiert wurde, fester Bestandteil der Ergebnisliste bleiben, dann speichern Sie die Ergebnisliste (durch Klick auf den Speichern-Knopf).

Standardabweichungen  
(Standardabweichung ist mit n  
nicht mit n-1 dividiert)

	A1	0.9934
	B1	0.7243
	B2	0.8320
	A1 B1	0.7243
	A1 B2	0.8512
Leistung	V5	1.9086

Die Streuungen der Dummies sind aus einer 0,1,-1 -Kodierung hervorgegangen und deswegen kaum interpretierbar

Mittelwerte

	A1	0.1148
	B1	0
	B2	0.2131
	A1 B1	0
	A1 B2	0.1148
Leistung	V5	3.8852

Die Mittelwerte der Dummies sind aus einer 0,1,-1 -Kodierung hervorgegangen und deswegen nicht interpretierbar

\*\*\* **Erläuterung:** Beachte, dass hier nicht alle, sondern nur die **notwendigen** Dummies verwendet werden. Die jeweils letzte Dummy fehlt. Die Mittelwerte der Dummies sind gleichzeitig die Anteilswerte. Dies gilt allerdings nicht, wenn wie in unserem Beispiel das Verfahren der „weighted squares of means“ gerechnet wird. Hätten wir mit „fitting constants I“ gerechnet, dann hätten wir als Mittelwert für A1=0.5574 gefunden. Das heißt, dass in der Ausprägung A1 sich 55,74 % der Untersuchungseinheiten befinden (und der Rest in der nicht ausgewiesenen Ausprägung A2).

**Zeige Ausgabe: Quadratsummen-Matrix**

\*\*\* **Erläuterung:** Wenn der Benutzer auf den Knopf klickt, dann zeigt Almo die nachfolgend abgebildete Tabelle.

Nachdem die Tabelle inkludiert wurde erscheint dann folgender Knopf:

**Verberge Ausgabe: Quadratsummen-Matrix**

Durch Klick auf diesen Knopf löscht Almo die Tabelle wieder.

Soll die Tabelle jedoch, nachdem sie inkludiert wurde, fester Bestandteil der Ergebnisliste bleiben, dann speichern Sie die Ergebnisliste (durch Klick auf den Speichern-Knopf).

Abweichungs-Quadratsummen-Matrix  
 (die Dummies sind aus einer 0,1,-1 -Kodierung hervorgegangen. Ihre Streuungen und Co-Streuungen sind deswegen kaum interpretierbar)

		Leistung					
		A1	B1	B2	A1B1	A1B2	V5
A1		60.1967	0	5.5082	0	12.1967	-36.1967
B1		0	32.0000	16.0000	0	0	20.0000
B2		5.5082	16.0000	42.2295	0	5.5082	16.4918
A1 B1		0	0	0	32.0000	16.0000	4.0000
A1 B2		12.1967	0	5.5082	16.0000	44.1967	7.8033
Leistung	V5	-36.1967	20.0000	16.4918	4.0000	7.8033	222.1967

\*\*\*\*\* MITTEILUNG  
 Grafik zu Streuungsmatrix wird nur gezeichnet wenn keine Interaktionen zwischen den nominalen Variablen

\*\*\*\*\* MITTEILUNG  
 Fuer Analyse-Variable mit "Kein\_Wert" wurde zur Berechnung der Abweichungs-Quadratsummen-Matrix folgende Kein-Wert-Behandlung durchgefuehrt  
 Kein-Wert-Behandlung=1: "Paarweises Ausscheiden"

\*\*\*\*\* MITTEILUNG  
 Allgemeines lineares Modell wird mit folgenden Einstellungen gerechnet:

- Verfahren: weighted squares of means (=Typ III; Abschnitt P20.7.5.1)
- Analysiert wird die Matrix der Abweichungsquadrate
- Die Streuungen sind Abweichungsquadrate
- Es entstehen nicht-standardisierte Koeffizienten
- Die Teststaerke von F bzw. t wird mit alpha= 0.05 ermittelt

=====  
 Diagonalglieder der Choleskymatrix zur Ermittlung und zum Ausschluss linearer Abhaengigkeiten

Unabhaengige Variable

A1 60.196721  
 B1 32.000000  
 B2 33.725490  
 A1 B1 32.000000  
 A1 B2 33.153488

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

Keine Variable wird eliminiert

\*\*\* **Erläuterung:** ALMO überprüft die Abweichungs-Quadratsummenmatrix auf lineare Abhängigkeit, indem es die Diagonalglieder der einzelnen Variablen aus der Cholesky-Matrix ermittelt. Siehe dazu die ausführliche Darstellung in Abschnitt P20.7.7. Ist ein Wert kleiner als  $0.0001 \cdot q_{ii}$ , dann wird die entsprechende Variable eliminiert. Genauer: Ihre Wirkung innerhalb des untersuchten Modells wird auf 0 gesetzt.  $q_{ii}$  ist das

entsprechende Diagonalglied aus der Streuungsmatrix.

Beispiel:

$q_{33} = 42.2295$  in der Quadratsummenmatrix. Das Diagonalglied aus der Choleskymatrix ist  $33.72549$ . Es ist größer als  $0.0001 * 42.2295$ .

Es kann nun vorkommen, dass eine Variable im Vergleich zu den anderen Variablen ein sehr kleines Diagonalglied besitzt, von ALMO jedoch nicht eliminiert wird, da es gerade über der Eliminationsschwelle liegt. Der Benutzer kann dann das Programm noch einmal laufen lassen und dabei die betreffende Variable weglassen, bzw. wenn es sich bei ihr um eine Dummy-Variable handelt, durch eine Umkodierungsanweisung (d.h. durch Zusammenfassung von Ausprägungen) diese Dummy aus der Analyse herausnehmen. Sollten bei nominalen Variablen Ausprägungen nicht besetzt sein, dann müssen durch entsprechende Umkodierungen diese leeren Ausprägungen beseitigt werden.

## Gesamte erklärte Streuung

Alle im Folgenden angegebenen Streuungen und erkläerten Streuungen sind Abweichungsquadratsummen

=====

Gesamtstreuung	222.196721
----------------	------------

=====

Koeffizienten fuer Gesamt-Modell

Durch alle unabh. Variable	
erklärte Streuung	41.858337
Fehlerstreuung	180.338384
multipler Korrelat.koeff.	0.434032
korrigierter mult. Korr.koeff.	0.338528
F-Wert f. erklärte Streuung	2.553210
Freiheitsgrade Nenner =	5
Zähler =	55
Signifikanz: p	0.037377
Signifikanz: (1-p)*100	96.262307 %
Teststärke von F	0.750368

### \*\*\* Erläuterung

Die Gesamtstreuung beträgt 222.1967. Davon sind 180.3384 Fehlerstreuung und 41.8583 erklärte Streuungen (durch alle unabhängigen Variablen erklärte Streuung). Das heißt 18.8% der Gesamtstreuung sind erklärte Streuungen. Der Korrelationskoeffizient von 0.4340 ergibt sich als Wurzel aus diesem (zuvor mit 100 dividierten) Prozentsatz. Umgekehrt: der Prozentsatz erklärter Streuung entsteht aus dem quadrierten Korrelationskoeffizient (mal 100). Wir bezeichnen den Korrelationskoeffizienten als „multiplen“, da er sich auf alle eingeführten unabhängigen Variablen bezieht. Der F-Wert von 2.5532 ist mit  $(1-p)*100 = 96.2623$  % Wahrscheinlichkeit „überzufällig“.

### Signifikanz

Bevor wir den Begriff der Signifikanz definieren können, müssen wir die Begriffe "Nullhypothese"  $H_0$  und "Alternativhypothese"  $H_1$  erklären. In unserem Falle lautet die Nullhypothese: Die unabhängigen Variablen V1 (= Geschlecht) und V3 (= Beruf) können die abhängige Variable V5 (= Leistung) nicht erklären. Die erklärte Streuung ist in Wirklichkeit 0. Wir könnten auch sagen: Die multiple Korrelation ist in Wirklichkeit (genauer: in der Grundgesamtheit) gleich 0. Die gefundenen Werte für die erklärte Streuung von 41.8 und für die multiple Korrelation von 0.43 sind zufällig entstanden, weil unsere Untersuchungseinheiten als Zufallsstichprobe aus der Grundgesamtheit gezogen wurden.

Die Alternativhypothese  $H_1$  ist die genaue Alternative zu  $H_0$ . Die erklärte Streuung bzw. die multiple Korrelation sind (in der Grundgesamtheit) größer als 0.

Die Signifikanz  $p$  nun, die ALMO ausgibt ist das Risiko, das wir eingehen, wenn wir die

Nullhypothese  $H_0$  aufgeben und die Alternativhypothese  $H_1$  als richtig akzeptieren. Immerhin kann die erklärte Streuung bzw. die multiple Korrelation, die wir ermittelt haben, auch bei Gültigkeit der Nullhypothese zufällig entstanden sein. In unserem Beispiel ist  $p = 0.037$ . Wir könnten auch sagen: Das Risiko beträgt  $p \cdot 100 = 3,7$  Prozent. Umgekehrt formuliert: Die Sicherheit  $(1-p) \cdot 100$  ist 96,3 Prozent.

Man sollte den empirischen p-Wert vergleichen mit dem üblicherweise vorgegebenen Signifikanzniveau  $\alpha$ . Üblich sind hier die Werte  $\alpha = 0.05$  oder  $\alpha = 0.01$  oder  $\alpha = 0.001$ . In unserem Beispiel wird man sagen: Der empirische p-Wert liegt unter  $\alpha = 0.05$  aber über  $\alpha = 0.01$  und über  $\alpha = 0.001$ . Unser Ergebnis ist also mindestens auf dem 5%-Niveau signifikant.

Das Risiko, das wir hier beschrieben haben, wird in der statistischen Literatur auch Risiko I genannt, gelegentlich auch Irrtumswahrscheinlichkeit.  $\alpha$  wird als das zulässige Risiko I und  $p$  als das resultierende Risiko I bezeichnet.

Auf folgendes ist noch hinzuweisen: Mit immer größer werdenden Stichproben werden immer kleinere Abweichungen von der Nullhypothese als solche erkannt. Es könnte in unserem Falle dann als Ergebnis entstehen, dass Geschlecht und Beruf zwar signifikante Determinanten der Leistung sind, aber die Stärke des Wirkens dieser beiden Variablen so schwach ist, dass unser Ergebnis bedeutungslos ist.

### Teststärke

Die Teststärke  $s$  ist die Wahrscheinlichkeit eine richtige Alternativhypothese  $H_1$  mit dem gewählten Signifikanztest, in unserem Falle dem F-Test, auch als solche zu erkennen. Der Betafehler  $\beta$  (gelegentlich auch Risiko II genannt) ist  $1-s$ , also die Wahrscheinlichkeit eine richtige Hypothese  $H_1$  fälschlicherweise abzulehnen. Siehe hierzu z.B. Botz, Lienert, Boehnke, Kap. 2.2, insbesondere Abschnitt 2.2.7

In unserem Falle ist die Teststärke  $s = 0.75$ . Dieser Wert kann auch in folgender Weise interpretiert werden: Würden aus der Grundgesamtheit viele Stichproben gezogen werden, dann wäre der verwendete F-Test in 75% der Fälle in der Lage eine auf dem 5%-Niveau ( $\alpha = 0.05$ ) signifikante erklärte Streuung bzw. multiple Korrelation nachzuweisen. Dies ist ein unbefriedigender Wert. Zur Berechnung der Teststärke ist es notwendig einen  $\alpha$ -Wert vorzugeben. In Almo ist als Signifikanzniveau  $\alpha = 0.05$  voreingestellt. Der Benutzer kann einen anderen Wert im Maskenprogramm Prog20mo in die Eingabe-Box „Programm-Optionen lt. Handbuch“ eintragen.

### Effekt der Konstanten

```
Koeffizienten fuer Konstante
hinsichtlich der abh.Variablen   V5 Leistung
Effekt (Regressionskoeffizient)  3.872054
```

\*\*\* **Erläuterung:** Im Falle der Varianzanalyse (wenn die unabhängigen Variablen nur nominal sind) ist der Konstanten-Effekt.

beim Verfahren der "fitting constants I" gleich dem Mittelwert der abhängigen Variablen.

beim Verfahren der "weigted squares of means" gleich dem Mittelwert aus den Zellenmittelwerten der abhängigen Variablen. (siehe obige Tabelle "Zellenmittelwerte der abhaengigen Variablen").

---

### Durch nominale Variable erklärte Streuung

```
Koeffizienten fuer Variable   V1 Geschlecht
```

```

Korrelat.koeff.          0.360780
erklaerte Streuung      26.985835
F-Wert f. erklarte Streuung  8.230200
Freiheitsgrade Nenner =  1
                      Zaehler=  55
Signifikanz: p          0.005669
Signifikanz: (1-p)*100  99.433074 %
Teststaerke von F      0.804677
=====

```

Koeffizienten fuer Variable V3 Beruf

```

Korrelat.koeff.          0.272589
erklaerte Streuung      14.475570
F-Wert f. erklarte Streuung  2.207396
Freiheitsgrade Nenner =  2
                      Zaehler=  55
Signifikanz: p          0.117691
Signifikanz: (1-p)*100  88.230887 %
Teststaerke von F      0.431839

```

\*\*\* **Erläuterung:** Die durch die Variable V1, Geschlecht (oder A) erklärte Streuung ist (wie alle erklärten Streuungen in Programm 20 ) eine **partielle** erklärte Streuung. D.h. würde A aus dem Modell herausgenommen werden, dann würde sich die Streuung, die durch die noch verbleibenden Variablen V3 (oder B) und AB insgesamt erklärt wird, um 26.9858 reduzieren. Das ist also jener Streuungsanteil, den man der Variablen A "gutschreiben" kann.

Entsprechendes gilt dann selbstverständlich auch für Variable B, Beruf. Der Korrelationskoeffizient, der auf der erklärten Streuung basiert (siehe Abschnitt P20.6.3), ist demzufolge auch ein partieller Korrelationskoeffizient. In der Literatur wird dieser Koeffizient auch "Eta-Korrelation" genannt.

Koeffizienten fuer Variable : Interaktion V1\*V3

```

Korrelat.koeff.          0.160083
erklaerte Streuung      4.743012
F-Wert f. erklarte Streuung  0.723267
Freiheitsgrade Nenner =  2
                      Zaehler=  55
Signifikanz: p          0.494038
Signifikanz: (1-p)*100  50.596222 %
Teststaerke von F      0.166305
=====

```

## Effekt der nominalen Variablen

Koeffizienten der Dummies  
hinsichtlich der abh. Var. V5 Leistung

Effekte von A Geschlecht

	Effekte	Standard- fehler	erklarte Streuung	partielle Korrelat.	t-Wert	Signifikanz p	Signifikanz (1-p)100	Test- staerke
A1 männlic	-0.6961	0.2427	26.9858	-0.3608	2.8688	0.0057	99.43%	0.8052
A2 weiblic	0.6961	0.2427	26.9858	0.3608	2.8688	0.0057	99.43%	0.8052

Paarweise Vergleiche (Kontraste) von A Geschlecht

	Differenz	Standard- fehler	erklarte Streuung	t-Wert (LSD)	Signifikanz p	Signifikanz (1-p)100	Test- staerke
A1 - A2	-1.3923	0.4853	26.9858	2.8688	0.0057	99.43%	0.8052

Freiheitsgrade fuer t-Wert: 55

Effekte von B Beruf

	Effekte	Standard- fehler	erklärte Streuung	partielle Korrelat.	t-Wert	Signifikanz p	(1-p)100	Test- staerke
B1 Arbeit	0.5029	0.3566	6.5210	0.1868	1.4102	0.1643	83.57%	0.2836
B2 Angeste	0.2441	0.3145	1.9756	0.1041	0.7762	0.4409	55.91%	0.1188
B3 Selbstä	-0.7471	0.3566	14.3872	-0.2718	2.0947	0.0408	95.92%	0.5399

Paarweise Vergleiche (Kontraste) von B Beruf

	Differenz	Standard- fehler	erklärte Streuung	t-Wert (LSD)	Signifikanz p	(1-p)100	Test- staerke
B1 - B2	0.2588	0.5701	0.6759	0.4540	0.6513	34.87%	0.0731
B1 - B3	1.2500	0.6402	12.5000	1.9525	0.0559	94.41%	0.4842
B2 - B3	0.9912	0.5701	9.9116	1.7386	0.0878	91.22%	0.4013

Freiheitsgrade fuer t-Wert: 55


\*\*\* **Erläuterung:** Die Effekte sind die Koeffizienten der linearen Gleichung, die wir eingangs dieses Abschnitts P20.9.1 beschrieben haben. Der Benutzer lese noch einmal diese Ausführungen, sowie die ausführliche Erörterung der Effekte in Abschnitt P20.6.5 und P20.7.5. Die "erklärten Streuungen" der Effekte (und die aus ihnen abgeleiteten partiellen Korrelationskoeffizienten) sind mit Sorgfalt zu interpretieren. Sie sind nicht additiv, da die Dummies einer nominalen Variablen miteinander korrelieren. Betrachten wir die Variable B. Die Addition der erklärten Streuungen der Dummies B1, B2, B3 ist mit 22.8838 (erheblich) größer als die durch die nominale Variable B tatsächlich erklärte Streuung von 14.4755. Sehr wohl sind jedoch die erklärten Streuungen und die partiellen Koeffizienten einer nominalen Variablen untereinander zu vergleichen. Wir erkennen, dass der Ausprägung B3 die größte Bedeutung zukommt. Die Erklärung dafür liefern uns die paarweisen Vergleiche (Kontraste). B3 kontrastiert mit den anderen Ausprägungen B1 und B2 sehr viel stärker als diese dies untereinander tun. Der t-Wert und dessen Signifikanz geben an, ob der betreffende Effekt von 0 signifikant verschieden ist. Der Effekt B3, z.B.: besitzt einen t-Wert von 2.0947 und ist damit von 0 mit einer Sicherheit von 95.92 % verschieden.

Effekte von AB

	Effekte	Standard- fehler	erklärte Streuung	partielle Korrelat.	t-Wert	Signifikanz p	(1-p)100	Test- staerke
A1 B1	-0.0539	0.3566	0.0748	-0.0204	0.1511	0.8802	11.98%	0.0525
A1 B2	0.3577	0.3145	4.2430	0.1516	1.1376	0.2603	73.97%	0.2010
A1 B3	-0.3039	0.3566	2.3804	-0.1141	0.8520	0.3978	60.22%	0.1333
A2 B1	0.0539	0.3566	0.0748	0.0204	0.1511	0.8802	11.98%	0.0525
A2 B2	-0.3577	0.3145	4.2430	-0.1516	1.1376	0.2603	73.97%	0.2010
A2 B3	0.3039	0.3566	2.3804	0.1141	0.8520	0.3978	60.22%	0.1333

## Randmittel

In der Ergebnisliste erscheint folgende Textzeile

 **Zeige Ausgabe:      Geschaetzte Randmittel**

\*\*\* **Erläuterung:** Wenn der Benutzer auf den Knopf klickt, dann fügt Almo die nachfolgend abgebildete Tabelle der geschätzten Randmittel in die Ergebnisliste ein.

Nachdem die Tabelle inkludiert wurde erscheint dann folgender Knopf:

**↓ Verberge Ausgabe: Geschaetzte Randmittel**

Durch Klick auf diesen Knopf löscht Almo die Tabelle wieder. Soll die Tabelle jedoch, nachdem sie inkludiert wurde, fester Bestandteil der Ergebnisliste bleiben, dann speichern Sie die Ergebnisliste (durch Klick auf den Speichern-Knopf am Oberrand des Almofensters).

Wir haben das Programm zwei Mal gerechnet. In der nachfolgenden Tabelle werden in der 1. Spalte die Randmittel (geschätzt nach dem standardmäßig eingestellten Verfahren der "weighted squares of means") ausgegeben und in der 2. Spalte die Randmittel geschätzt nach dem Verfahren der "fitting constants I" bzw. nach dem bei dieser Variablen-Konstellation identischen "fitting constants II" (=SS-Typ II).

geschätzte Rand- und Zellenmittel  
hinsichtlich der abhaengigen Variablen Leistung

	Verfahren weighted squares of means -----	Verfahren fitting constants I+II -----
modellreproduzierter Gesamtmittelwert	3.872054	3.885246
=====		
V1 Geschlecht		
A1 männlich	3.175926	3.329394
A2 weiblich	4.568182	4.585208
V3 Beruf		
B1 Arbeiter	4.375000	4.302945
B2 Angestellt	4.116162	4.113992
B3 Selbständi	3.125000	3.052945
=====		
2-er Randmittel AB Geschlecht*Beruf		
A1 B1	3.625000	3.625000
A1 B2	3.777778	3.777778
A1 B3	2.125000	2.125000
A2 B1	5.125000	5.125000
A2 B2	4.454545	4.454545
A2 B3	4.125000	4.125000
=====		

Wir wollen die *Randmittel* mit den *Zellenmittel* vergleichen. Die empirischen Zellenmittel werden von Prog20mx oder Prog20mo standardmäßig ermittelt und ausgegeben. Almo liefert folgende Tabelle

Zellenmittelwerte der  
abhaengigen Variablen

		+-----+
		Leistun
		V5
Geschlec	Beruf	
+-----+		+-----+
männlich	Arbeiter	3.6250
	Angestel	3.7778
	Selbstän	2.1250
+-----+		+-----+
weiblich	Arbeiter	5.1250
	Angestel	4.4545
	Selbstän	4.1250
+-----+		+-----+
Gesamt	mittel	3.8852

+-----+  
Mittelwert aus Zellenmittelwerten  
Leistung 3.8721

Wir stellen diese Tabelle etwas um und ermitteln die Randmittel von Geschlecht und Beruf als Mittelwerte aus den Zellenmittelwerten

		empirische Zellenmittel Geschlecht*Beruf Z			Randmittel aus Zellenmittel Geschlecht RG2
		Beruf			
		Arbeiter 1	Angestel 2	Selbstän 3	
Geschl männlich	1	3.6250	3.7778	2.1250	3.1759
weiblich	2	5.1250	4.4545	4.1250	4.5682
Randmittel aus Zellenmittel Beruf RB2		4.3750	4.1162	3.1250	3.8721

Gesamtmittel  
aus Zellenmittel

Man erkennt, dass diese Randmittel identisch sind mit denen aus obiger Tabelle. Dies gilt allerdings nur, wenn mit dem Verfahren der "weighted squares of means" gerechnet wird.

Das geschätzte Randmittel, z.B. für die Ausprägung j des Berufs ergibt sich gemäß folgender Gleichung:

$$R(j) = b(j) + \text{Konstante}$$

$R(j)$  = geschätztes Randmittel der Ausprägung j der Variablen Beruf

$b(j)$  = Effekt der Ausprägung j der Variablen Beruf

Für die drei Berufe erhält man so

B1 Arbeiter:	$0.5029 + 3.8721 = 4.3750$
B2 Angestellter:	$0.2441 + 3.8721 = 4.1162$
B3 Selbständiger:	$-0.7471 + 3.8721 = 3.1250$

Das Randmittel für die Ausprägung j des Berufs kann also betrachtet werden als Prognosewerte, bei dem alle anderen Effekte nicht berücksichtigt werden.

Dasselbe gilt auch für die Interaktionen. Das geschätzte Randmittel der 2-er Interaktion ergibt sich gemäß

$$R(ij) = a(i) + b(j) + ab(ij) + \text{Konstante}$$

also aus den Interaktionseffekt plus  
den Haupteffekten, welche die Interaktion bilden plus  
der Konstanten

Das Randmittel für beispielsweise die Interaktion A2B3 wird so berechnet:

$$R(A2B3) = A2+B3+A2B3+konst = 0.6961-0.7471+0.3039+3.872054 = 4.125$$

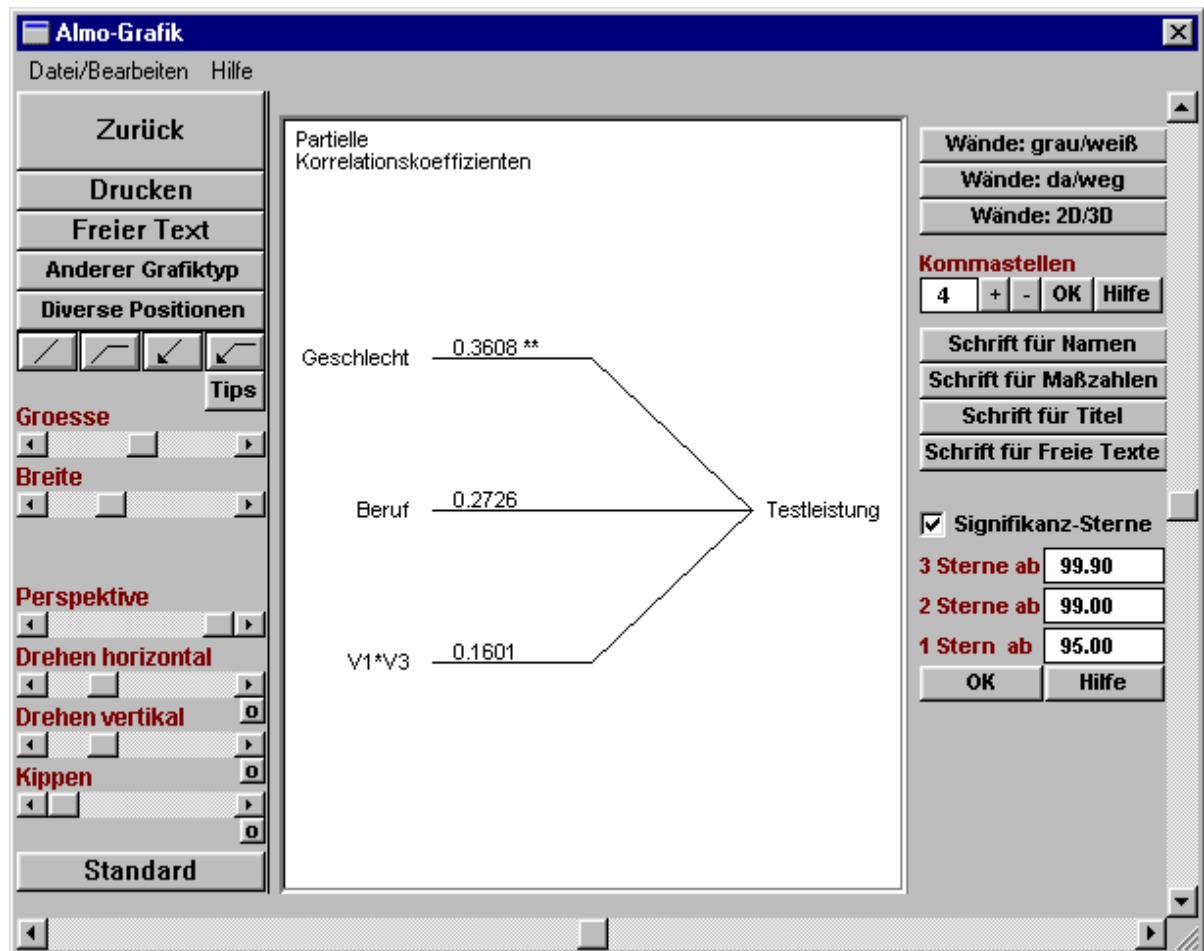
Da die Randmittel von manchen Autoren als ein wichtiges Ergebnis aus der Varianz-Kovarianzanalyse eingeschätzt werden und in der Ergebnisausgabe von z.B. SPSS einen großen Raum einnehmen, werden wir im anschließenden Abschnitt P20.9.1.0 die Randmittel nochmals ausführlich betrachten.

### Zusammenfassung: Erklärte Streuungen und Diagramme

Zusammenfassung

Streuungsquelle	Streuung	Korrel. Koeff.	F-Wert	df	Signifikanz p	(1-p) 100	Test-staerke
Gesamtstreuung	222.1967						
Fehlerstreuung	180.3384			55			
alle unabh. Var. zusammen	41.8583	0.4340	2.5532	5	0.0374	96.2623	0.7504
V1 Geschlecht	26.9858	0.3608	8.2302	1	0.0057	99.4331	0.8047
V3 Beruf	14.4756	0.2726	2.2074	2	0.1177	88.2309	0.4318
V1*V3	4.7430	0.1601	0.7233	2	0.4940	50.5962	0.1663

Almo zeichnet folgendes Flussdiagramm der partiellen Korrelationskoeffizienten:



Die Grafik kann in vielfältiger Weise verändert werden. Wir wollen hier nur auf folgende Möglichkeit hinweisen: Der Koeffizient "Geschlecht" mit 0.3608 hat 2 Sterne. Damit soll ausgedrückt werden, dass er mit 99% (p=0.01) signifikant ist. Auf der rechten Editorseite gibt es nun eine Checkbox "Signifikanz-Sterne". In den 3 Feldern darunter kann der Benutzer

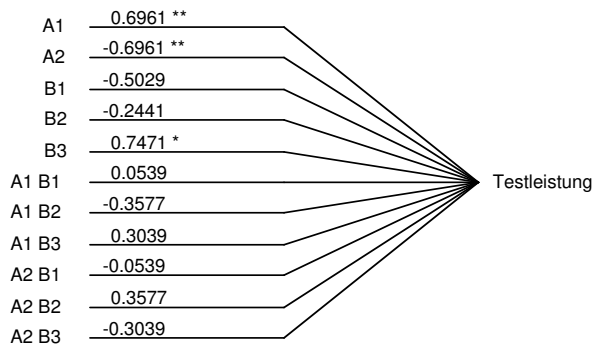
festlegen, bei welchem Signifikanzniveau  $(1-p)*100$  ein oder zwei oder drei Sterne hinter den Koeffizienten gesetzt werden. Danach muss noch auf OK geklickt werden.

Zusammenfassung: Effekte  
und ihre Signifikanzen  
hinsichtlich der abhaengigen Variablen  
Leistung

	Effekte	Signifikanz (1-p) *100
A1 männlich	-0.696128	99.428574
A2 weiblich	0.696128	99.428574
B1 Arbeiter	0.502946	83.574330
B2 Angestellt	0.244108	55.907068
B3 Selbständig	-0.747054	95.920065
A1 B1	-0.053872	11.979539
A1 B2	0.357744	73.972424
A1 B3	-0.303872	60.221999
A2 B1	0.053872	11.979539
A2 B2	-0.357744	73.972424
A2 B3	0.303872	60.221999
Konstante	3.872054	-

Almo zeichnet folgendes Flussdiagramm der Effekte und Regressionskoeffizienten

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



### P20.9.1.0 Randmittel


Siehe hierzu auch die ausführliche Darstellung in Abschnitt P20.6.6.1.

Den Forscher interessiert primär


1. welche (ursächlichen) Variable seine Zielvariable *signifikant* erklären, bzw. welche von ihm in die Analyse eingeführten erklärenden Variablen dies nicht leisten können
2. und wie *unterschiedlich stark* die signifikanten Variablen die Zielvariable erklären.

Die geschätzten Randmittel sind in diesem Zusammenhang eher von zweitrangiger Bedeutung. In Almo werden deswegen die Randmittel zwar berechnet und intern gespeichert, aber nur ausgegeben, wenn der Benutzer sie anfordert. In der Ergebnisliste muss der Benutzer

auf diesen Knopf klicken

 **Zeige Ausgabe: Geschaetzte Randmittel**

Wenn der Benutzer auf den Knopf klickt, dann zeigt Almo die nachfolgend abgebildete Tabelle der geschätzten Randmittel. Nachdem die Tabelle inkludiert wurde erscheint dann folgender Knopf:

 **Verberge Ausgabe: Geschaetzte Randmittel**

Durch Klick auf diesen Knopf löscht Almo die Tabelle wieder aus der Ergebnisliste

Soll die Tabelle jedoch, nachdem sie inkludiert wurde, fester Bestandteil der Ergebnisliste bleiben, dann speichern Sie die Ergebnisliste (durch Klick auf den Speichern-Knopf).

Wir haben das Programm nach dem Verfahren der "weighted squares of means" (bzw. SS-Typ III) und ein 2. Mal nach dem Verfahren der "fitting constants I" gerechnet. Bei 2 nominalen Variablen und ihrer Interaktion sind fitting constants I und II identisch. Fitting constants II entspricht SS-Typ II bei SAS bzw. SPSS.

In der nachfolgenden Tabelle werden in der 1. Spalte die Randmittel ausgegeben, wie sie vom standardmäßig eingestellten Verfahren der "weighted squares of means" (bzw. SS-Typ III) geschätzt werden und in der 2. Spalte die Randmittel, wie sie vom Verfahren der "fitting constants I" oder II geschätzt werden. Die beiden Ergebnisse sind bei den Haupteffekten ähnlich, stimmen aber nicht überein. Bei der 2-er Interaktion sind sie gleich. Generell gilt, dass die Randmittel der höchsten Interaktion bei den Verfahren gleich sind. Sie sind identisch mit den empirischen Zellenmitteln. Wir werden das noch darstellen.

geschätzte Randmittel  
hinsichtlich der abhaengigen Variablen Leistung

	Verfahren weighted squares of means -----	Verfahren fitting constants I+II -----
modellreproduzierter Gesamtmittelwert (Konstante)	3.872054	3.885246
=====		
V1 Geschlecht		
A1 männlich	3.175926	3.329394
A2 weiblich	4.568182	4.585208
V3 Beruf		
B1 Arbeiter	4.375000	4.302945
B2 Angestellt	4.116162	4.113992
B3 Selbständi	3.125000	3.052945
=====		
2-er Randmittel AB Geschlecht*Beruf		
A1 B1	3.625000	3.625000
A1 B2	3.777778	3.777778
A1 B3	2.125000	2.125000
A2 B1	5.125000	5.125000
A2 B2	4.454545	4.454545
A2 B3	4.125000	4.125000
=====		

Das geschätzte Randmittel, z.B. für die Ausprägung  $j$  des Berufs ergibt sich gemäß folgender Gleichung:

$$(1) \quad R(j) = B(j) + \text{Konstante}$$

$R(j)$  = geschätztes Randmittel der Ausprägung  $j$  der Variablen Beruf

$B(j)$  = Effekt der Ausprägung  $j$  der Variablen Beruf

Dasselbe gilt auch für die Variablenkombination AB. Das geschätzte Randmittel  $R(ij)$  ergibt sich gemäß

$$(2) \quad R(ij) = A(i) + B(j) + AB(ij) + \text{Konstante}$$

also aus dem Interaktionseffekt  $AB(ij)$  plus

den Haupteffekten  $A(i)$  und  $B(j)$ , welche die Variablenkombination AB bilden plus der Konstanten

Für die drei Berufe erhält man so gemäß Gleichung (1)

B1 Arbeiter:	0.5029 + 3.8721 =	4.3750
B2 Angestellter:	0.2441 + 3.8721 =	4.1162
B3 Selbständiger:	-0.7471 + 3.8721 =	3.1250

Das Randmittel für die Ausprägung  $j$  des Berufs kann also betrachtet werden als der Prognosewerte für die Probanden mit Beruf  $j$ , bei dem alle anderen Effekte (Geschlecht  $i$  und Interaktion  $ij$ ) "auspartielliert" sind.

Das Randmittel für beispielsweise die Kombination A2B3 wird gemäß Gleichung (2) so berechnet:

$$R(A2B3) = A2 + B3 + A2B3 + \text{konst} = 0.6961 - 0.7471 + 0.3039 + 3.872054 = 4.125$$

Die Formeln für geschätzte Randmittel höherer Ordnung als 2 werden wir weiter unten angeben. Sind Kovariate vorhanden (Fall der Kovarianzanalyse) dann werden in die Formeln die kovarianzadjustierten Effekte eingesetzt.

Wir unterscheiden folgende zwei Begriffe

1. geschätzte **Randmittel**. Sie werden von Almo, wie gezeigt, im Rahmen einer Varianz-Kovarianzanalyse aus den Effekten *geschätzt*.
2. **Zellenmittelwerte**. Das sind die aus den Probanden einer Zelle (z.B. der Zelle A2B3 also „Geschlecht-weiblich \* Beruf-Selbständig“) errechneten Mittelwerte. Sie sind *empirische* Mittelwerte.

Die empirischen Zellenmittelwerte werden von Prog20mx oder Prog20mo standardmäßig ermittelt und ausgegeben. Almo liefert dabei folgende Tabelle

Zellenmittelwerte der  
abhaengigen Variablen

A	B	Leistung
Geschlecht	Beruf	
männlich	Arbeiter	3.6250
	Angestel	3.7778
	Selbstän	2.1250
weiblich	Arbeiter	5.1250
	Angestel	4.4545
	Selbstän	4.1250
Gesamtmittel		3.8852

Mittelwert aus Zellenmittelwerten  
Leistung 3.8721

Wir stellen diese Tabelle etwas um und ermitteln die Zeilen-Mittelwerte von Geschlecht und die Spaltenmittelwerte von Beruf - als Mittelwerte aus den Zellenmittelwerten

		empirische Zellenmittel Geschlecht*Beruf			Zeilenmittel aus Zellenmittel Geschlecht
		Beruf			
		Arbeiter 1	Angestel 2	Selbstän 3	
Geschl	männlich 1	3.6250	3.7778	2.1250	3.1759
	weiblich 2	5.1250	4.4545	4.1250	4.5682
Spaltenmittel aus Zellenmittel Beruf		4.3750	4.1162	3.1250	3.8721
					Gesamtmittel aus Zellenmittel

Das *Zeilenmittel* z.B. für "Geschlecht männlich" von 3.1759 entsteht aus

$$(3.6250+3.7778+2.1250)/3 = 3.1759$$

Man erkennt, dass diese so berechneten Zeilen- und Spaltenmittel identisch sind mit den "geschätzten Randmitteln". Dies gilt allerdings nur, wenn mit dem Verfahren der "weighted squares of means" gerechnet wird. Dann gilt es auch prinzipiell immer für jede beliebige Zahl von unabhängigen nominalen Variablen. Wir werden dies anschliessend in einer Analyse mit 4 nominalen Variablen zeigen.

### Randmittel bei fitting constants I

Wird jedoch mit dem Verfahren der "fitting constants I" oder II gerechnet, dann gilt dies nur für die 2-er Interaktion AB. Betrachten wir A1 Geschlecht – männlich. Der Wert des Randmittels A1 (=männlich) ist nicht der empirische Mittelwert aus allen Männern, wie man vermuten könnte. Der empirische Mittelwert in der Zielvariablen Leistung aus allen 34 Männern ist 3.3529. Das Randmittel ist jedoch 3.329394.

Für die 27 Frauen ist der empirische Mittelwert 4.5556 und das Randmittel 4.585208.

*Das Randmittel ist der Wert, den das Verfahren (in unserem Beispiel) für die Männer bzw. die Frauen reproduziert, wobei die andere Variablen B (Beruf) „auspartiielliert“ ist.*

Dies gilt für den Fall ungleicher Zellenhäufigkeiten. Bei gleichen bzw. balancierten Zellenhäufigkeiten, wenn also A und B unabhängig sind, dann fällt alles zusammen:

Randmittel der Gruppe i = empirischer Mittelwert für i = Mittelwert aus Zellenmittel für i

**Randmittel aus Analyse mit 4 nominalen Variablen**

Wir rechnen mit dem Verfahren der "weighted squares of means" ein Beispiel mit unseren Testdaten mit 4 nominalen Variablen.

- A Geschlecht: (1) männlich, (2) weiblich;
- B Wohnort: (1) Stadt, (2) Land;
- C Schulbildung: (1) niedrig, (2) hoch;
- D Beruf: (1) Arbeiter, (2) Angestellter, (3) Selbständiger;

und Leistung als abhängiger Variablen.

Das Programm ist unter dem Namen "Prog20\_Randmittel.Alm" nach Klick auf den Knopf "alle Progs" am Oberrand des Almofensters zu finden.

Almo liefert in der Ergebnisliste zu diesem Programm folgende Tabelle der Zellenmittelwerte. Hinter diese Tabelle haben wir die „Tabelle als Datenmatrix“ angefügt.

A	B	C	D	Leistung	Tabelle als Datenmatrix "ABCDdat.fre"				
Geschlec	Wohnort	Schulbild	Beruf		Datei der nominalen Var. und der Zellenmittelwerte				
männlich	Stadt	niedrig	Arbeiter	4.5000	1	1	1	1	4.5000
			Angestel	4.0000	1	1	1	2	4.0000
			Selbstän	4.0000	1	1	1	3	4.0000
		hoch	Arbeiter	3.0000	1	1	2	1	3.0000
			Angestel	3.0000	1	1	2	2	3.0000
			Selbstän	1.5000	1	1	2	3	1.5000
	Land	niedrig	Arbeiter	3.5000	1	2	1	1	3.5000
			Angestel	3.5000	1	2	1	2	3.5000
			Selbstän	2.0000	1	2	1	3	2.0000
		hoch	Arbeiter	3.5000	1	2	2	1	3.5000
			Angestel	3.9167	1	2	2	2	3.9167
			Selbstän	1.0000	1	2	2	3	1.0000
weiblich	Stadt	niedrig	Arbeiter	6.0000	2	1	1	1	6.0000
			Angestel	5.5000	2	1	1	2	5.5000
			Selbstän	6.0000	2	1	1	3	6.0000
		hoch	Arbeiter	3.5000	2	1	2	1	3.5000
			Angestel	4.0000	2	1	2	2	4.0000
			Selbstän	3.5000	2	1	2	3	3.5000
	Land	niedrig	Arbeiter	6.0000	2	2	1	1	6.0000
			Angestel	4.0000	2	2	1	2	4.0000
			Selbstän	4.5000	2	2	1	3	4.5000
		hoch	Arbeiter	5.0000	2	2	2	1	5.0000
			Angestel	5.0000	2	2	2	2	5.0000
			Selbstän	2.5000	2	2	2	3	2.5000
Gesamtmittel				3.8852					

Mittelwert aus Zellenmittelwerten  
V5      Leistung                      3.8715

Aus den Codeziffern der 4 nominalen Variablen und den Zellenmittelwerten erzeugen wir die Datei ABCDdat.fre. Wir nennen sie "Datei der Zellenmittelwerte". Ein Datensatz besteht aus einem Zellenmittelwert und den die Zelle definierenden Variablenausprägungen. Der Benutzer findet die Datei im Ordner TESTDAT. Wir haben sie oben parallel neben die Tabelle der Zellenmittelwerte geschrieben. Die Datenmatrix ABCDdat.fre kann mit **Prog20tc** erstellt, ausgegeben und in eine Datei gespeichert werden. Dieses Programm wird ebenfalls nach Klick auf den Knopf "alle Progs" gefunden.

Von den geschätzten Randmittel, die Almo berechnet und ausgibt, wollen wir hier nur die für die Haupteffekte, die für die 2-er Interaktion BC und die für die 3-er Interaktion ABD zeigen.

```

Randmittel
=====
Gesamtmittelwert (Konstante)      3.871528
=====
A Geschlecht
  A1 männlich                      3.118056
  A2 weiblich                      4.625000
B Wohnort
  B1 Stadt                        4.041667
  B2 Land                         3.701389
C Schulbildung
  C1 niedrig                      4.458333
  C2 hoch                        3.284722
D Beruf
  D1 Arbeiter                    4.375000
  D2 Angestellt                  4.114583
  D3 Selbständi                 3.125000
=====
. .
. .
=====
2-er Randmittel BC Wohnort*Schulbildung
  B1 C1                          5.000000
  B1 C2                          3.083333
  B2 C1                          3.916667
  B2 C2                          3.486111
=====
. .
. .
=====
3-er Randmittel ABD Geschlecht*Wohnort*Beruf
  A1 B1 D1                       3.750000
  A1 B1 D2                       3.500000
  A1 B1 D3                       2.750000
  A1 B2 D1                       3.500000
  A1 B2 D2                       3.708333
  A1 B2 D3                       1.500000
  A2 B1 D1                       4.750000
  A2 B1 D2                       4.750000
  A2 B1 D3                       4.750000
  A2 B2 D1                       5.500000
  A2 B2 D2                       4.500000
  A2 B2 D3                       3.500000
=====
. .
. .
=====
4-er Randmittel ABCD Geschl*Wohn*Schulb*Beruf
  A1 B1 C1 D1                    4.500000
  A1 B1 C1 D2                    4.000000
. . .
. . .
  A2 B2 C2 D2                    5.000000

```

Wir wollen am Beispiel dieser Analyse mit 4 nominalen Variablen A, B, C, D die allgemeine Formel für die 2-er Randmittel, für die 3-er Randmittel und das 4-er Randmittel anschreiben. Dabei verwenden wir folgende Notation:

R = Randmittel

E = Effekt

mit  $i, j, k, l$  werden die Ausprägungen der nominalen Variablen A, B, C, D bezeichnet

$E(i), E(j), E(k), E(l)$  = Haupteffekte von A, B, C, D

$E(ij)+E(ik)+E(il), E(jk), \dots$  = 2-er Interaktionseffekte

$E(ijk)$  = Effekt der 3-er Interaktion  $ijk$

$E(ijkl)$  = Effekt der 4-er Interaktion  $ijkl$

K = Konstante

G1 = Gruppe der Haupteffekte

G2 = Gruppe der 2-er Interaktionseffekte

G3 = Gruppe der 3-er Interaktionseffekte

G4 = Interaktionseffekte 4. Ordnung

**2-er Randmittel R(AB):**

$$R(Ai.Bj) = K + E(Ai) + E(Bj) + E(Ai.Bj)$$

**2-er Randmittel R(AC):**

$$R(Ai.Ck) = K + E(Ai) + E(Ck) + E(Ai.Ck)$$

.....  
.....

**2-er Randmittel R(CD):**

$$R(Ck.Dl) = K + E(Ck) + E(Dl) + E(Ck.Dl)$$

Was wird nun durch ein geschätztes 2-er Randmittel, beispielsweise  $R(C1D2)$  ausgedrückt?

**3-er Randmittel R(ABC):**

$$R(Ai.Bj.Ck) = K + \frac{E(Ai) + E(Bj) + E(Ck)}{G1} + \frac{E(Ai.Bj) + E(Ai.Ck) + E(Bj.Ck)}{G2} + \frac{E(Ai.Bj.Ck)}{G3}$$

.....  
.....

**3-er Randmittel R(BCD):**

$$R(Bj.Ck.Dl) = K + \frac{E(Bj) + E(Ck) + E(Dl)}{G1} + \frac{E(Bj.Ck) + E(Bj.Dl) + E(Ck.Dl)}{G2} + \frac{E(Bj.Ck.Dl)}{G3}$$

**4-er Randmittel R(ijkl)**

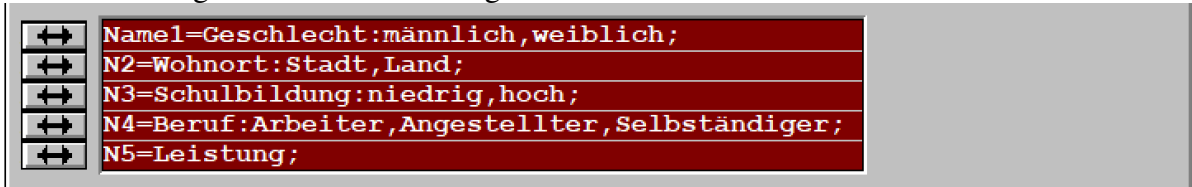
Wir vereinfachen die Formel, indem wir nur die Indices  $i, j, k, l$  der 4 Variablen A B C D schreiben

$$R(ijkl) = K + \frac{E(i)+E(j)+E(k)+E(l)}{G1} + \frac{E(ij)+E(ik)+E(il)+E(jk)+E(jl)+E(kl)}{G2} + \frac{E(ijk)+E(ijl)+E(ikl)+E(jkl)}{G3} + \frac{E(ijkl)}{G4}$$

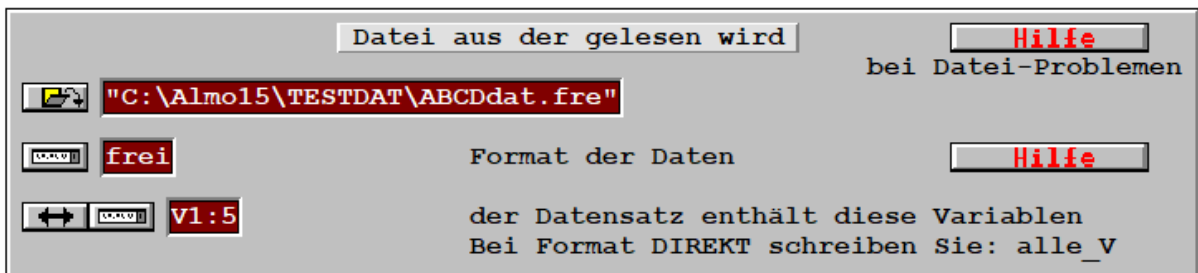
Wir haben oben aus den Codeziffern der 4 nominalen Variablen und den Zellenmittelwerten die Datei ABCDdat.fre erzeugt. Dazu wurde die Programm-Maske **Prog20tc** verwendet. Aus der Programm-Maske für Mittelwertsdifferenzen Prog18m2 haben wir das **Prog18\_Randmitt.Alm** gebildet. Es liest die Datei ABCDdat.fre ein. Mit ihm (und dieser Datei) können nun sämtliche Randmittel als Mittelwerte der jeweiligen Zellenmittelwerte errechnet werden

1. für die Haupteffekte A Geschlecht, B Wohnort, C Schulbildung und D Beruf
2. für die 2-er Variablenkombination AB, AC, AD, BC, BD, CD
3. für die 3-er Variablenkombination ABC, ABD, BCD
4. für die 4-er Variablenkombination ABCD. Diese ist identisch mit der oben bereits abgebildeten Tabelle der Zellenmittelwerte.

Die Datei ABCDdat.fre besitzt 5 Spalten. Sie erhalten in der Eingabebox "Freie Namensfelder" von Prog18\_Randmitt.Alm folgende Variablenamen



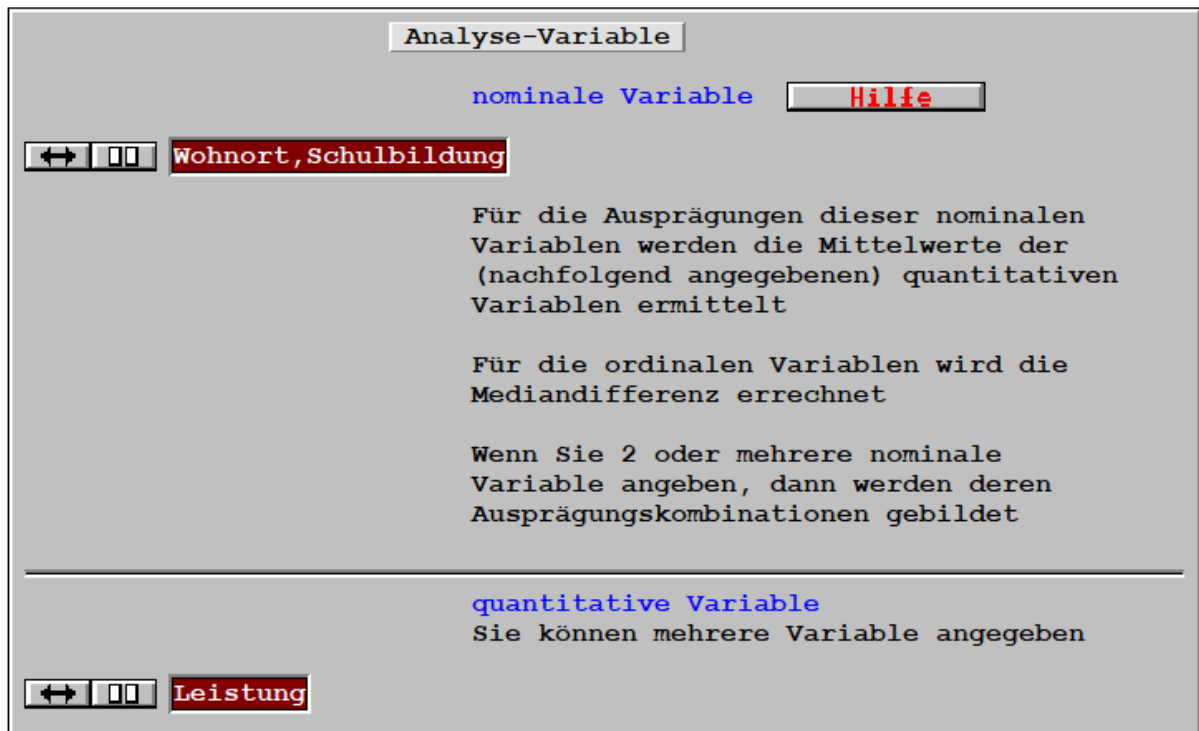
Die Dateibox von Prog18\_Randmitt.Alm ist folgende



In der Eingabebox "Analysevariable - nominale Variable" müssen die nominalen Variablen eingetragen werden, die die jeweilige Variablenkombination bilden. Das Programm wird gefunden durch Klick auf den Knopf "alle Progs" am Oberrand des Almofensters. Es ist ausgefüllt für die

2-er Variablenkombination BC „Wohnort\*Schulbildung“.

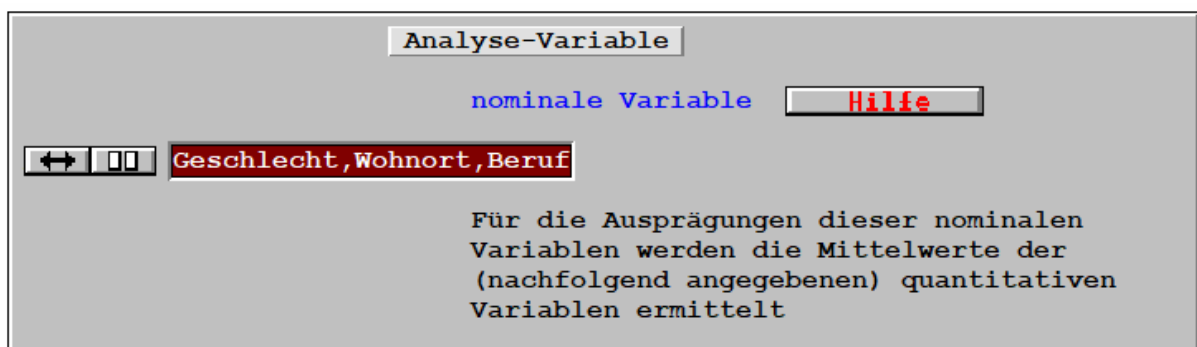
Die Eingabebox für die Analysevariablen ist folgende



Das Programm liefert folgende Zellenmittel für die 2-er Variablenkombination BC

		Leistung
B	C	
Wohnort	Schulbild	
Stadt	niedrig	5.000000
	hoch	3.083333
Land	niedrig	3.916667
	hoch	3.486117
Gesamtmittel		3.871529

Der Zellenmittelwert z.B. der 2-er Variablenkombination B1C2 ist 3.083333. Wir erkennen dass dieser mit dem oben ausgegeben Randmittel B1C2 identisch ist. Wir rechnen noch eine zweite Analyse für die 3-er Variablenkombination ABD. In "Prog18\_Randmitt.Alm" muss in der Eingabebox "Analyse-Variable - nominale Variable" dazu folgendes eingegeben werden.



Almo gibt dann diese Zellenmittelwerte aus:

A	B	D	Leistung
Geschlec	Wohnort	Beruf	
männlich	Stadt	Arbeiter	3.750000
		Angestel	3.500000
		Selbstän	2.750000
	Land	Arbeiter	3.500000
		Angestel	3.708350
		Selbstän	1.500000
weiblich	Stadt	Arbeiter	4.750000
		Angestel	4.750000
		Selbstän	4.750000
	Land	Arbeiter	5.500000
		Angestel	4.500000
		Selbstän	3.500000
Gesamtmittel			3.871529

Wir erkennen wieder, dass diese Zellenmittelwerte identisch mit den Randmittel für ABD sind.

Dies gilt allgemein: Die Randmittel aus einer Varianzanalyse sind identisch mit den Zellenmittelwerten der entsprechenden Variablenkombination. Dies gilt allerdings nur unter folgenden 2 Bedingungen: (1) Die Randmittel wurden in einer Analyse nach dem Verfahren der "weighted squares of means" errechnet und (2) es wurde ein volles Modell mit allen Interaktionen bis zur höchsten Ordnung gerechnet. In diesem Fall können also mit den beiden Programm-Masken Prog20tc und Prog18\_Randmitt.Alm die Randmittel für Interaktionen jeglicher Ordnung errechnet werden.

### Beispiel mit fitting constants I gerechnet

Wir rechnen nun dieses 4-Variablen-Beispiel nach dem Verfahren der fitting constants I. Im Programm "Prog20\_Randmittel.Alm" muss nur die Optionsbox „Verfahren“ geöffnet werden und dort „fitting constants I“ eingesetzt werden. Das Programm liefert folgende Ausgabe:

hinsichtlich der abhaengigen Variablen Leistung		aus "weighted squares of means"
modellreproduzierter Gesamtmittelwert (Konstante)		3.871528
=====		
V1 Geschlecht		
A1	männlich	3.897667   3.118056
A2	weiblich	3.735404   4.625000
V2 Wohnort		
B1	Stadt	4.022667   4.041667
B2	Land	3.872825   3.701389
V4 Schulbildung		
C1	niedrig	4.035088   4.458333
C2	hoch	3.747825   3.284722
V3 Beruf		
D1	Arbeiter	3.872825   4.375000

D2	Angestellt	4.035088		4.114583
D3	Selbständi	3.747825		3.125000

Für die Interaktionen 2-ter, 3-ter und 4-ter Ordnungen werden keine Randmittel ausgegeben. Sie können nicht berechnet werden, da (bei ungleichen Zellenhäufigkeiten) die Effekte – aus denen die Randmittel berechnet werden – nicht bestimmbar sind, bzw. „wechselnde Werte“ besitzen. Siehe dazu nachfolgenden Abschnitt. Vergleicht man die ausgegebenen Randmittel mit denen, die aus dem Verfahren der "weighted squares of means" hervor gegangen sind, erkennt man doch deutliche Unterschiede.

### ***P20.9.1.1 "Wechselnde Werte" für die Interaktionseffekte bei "fitting constants I+II"***

Siehe hierzu auch unsere Ausführungen in P20.6.5.1. Bei Analysen mit 3 und mehr unabhängigen nominalen Variablen tritt eine Besonderheit auf sofern mit dem Verfahren der "fitting constants I" gerechnet wurde und sofern ungleiche Zellenhäufigkeiten vorliegen. Betrachten wir ein Beispiel mit 3 unabhängigen nominalen Variablen A, B, C.

<u>Nominale Variable</u>	<u>Zweier- Interaktion</u>	<u>Dreier- Interaktionen</u>
A,B,C	AB, AC, BC	ABC

Diese 2er-Interaktionen besitzen "wechselnde Werte" - sofern ungleiche Zellenhäufigkeiten vorliegen

Prinzipiell gilt: Die Interaktionen - mit Ausnahme derer letzter Ordnung - besitzen beim Verfahren fitting constants I wechselnde Werte (sofern ungleiche Zellenhäufigkeiten bestehen).

ALMO bringt beispielsweise für die Interaktion AB folgende Effekte:

Effekte von AB	
-----	
A1 B1 C1	0.0441
A1 B1 C2	0.0627
A1 B1 C3	0.0441
-----	
A1 B2 C1	-0.0393
A1 B2 C2	-0.0207
A1 B2 C3	-0.0393
-----	
A2 B1 C1	-0.0489
A2 B1 C2	-0.0303
A2 B1 C3	-0.0489
-----	
A2 B2 C1	0.0441
A2 B2 C2	0.0627
A2 B2 C3	0.0441

Der Interaktionseffekt  $A_1B_1$  beispielsweise ist davon abhängig, welche Ausprägung die andere Variable C besitzt. Er ist -0.0441, wenn die andere Variable C die Merkmalsausprägung  $C_1$  besitzt. Er ist -0.0627, wenn C die Ausprägung  $C_2$  besitzt.

Es ist also nicht möglich die eigenständigen Interaktionseffekte  $AB_{ij}$  zu ermitteln, wenn ungleiche Zellenhäufigkeiten vorliegen. Das Modell der „fitting constants I“ besitzt dafür keine Lösung.

Im Beispielpogramm "Gleich3.Alm" wird eine Varianzanalyse mit *gleichen* Zellenhäufigkeiten gerechnet. Das Programm ist in der Almo-Programmiersprache geschrieben.

Man findet es durch Klick auf den Knopf "alle Progs" am Oberrand des Almo-Fensters. Der Benutzer muss noch Verfahren auf "fitting constants I" umschalten. Für z.B. die 2-er Interaktion AB liefert das Programm folgende Ausgabe

```
Effekte von AB
-----
A1 B1 C1      -0.9083
A1 B1 C2      -0.9083
A1 B2 C1       0.9083
A1 B2 C2       0.9083
A2 B1 C1       0.9917
A2 B1 C2       0.9917
A2 B2 C1      -0.9917
A2 B2 C2      -0.9917
A3 B1 C1      -0.0833
A3 B1 C2      -0.0833
A3 B2 C1       0.0833
A3 B2 C2       0.0833
```

Man erkennt, dass die 3. Variable C keine Rolle spielt. So ist z.B. A1B1  $-0.9083$ , egal ob C die Ausprägung C1 oder C2 besitzt, A1B2 ist  $0.9083$  egal ob C1 oder C2 usw. Tatsächlich liegen also bei *gleichen* Zellenhäufigkeiten keine "wechselnden" Werte vor.

Beim Verfahren der "weighted squares of means" treten keine "wechselnde Werte" auf. Es ist also bei Analysen mit 3 und mehr Variablen, bei denen Interaktionen gebildet werden sollen, vorzuziehen. Das Verfahren der fitting constants I sollte nur verwendet werden, wenn keine Interaktionen benötigt werden oder nur für ein Modell mit zwei Faktoren A, B und deren Interaktion AB.

### ***P20.9.1.2 Verschiedene Ergebnisse für Abweichungs-Quadratsummen und Korrelationsmatrix***

Die erklärten Streuungen, die Regressionskoeffizienten der Kovariaten und die Effekte der nominalen Variablen sowie deren Standardabweichungen sind natürlich verschieden, wenn man ein Mal eine Analyse mit Abweichungs-Quadratsummenmatrix und ein anderes Mal eine Analyse mit der Korrelationsmatrix rechnet (Siehe P20.8.5). Die F- bzw. t-Werte, die Signifikanzen und die (partiellen) Korrelationskoeffizienten sind aber gleich. Dies gilt jedoch nicht, wenn in den Daten Kein\_Wert- Fälle enthalten sind und das paarweise Ausscheiden erfolgte, d.h. mit der (voreingestellten) Option KW\_Behandlung=1 (P20.8.5) gerechnet wurde. Wird das vollständige Ausscheiden (KW\_Behandlung=3) vorgenommen oder wird MATRIX= QUASIKORRELATION und KW\_BEHANDLUNG=1 gesetzt, dann sind die Ergebnisse in diesen Koeffizienten wieder gleich. Der ALMO-Benutzer wird natürlich wissen wollen, welche Vorgehensweise die "wahren" Ergebnisse erbringt, die Korrelationsmatrix mit paarweisen Ausscheiden oder die Quasikorrelationsmatrix bzw. die Abweichungs-Quadratsummenmatrix mit paarweisem Ausscheiden.

Unsere Empfehlung wäre, dass man das vollständige Ausscheiden wählt, den Datenverlust also in Kauf nimmt, wenn die Ergebnisse zu stark divergieren.

### **P20.9.2 Ausgabe bei Regressionsanalyse**

In den Maskenprogrammen Prog20mx und Prog20mo sind die Eingabe-Boxen für die abhängige und die unabhängigen Variablen folgende:

**Analyse-Variable: Abhängige Variable** **Hilfe**

Erlaubt sind:

1. Eine oder mehrere quantitative 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

**Leistung**

---

ordinale abhängige Zielvariable **Hilfe**

---

nominale abhängige Zielvariable **Hilfe**

**Analyse-Variable: Unabhängige Variable** **Hilfe**

nominale unabhängige Variable **Hilfe**

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

∅ =keine Interaktionen bilden

paarweise Vergleiche (Kontraste) für die nominalen unabhängige Variablen rechnen

---

quantitative unabhängige Variable **Hilfe**

**Alter, Einkommen, Kinderzahl**

---

ordinale unabhängige Variable **Hilfe**

Der Benutzer kann auch die beiden Programme Prog20mf oder Prog20mc, die eigens für die Regressionsanalyse entwickelt wurden, verwenden.

Klicken Sie auf Verfahren/Regressionsanalyse.

Auch die Syntax-Programme Prog20k oder Prog20j können verwendet werden. Klicken Sie auf das Menü „Almo/Liste aller Almo-Programme“ oder „Almo/Masken-Programm laden“.

Standardmäßig wird in den beiden Maskenprogrammen Prog20mx und Prog20mo mit der Abweichungs-Quadratsummen-Matrix gerechnet. Demzufolge werden auch nicht-standardisierte Regressionskoeffizienten ermittelt.

Sollen standardisierte Regressionskoeffizienten ermittelt werden, dann muss mit Prog20mo

gerechnet werden und dort die Optionsbox "Streuungsmatrix" geöffnet werden und auf "Korrelation" eingestellt werden. Almo errechnet dann die Korrelationsmatrix und bildet die standardisierten Regressionskoeffizienten.

Bei den beiden speziellen Kurzprogrammen ist bereits auf "Korrelation" und damit auf standardisierte Regressionskoeffizienten eingestellt.

Das Modell, das wir rechnen, ist folgendes:

$$V_5' = \beta_1 * V_6 + \beta_2 * V_7 + \beta_3 * V_8$$

Wir wollen die wichtigsten Ergebnisse vorwegnehmen:

	partieller Regr.Koeff	partieller. Korr.Koeff.
V6	$\beta_1 = 0.0329$	0.0383
V7	$\beta_2 = -0.0274$	-0.0357
V8	$\beta_3 = -0.1280$	-0.1687

Der multiple Korrelationskoeffizient beträgt 0.1737 und ist mit  $p=0.6274$  bzw.  $(1-p)100 = 37.2555\%$  signifikant. Das Modell muss also verworfen werden. Die Ausgabe von ALMO für dieses Programm entspricht in seinem 1. Teil, bis hin zur Quadratsummen-Matrix, der Ausgabe bei der Varianzanalyse. Wir bringen diesen 1. Teil deswegen hier nicht. Der 2. Teil der Ausgabe ist dann folgender:

Diagonalglieder der Choleskymatrix  
zur Ermittlung und zum Ausschluss  
linearer Abhaengigkeiten

Unabhaengige Variable

```
V6      297.737705
V7      368.564916
V8      385.697138
```

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

Keine Variable wird eliminiert

Alle im Folgenden angegebenen Streuungen und erklarten Streuungen sind  
Abweichungsquadratsummen

```
Gesamtstreuung                222.196721
```

Koeffizienten fuer Gesamt-Modell

```
Durch alle unabh. Variable
erklarte Streuung                6.702408
Fehlerstreuung                  215.494314
multipler Korrelat.koeff.        0.173679
F-Wert f. erklarte Streuung      0.590947
Freiheitsgrade Nenner =    3
                Zaehler=    57
Signifikanz: p                    0.627445
Signifikanz: (1-p)*100           37.255464 %
Teststaerke von F                0.165058
```

Koeffizienten fuer quantitat./ordinale Variable aus univariater Analyse  
hinsichtlich der abhaeng. Var. V5 Leistung

Variable	standard. Regr. koeff.	Regr. koeff.	Standard fehler	95% Konfidenz- bereich nach oben u.unten	erklärte Streuung	part. Korrel.
V6 Alter	0.0381	0.0329	0.1139	0.2282	0.3159	0.038
V7 Einkommen	-0.0357	-0.0274	0.1017	0.2037	0.2748	-0.036
V8 Kinderzahl	-0.1695	-0.1280	0.0990	0.1983	6.3149	-0.169

Variable	F-Wert	Signifikanz p	(1-p)100	df1	df2	Test- staerke
V6 Alter	0.084	0.775	22.51	1	57	0.0593
V7 Einkommen	0.073	0.790	21.01	1	57	0.0581
V8 Kinderzahl	1.670	0.201	79.85	1	57	0.2462

**\*\*\* Erläuterung:**

Almo gibt in der 1. Spalte die standardisierten und in der 2. Spalte die nicht-standardisierten Regressionskoeffizienten aus. Die Spalte "Standardfehler" und die Spalte "95% Konfidenzbereich" beziehen sich auf die nicht-standardisierten Regressionskoeffizienten. Wäre im Programm Prog20mo die Optiosbox "Streuungsmatrix" geöffnet worden und "Korrelation" oder "Quasi\_Korrelation" selektiert worden, dann rechnet Almo das ALM bezogen auf die Korrelationsmatrix. Die nicht-standardisierten Regressionskoeffizienten sind in diesem Fall gleich den standardisierten. Siehe dazu in Abschnitt P20.8.1.1 die Erläuterungen zur Eingabe-Box "Streuungsmatrix". Auch die Art der "erklärten Streuung" hängt davon ab, welche Streuungsmatrix analysiert wurde. Betrachten wir die Koeffizienten der unabhängigen Variablen V8. Ihr nicht-standardisierter Regressionskoeffizient ist  $b = -0.1280$ . Der Standardfehler des Regressionskoeffizienten ist  $\gamma = 0.0990$ . Wir können also ein Vertrauensintervall mit Konfidenzniveau 95 % bilden - indem wir  $\gamma$  mit 2 (genauer: 1.96) multiplizieren.

$$b (+/-) 1.96 * \gamma = -0.1280 (+/-) 1.96 * 0.0990 = -0.1280 (+/-) 0.1983$$

Dieser Konfidenzbereich von 0.1983 ist oben in der 3. Spalte angegeben. Dieses Vertrauensintervall schließt den Wert .0 mit ein - also ist der Regressionskoeffizient auf dem Konfidenzniveau 95 % nicht signifikant. Der F-Wert für den Regressionskoeffizienten beträgt  $F=1.67$  mit einer Signifikanz (1-p)100 von 79.85 %. Da bei  $df_1=1$  gilt:  $t=\sqrt{F}$  können wir auch den t-Wert bestimmen. Er ist  $t=1.292$ . Der Regressionskoeffizient liegt also in seinem Sicherheitsniveau - was wir aus der Betrachtung des Vertrauensintervalls schon wissen - unter der üblichen 95 %-Marke.

Die Variable V8 (=Kinderzahl) erklärt von der Gesamtstreuung der abhängigen Variablen (von 222.1967) 6.3149 Einheiten. Diese erklärte Streuung ist eine partielle erklärte Streuung. Würde die Kinderzahl aus der Analyse herausgenommen werden, dann würde sich die Fehlerstreuung der abhängigen Variablen um 6.3149 Einheiten erhöhen. Wir können auch so formulieren: Die durch die Kinderzahl erklärte Streuung ist von den Wirkungen befreit, die daher rühren, dass die Kinderzahl mit den anderen unabhängigen Variablen (Alter und Einkommen) korreliert. In diesem Sinne ist auch der Regressionskoeffizient und der Korrelationskoeffizient ein "partieller" Koeffizient.

Aus der erklärten Streuung ergibt sich schließlich der partielle (Produkt-Moment-) Korrelationskoeffizient von  $r = -0.1687$ , der wie der Regressionskoeffizient mit 79.85 % signifikant ist. Zur Berechnung des Korrelationskoeffizienten aus der erklärten Streuung siehe P20.6.3.

Koeffizienten fuer Konstante

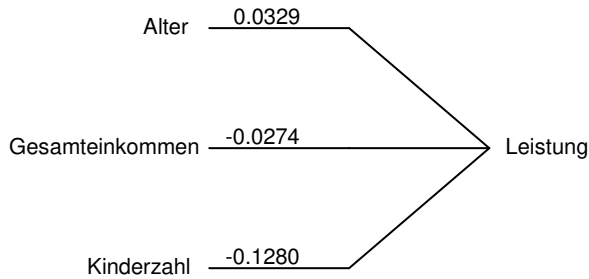
```

hinsichtlich der abh.Variablen   V5 Leistung
Effekt (Regressionskoeffizient)   4.453012
=====

```

\*\*\* **Erläuterung:** Wenn wir das allgemeine lineare Modell auf die Korrelationsmatrix anwenden, dann ist die Konstante immer 0. Das gilt für alle Submodelle des allgemeinen linearen Modells also auch für die Varianzanalyse und die Kovarianzanalyse.

Almo zeichnet folgendes Flussdiagramm der Regressionskoeffizienten



***P20.9.2.2 Maskenprogramm: 2D- Streudiagramm und Regressionsgerade für eine Analyse mit einer unabhängigen Variablen Prog02mb***

Der Benutzer findet das Programm durch Klick auf Verfahren/Regressionsanalyse.

Prog02mb.Msk  
 Regression mit 2D-Streudiagramm  
 fuer eine Analyse mit 1 unabhangigen Variablen

Die Daten werden in einem xy-Koordinatensystem als Punkte grafisch dargestellt

Beispiel: Untersuchungspersonen werden nach Alter und Leistung in einem Diagramm dargestellt. Almo erzeugt etwa folgendes Diagramm

Leistung  
y-Achse

Alter (x-Achse)

Das Programm ist nur dann sinnvoll, wenn die Zahl der Untersuchungseinheiten (=der Punkte im Streudiagramm) nicht zu gross ist. Grenze ungefahr n=50. Ist sie groer, dann besteht die Moglichkeit, nur eine Stichprobe zu verwenden

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

- 1    
 Vereinbare Variable=  ;
- 2  Option: Weitere Vereinbarungen - nur wenn Almo dazu auffordert
- 3    
  
      zeige = Namensdatei in Output zeigen  
                                  leer = nicht
- 4    
  
  
  
 erzeuge zusatzliche Namensfelder
- 5   bei Datei-Problemen  
  
      Format der Daten        
      der Datensatz enthalt diese Variablen  
                                  Bei Format DIREKT schreiben Sie: alle\_U
- 6  Wenn Dateiformat FIX oder Nicht-Standard-FREI

7

**Analyse-Variable**

unabhängige quantitative Variable  
maximal 2 möglich

---

abhängige quantitative Variable  
nur 1 möglich

Die Datenpunkte können nach einer Gruppierungsvariablen verschieden markiert werden

8

Option: Gruppierungsvariable

9

Option: Ein- und Ausschliessen von Untersuchungseinheiten

10

Option: Umkodierungen und Kein-Wert-Angaben

11

**Was soll gezeichnet werden ?**

1 = Regressionsgerade bzw. -ebene zeichnen  
0 = nicht, nur Punktwolke zeichnen

12

Grafik-Optionen

13

**Programmende**

## Erläuterung zu den Eingabe-Boxen

### Eingabe-Box: Vereinbarungen

Siehe P0.1.

### Eingabe-Box: Option: Weitere Vereinbarungen

Siehe P0.2.

### Eingabe-Box: Datei der Variablennamen

Siehe P0.3.

### Eingabe-Box: Freie Namensfelder

Siehe P0.3.

### Eingabe-Box: Datei aus der gelesen wird

Siehe P0.4.

### Eingabe-Box: Wenn Dateiformat FIX oder Nicht-Standard-FREI

Siehe P0.4.

### Eingabe-Box: Analyse-Variable

Analyse-Variable

unabhängige quantitative Variable  
maximal 2 möglich

↔ □ Alter

---

abhängige quantitative Variable  
nur 1 möglich

↔ □ Leistung

Werden 2 unabhängige quantitative Variable angegeben, dann zeichnet Also eine Regressions-Ebene. Siehe dazu nachfolgendes Programm.

### Eingabe-Box: Option: Gruppierungsvariable

↓

Option: Gruppierungsvariable

Die Datenpunkte können nach einer Gruppierungsvariablen verschieden markiert werden

Optionsbox geöffnet:



Der Zweck der Gruppierungsvariable ist folgender: Die Datenpunkte in der Grafik können nach der Gruppierungsvariablen verschieden markiert werden.

Die Gruppierungsvariable muss ganzzahlige Werte besitzen und fortlaufend, ohne Lücke, kodiert sein. Ist dies nicht der Fall, dann muss sie in nachfolgender Eingabe-Box

"Umkodierungen und Kein-Wert-Angaben"

auf Ganzzahligkeit und lückenlos fortlaufende Werte umkodiert werden.

**Eingabe-Box: Option: Ein- und Ausschließen von Untersuchungseinheiten**  
Siehe P0.7.

**Eingabe-Box: Option: Umkodierungen und Kein-Wert-Angaben**  
Siehe P0.5.

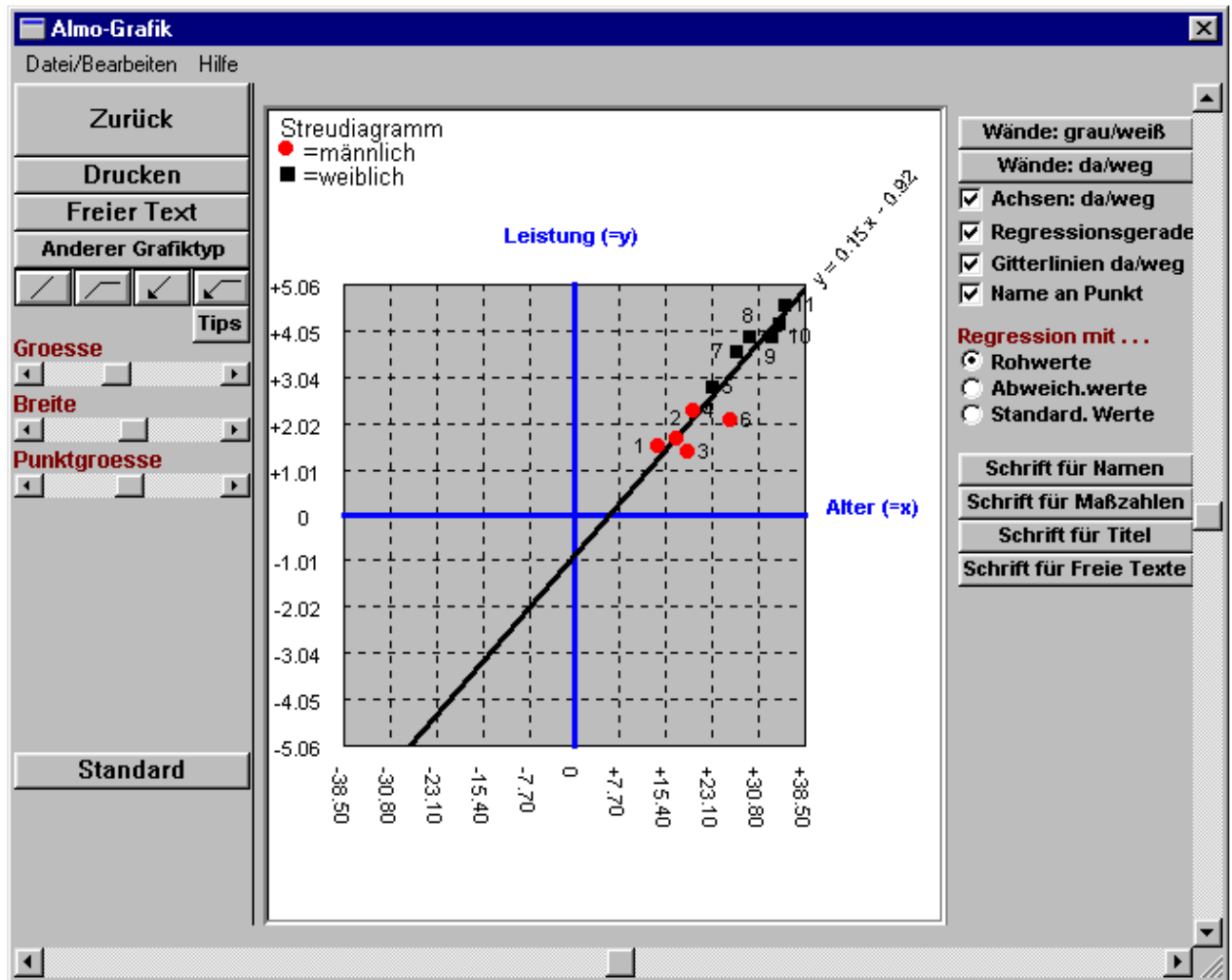
**Eingabe-Box: Was soll gezeichnet werden ?**



Die Regressionsgerade kann weg gelassen werden. Nur die Punktwolke wird dann gezeichnet.

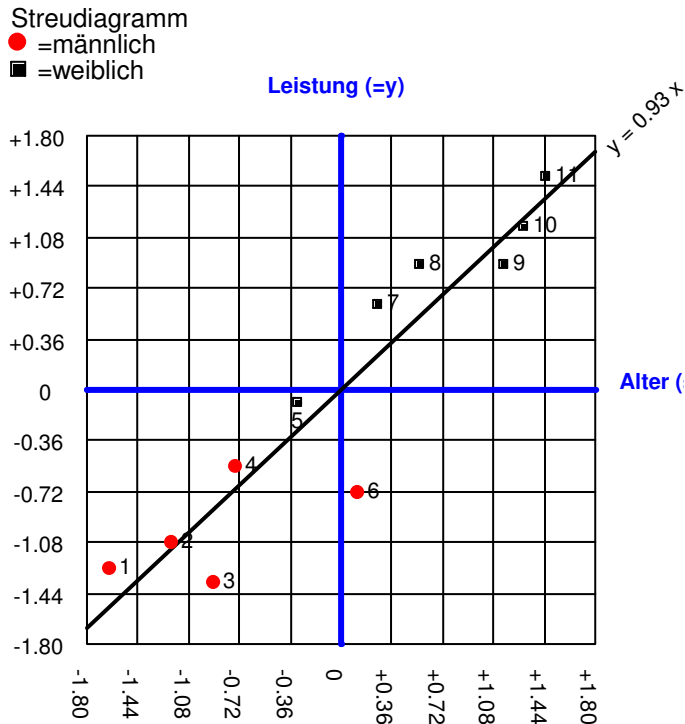
Die Daten, die in das Programm eingegeben wurden, sind folgende:

Alter	Leistung	Geschlecht
14	1.5	1
17	1.7	1
19	1.4	1
20	2.3	1
23	2.8	2
26	2.1	1
27	3.6	2
29	3.9	2
33	3.9	2
34	4.2	2
35	4.6	2



ALMO zeichnet die Daten als Punkte in ein Diagramm ein und legt eine Regressionsgerade durch die Punktwolke. Die Gleichung für diese Regressionsgerade wird in die Grafik eingeschrieben. Sie lautet:  $y = 0.15 * x - 0.92$ .

Wenn der Benutzer auf der rechten Seite des Grafikfensters den Radio-Button "Regression mit standardisierten Werten" wählt, dann werden die Daten standardisiert und die Grafik in folgender Weise verändert.



***P20.9.2.3 Maskenprogramm: 3D- Streudiagramm und Regressionsebene für eine Analyse mit 2 unabhängigen Variablen Prog02mc***

Das Programm ist identisch mit dem vorausgehend dargestellten Prog02mb.  
 Die Daten, die in diesem Maskenprogramm gerechnet werden, sind folgende:

Alter	Kraft	Leistung
24	0.10	1.5
15	0.08	1.7
19	-0.05	1.4
25	-0.14	2.8
23	-0.11	3.1
26	0.09	2.0
17	0.12	3.6
19	0.06	3.9
20	0.07	3.0
25	0.10	4.2
28	0.08	4.6

Wird das Maskenprogramm gerechnet, dann entsteht ein Output, der im Wesentlichen nur aus einem Grafikknopf besteht. Wird auf diesen geklickt, dann entsteht folgendes:

Alter und Kraft sind die unabhängigen Variablen. Leistung ist die abhängige Variable. Die Eingabe-Box „Analyse-Variable“ sieht folgendermaßen aus:

**Analyse-Variable**

unabhängige quantitative Variable  
maximal 2 möglich

← □ □ **Alter, Kraft**

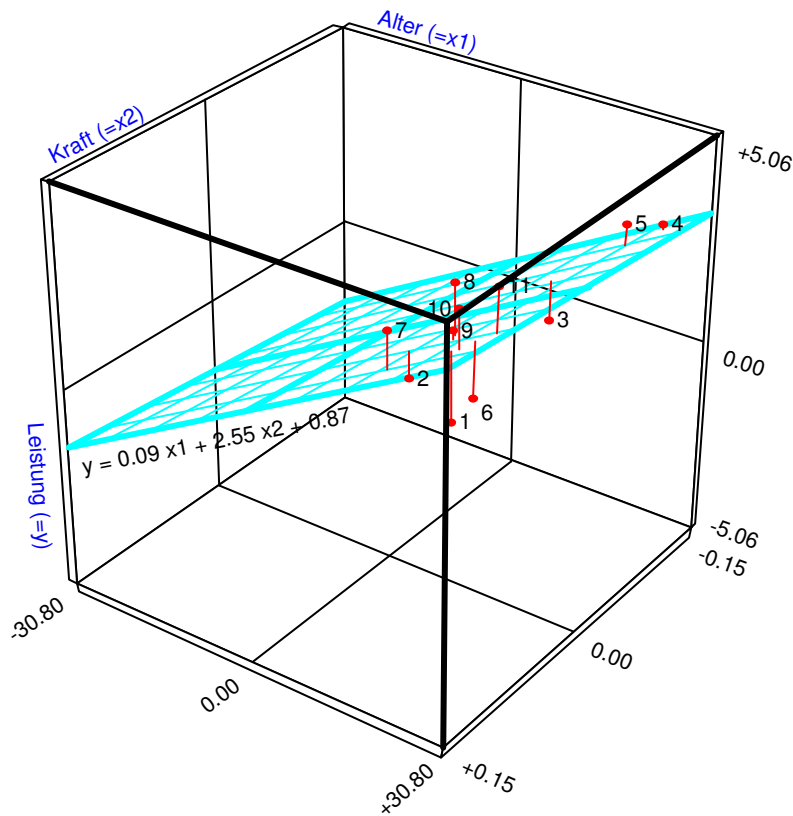
---

abhängige quantitative Variable  
nur 1 möglich

← □ □ **Leistung**

ALMO erstellt folgendes 3-dimensionales Streudiagramm:

Streudiagramm



Bei 2 unabhängigen Variablen entsteht keine Regressionsgerade mehr, sondern eine Regressionsebene. Diese ist in obiger Grafik eingezeichnet. Ihre Gleichung ist in der Grafik eingetragen. Sie lautet:  $y = 0.09 \cdot x_1 + 2.55 \cdot x_2 + 0.87$

Auch hier besteht wieder die Möglichkeit durch Mausklick auf eine Regression mit Abweichungswerten oder standardisierten Werten umzuschalten.

### P20.9.3 Ausgabe bei Kovarianzanalyse

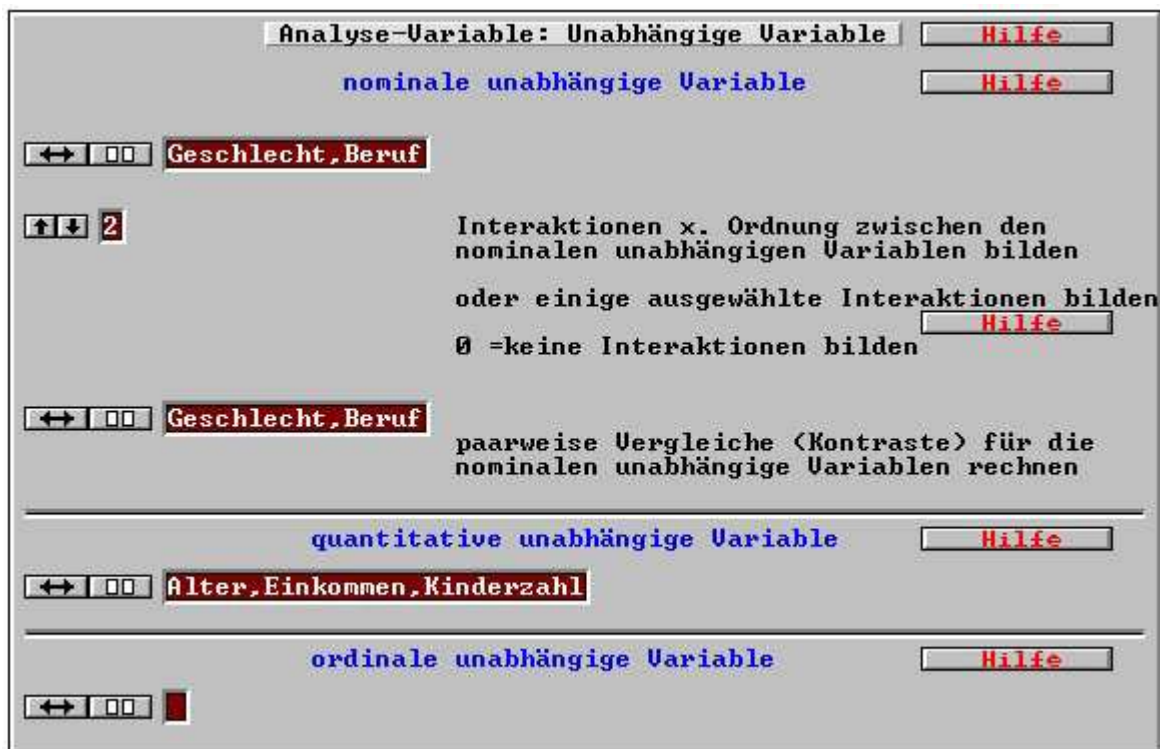
Wir wollen die beiden vorausgegangenen Beispiele nun zur Kovarianzanalyse zusammenbringen. Wir verwenden wieder unsere Datenmatrix aus P20.9.0. Die abhängige Variable ist V5.

Als unabhängige nominale Variable wollen wir aus unserer Datenmatrix V1 und V3 verwenden.

Als unabhängige quantitative Variable (=Kovariate) verwenden wir V6,7,8. Wir rechnen mit ungleichen Zellenhäufigkeiten. Es wird mit Rohwerten gerechnet, also mit der Abweichungs-Quadratsummen-Matrix.

In den Maskenprogrammen Prog20mx und Prog20mo sind in die Eingabe-Boxen für die abhängige und die unabhängigen Variablen einzutragen:

The screenshot shows a dialog box titled "Analyse-Variable: Abhängige Variable" with a "Hilfe" button in the top right corner. The main text reads: "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". Below this, there are three sections for selecting variables: "quantitative abhängige Variable" with a selection box containing "Leistung"; "ordinale abhängige Zielvariable" with a selection box containing a red square; and "nominale abhängige Zielvariable" with a selection box containing a red square. Each section has a "Hilfe" button to its right.



Der Benutzer kann auch die beiden Maskenprogramme Prog20mh und Prog20ma, die eigens für die Kovarianzanalyse entwickelt wurden, verwenden. Klicken Sie auf Verfahren / Kovarianzanalyse.

Auch die Syntaxprogramme Prog20c, Prog20k oder Prog20j können eingesetzt werden. Siehe Abschnitt P20.8.2. Klicken Sie auf das Menü "Almo / Liste aller Almo-Programme".

Die Ausgabe, die ALMO liefert, setzt sich im Prinzip aus der Ausgabe für die Varianzanalyse und der Regressionsanalyse zusammen. Selbstverständlich sind die erklärten Streuungen, Effekte und Regressionskoeffizienten etwas verschieden, weil die quantitativen unabhängigen Variablen (= die Kovariaten) und die nominalen unabhängigen Variablen leicht miteinander korrelieren und sich gegenseitig etwas an Erklärungsfähigkeit wegnehmen.

Zusätzlich wird von ALMO die Erklärungsfähigkeit der Gruppen der quantitativen und der nominalen Variablen ausgegeben:

```

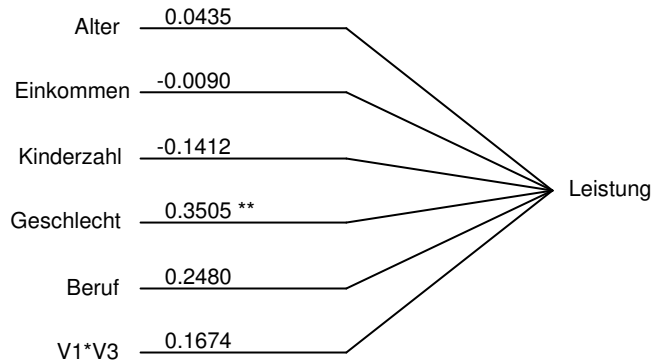
Durch die quantitativen/ordinalen Variablen insgesamt
erklärte Streuung                3.960829
multipler partieller Korrelat.koeff.  0.148200
F-Wert f. erklärte Streuung        0.389247
Freiheitsgrade Nenner =    3
                Zaehler=    52
Signifikanz: p                    0.764236
Signifikanz: (1-p)*100            23.576354 %
Teststaerke von F                  0.121782
=====

Durch die nominalen Variablen
und ihre Interaktionen insgesamt
erklärte Streuung                39.116759
multipler partieller Korrelat.koeff.  0.426053
F-Wert f. erklärte Streuung        2.306497
Freiheitsgrade Nenner =    5
                Zaehler=    52
Signifikanz: p                    0.056874

```



Partielle  
Korrelationskoeffizienten

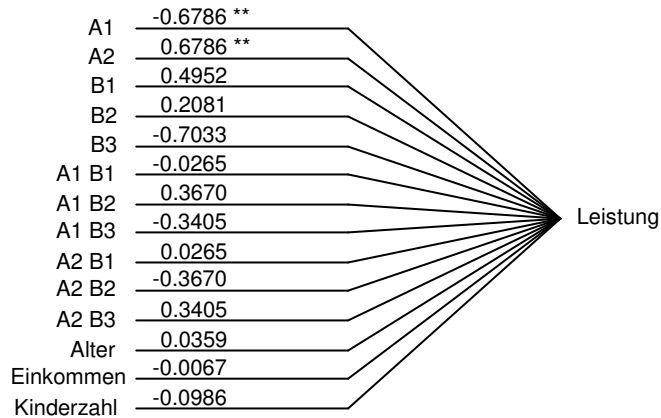


Zusammenfassung: Effekte und Regressionskoeffizienten  
und ihre Signifikanzen  
hinsichtlich der abhaengigen Variablen  
Leistung

	Effekte Regress.koeff	Signifikanz (1-p) *100
A1 männlich	-0.678572	99.078188
A2 weiblich	0.678572	99.078188
B1 Arbeiter	0.495232	81.677925
B2 Angestellt	0.208069	46.903505
B3 Selbständi	-0.703302	92.597197
A1 B1	-0.026518	5.839835
A1 B2	0.367000	73.457054
A1 B3	-0.340482	63.724037
A2 B1	0.026518	5.839835
A2 B2	-0.367000	73.457054
A2 B3	0.340482	63.724037
Alter	0.035945	24.405508
Einkommen	-0.006716	5.390303
Kinderzahl	-0.098593	69.142907
Konstante	4.219143	-

Almo zeichnet folgendes Flussdiagramm der Effekte und Regressionskoeffizienten

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



Almo zeichnet die linearen Funktionen der unabhängigen quantitativen Variablen Alter, Einkommen, Kinderzahl hinsichtlich der abhängigen Variablen Leistung. Dabei wird nach der Kombination der nominalen Variable Geschlecht und Beruf als Gruppierungsvariable gezeichnet. Siehe hierzu die ausführliche Darstellung in Abschnitt P20.8.1.1, Eingabe-Box "Grafik-Optionen".

In der Ergebnisliste ist folgender Grafikknopf enthalten

```

Lineare Funktion fuer
abhaengige Variable: U5 Leistung
unabhaengige Variable: U6 Alter
Gruppierungsvariable: U1 Geschlecht: 1.Auspraegung: männlich
mit
U3 Beruf: 1.Auspraegung: Arbeiter
  
```

Hilfe

Grafik



### P20.9.3.1 Prognosewerte und Residuen

Der Leser gehe noch einmal zurück zu P20.9.1. Residuen sind die Differenz zwischen prognostiziertem Wert  $y'$  und tatsächlichem  $y$ -Wert. In unserem Beispiel in P20.9.1 haben wir für  $y'$  einen Wert von 5.5455 ermittelt. Wenn der tatsächliche  $y$ -Wert 5.7455 ist, dann ist das Residuum  $e=0.2$ . Diese Residuen können für verschiedene Analysen weiter verwendet werden. Sie können z.B. auf Normalverteilung getestet werden.

Mit dem Maskenprogramm Prog20mo aus Abschnitt P20.8.0 können Prognosewerte und Residuen errechnet werden. Wir betrachten die Eingabe-Box „Prognosewerte und Residuen berechnen“ aus diesem Programm.

Loesche wieder diese Box

**Prognosewerte und Residuen ermitteln**

↑ ↓ !

0 =keine Prognosewerte und Residuen  
1 =ermitteln und ausgeben  
2 =ermitteln aber nur Prognoseerfolg ausgeben  
(nur wenn abhängige Variable nominal ist)

↔ !

Prognosewerte und Residuen nur ermitteln für  
eine Stichprobe von ca x %

↔ ! "C:\Almo7\Progs\Residuen.fre"

Schreibe die Prognosewerte und Residuen  
in eine Datei im Format FREI  
Wenn nicht, dann löschen Sie den Dateinamen  
durch Klick auf den doppelköpfigen Pfeil

Almo rechnet Prognosewerte und Residuen, wenn das Verfahren „weighted squares of means“ ist.

Ist das Verfahren „fitting constants I“, dann werden Prognosewerte und Residuen nur ermittelt, wenn keine Interaktionen in die Analyse eingeschlossen werden – oder wenn nur 2 unabhängige nominale Variable und ihre Interaktionen vorhanden sind. Diese Einschränkung ist notwendig, da bei höheren als 2-er Interaktionen wechselnde Werte auftreten. Siehe P20.6.5.1 und P20.9.1.1. Die Prognosewerte und Residuen sind für beide Verfahren exakt dieselben. Für das Verfahren „fitting\_constants\_II“ werden keine Prognosewerte und Residuen gerechnet.

#### Erläuterung:

*Eingabefeld 1:* Geben Sie 1 ein, wenn Prognosewerte und Residuen ermittelt und ausgegeben werden sollen.

*Eingabefeld 2:* Geben Sie eine Zahl zwischen 0 und 100 ein, z. B. 80. Almo berechnet dann die Prognosewerte und Residuen nur für eine Stichprobe von 80% der Datensätze.

*Eingabefeld 3:* Wenn Sie die Prognosewerte und Residuen in eine Datei speichern wollen, dann geben Sie hier den Dateinamen (mit Pfad) an. Wenn das Eingabefeld leer bleibt, dann nimmt Almo an, dass nicht gespeichert werden soll.

Wir wollen am Beispiel der Kovarianzanalyse in P20.9.3 zeigen, wie mit Prog20mo Prognosewerte und Residuen berechnet werden.

Die Kovarianzanalyse hat uns folgende Ergebnisse geliefert (gerechnet mit dem Verfahren der „weighted squares of means“):

Zusammenfassung: Effekte und Regressionskoeffizienten  
und ihre Signifikanzen  
hinsichtlich der abhaengigen Variablen  
Leistung

	Effekte Regress.koeff	Signifikanz (1-p) *100
	-----	
A1 männlich	-0.678572	99.078188
A2 weiblich	0.678572	99.078188
B1 Arbeiter	0.495232	81.677925
B2 Angestellt	0.208069	46.903505
B3 Selbständi	-0.703302	92.597197
A1 B1	-0.026518	5.839835
A1 B2	0.367000	73.457054
A1 B3	-0.340482	63.724037
A2 B1	0.026518	5.839835
A2 B2	-0.367000	73.457054
A2 B3	0.340482	63.724037
Alter	0.035945	24.405508
Einkommen	-0.006716	5.390303
Kinderzahl	-0.098593	69.142907
Konstante	4.219143	-

Almo erzeugt folgende Ausgabe (gekürzt):

```

***** MITTEILUNG
Almo gibt die Prognosewerte und die Residuen in die Datei
"C:\Almo7\Progs\Residuen.fre"
mit Doppelklick auf den Namen laden Sie die Datei !!

```

```

fuer folgende Variable
V5 |
Leistung |

```

Dabei werden zuerst die Prognosewerte ausgegeben  
dann anschliessend in der gleichen Zeile die Residuen

Datensatz	tatsaechlicher Wert in der abaengigen Variablen V5 Leistung	prognostizierter Wert in der abaengigen Variablen V5 Leistung	Residuen (Differenz)  V5 Leistung
1	4.00000	3.74526	0.25474
2	5.00000	3.51884	1.48116
3	4.00000	3.96243	0.03757
4	2.00000	3.82118	-1.82120
5	4.00000	3.42801	0.57199
6	4.00000	3.40131	0.59869
7	2.00000	3.87160	-1.87160
8	4.00000	4.13396	-0.13400
9	3.00000	1.90103	1.09897
10	5.00000	2.24200	2.75800
.	.	.	.
.	.	.	.
.	.	.	.

Mittelwert und Standardabweichung der Residuen

	Mittelwert	Standardabweichung
V5 Leistung	2.72754e-006	1.70042

Der Mittelwert der Residuen ist (was er auch sein sollte) nahe 0.

Mit Prog 4 können nun die Residuen auf Normalverteilung überprüft werden. Wir verwenden dazu das Maskenprogramm Pro04m2. Der Benutzer findet dieses Programm durch Klick auf Verfahren / Häufigkeitsverteilung oder durch Klick auf das Menü "Almo / Liste aller Almo-Programme".

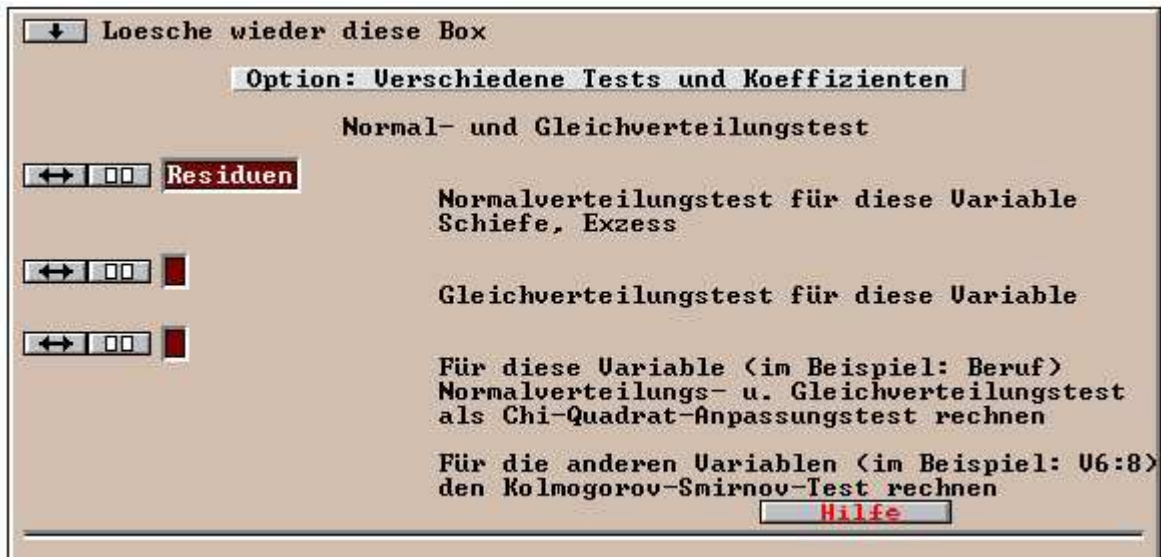
Das Programm ist im Almo-Dokument N3. 1a „Eindimensionale Tabellierung“, Abschnitt P4.4.1 beschrieben.

Wir zeigen hier nur, wie folgende Eingabe-Boxen dieses Programms auszufüllen sind.

The image shows four sequential input panels from the Pro04m2 program:

- Panel 1: Freie Namensfelder** (Free Name Fields). It contains two text boxes: "Name 1=Prognoswert;" and "Name 2=Residuen;". Below them is a button labeled "erzeuge zusätzliche Namensfelder" (create additional name fields).
- Panel 2: Datei aus der gelesen wird** (File to be read). It shows a file path "C:\Almo7\Progs\Residuen.fre" and a format dropdown set to "frei". A note says "Format der Daten" (Data format) and "der Datensatz enthält diese Variablen" (the dataset contains these variables). A note below says "Bei Format DIREKT schreiben Sie: alle\_U" (For format DIRECT write: all\_U).
- Panel 3: Wenn Dateiformat FIX oder Nicht-Standard-FREI** (If file format is FIX or non-standard-FREI). This panel has a dropdown arrow pointing down and a "Hilfe" (Help) button.
- Panel 4: die auszuzählenden Variablen** (Variables to be counted). It has a text box containing "Prognoswert,Residuen" and a "Hilfe" (Help) button.

Die Optionsbox "Option: Verschiedene Tests und Koeffizienten" wird geöffnet und in ihrem oberen Teil folgendermaßen ausgefüllt:



Almo liefert einen längeren Output. Wesentlich ist folgender Teil:

```

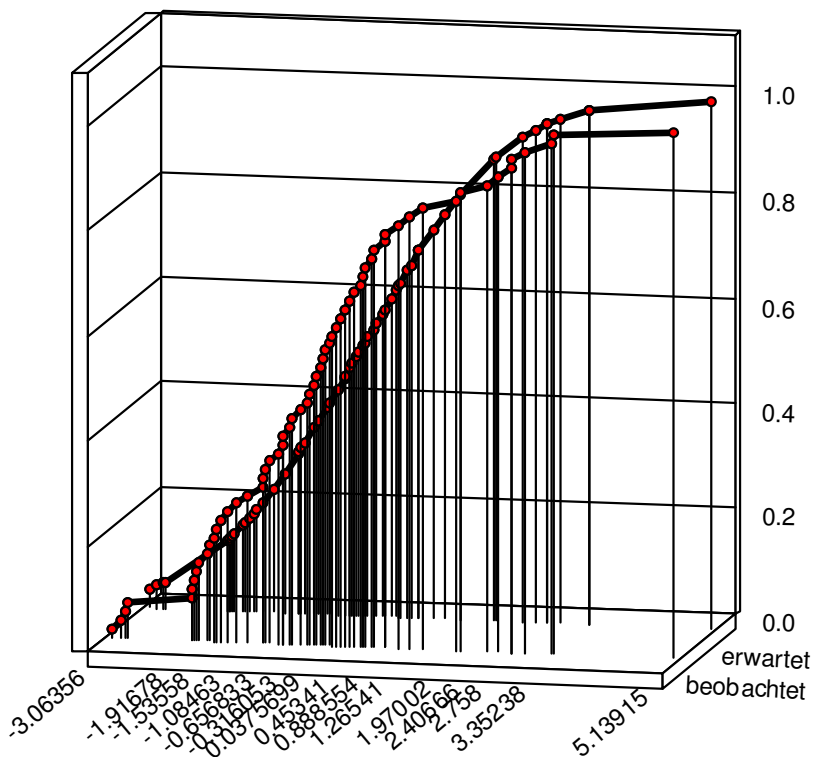
Wahrscheinlichkeit (1-p)*100
dass keine Normalverteilung 27.5724 %
(zweiseitiger Test)

```

Die Normalverteilung kann angenommen werden, wenn obiger Wert kleiner ca. 95 % ist.

Almo zeichnet dann noch folgende Grafik:

Erwartete und beobachtete  
kumulierte Haeufigkeiten



Die hintere Kurve ist die der erwarteten Werte, der Werte, die sich ergeben müssten, wenn die Residuen exakt normalverteilt wären. Es ist dies die Ogive.

Die vordere Kurve ist die der tatsächlichen Residuenwerte. Man erkennt, dass die beiden Kurven weitgehend übereinstimmen.

### Prognosewerte bei Verfahren "fitting constants I"

Wir haben ausgeführt, dass beim Verfahren der fitting constants I "wechselnde Werte" auftreten, sofern 3 und mehr nominale Variable und ihre Interaktionen in die Analyse einbezogen werden und sofern ungleiche Zellenhäufigkeiten vorliegen. Siehe dazu Abschnitt P20.9.1.1. Wir haben empfohlen, bei dieser Konstellation die fitting constants I nicht zu verwenden, dafür mit dem Verfahren der "weighted squares of means" zu rechnen.

Trotz dieser "wechselnden Werte" ist es jedoch möglich korrekte Prognosewerte zu ermitteln. Wir wollen die Berechnung an einem Beispiel vorführen.

Wir rechnen mit Prog20mo eine Analyse mit unseren Testdaten. Abhängige Variable ist V5 (Leistung). Die unabhängigen nominalen Variablen sind V1 (Geschlecht), V2 (Wohnort) und V3 (Beruf). Der Prognosewert für eine Person mit

```
Geschlecht = 1 (A1)
Wohnort    = 1 (B1)
Beruf      = 1 (C1)
```

Aus der Konstanten, sowie den Haupt- und Interaktionseffekten, die Almo liefert errechnen wir den Prognosewert für diese Person:

Konstante		3.8852	
Haupteffekt	A1	-0.5495	
Haupteffekt	B1	0.0417	
Haupteffekt	C1	0.4142	
2-er Interaktionseffekt	A1B1	0.0441	(wenn C1)
2-er Interaktionseffekt	A1C1	-0.0396	(wenn B1)
2-er Interaktionseffekt	B1C1	-0.2521	(wenn A1)
3-er Interaktionseffekt	A1B1C1	0.2059	
Summe (=Prognosewert)		3.7499	

Wird mit mehr als 4 Kommastellen gerechnet, dann entsteht 3.75. Dies ist - wie es sein muß - der Mittelwert der Variablen V5 (Leistung) in der Zelle A1 B1 C1.

**BEACHTEN:** In Almo ist die Berechnung von Prognosewerten bei "wechselnden Werten" nicht einprogrammiert. Werden Prognosewerte und Residuen benötigt, dann muß mit dem Verfahren der "weighted squares of means" gerechnet werden.

### Prognosewerte und Residuen bei nominaler abhängiger Variablen

Im Almo-Dokument Nr. 25 „Statistische Datenanalyse II“, Abschnitt P45.15.1.7 wird gezeigt, wie die Prognosewerte und Residuen zu interpretieren sind, wenn die abhängige Variable nominal-dichotom ist.

### P20.9.3.2 Gewichtete Kleinste-Quadrate

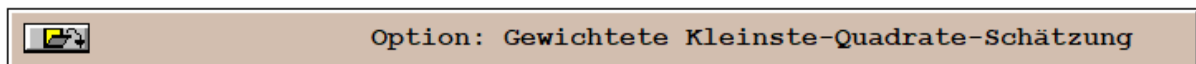
Ist die abhängige Variable nominal-dichotom, dann besteht modell-bedingte Varianzheterogenität. Diese kann durch die Methode der "gewichteten Kleinste-Quadrate" beseitigt werden. Siehe dazu die ausführliche Diskussion im Almo-Dokument Nr. 25 „Statistische Datenanalyse II“, Abschnitt P45.15.1.0.

Wir wollen zuerst den einfacheren Fall betrachten, dass die abhängige Variable nominal-dichotom ist und dann den etwas komplizierteren Fall, dass die abhängige Variable nominal-polytom ist.

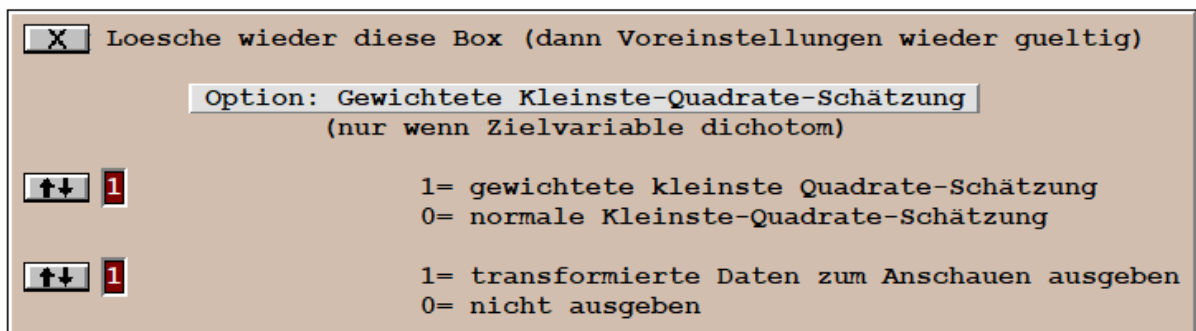
#### P20.9.3.2.1 Abhängige Variable ist nominal-dichotom

Betrachten wir folgendes Beispiel: Der Erfolg in einem Test (dichotom: Ja, Nein) ist abhängig von der Schulnote (quantitativ), dem Alter (quantitativ) und dem Geschlecht (nominal).

Im Maskenprogramm Prog20mo muss die Optionsbox „Gewichtete Kleinste-Quadrate-Schätzung“ geöffnet werden.



Nach der Öffnung sieht man folgendes:



Siehe hierzu die Erläuterungen zu dieser Eingabe-Box in Abschnitt P20.8.1.1.

Almo rechnet zuerst eine normale Analyse (in unserem Beispiel eine Kovarianzanalyse). Dabei werden folgende Effekte für die unabhängigen nominalen Variablen bzw. Regressionskoeffizienten für die unabhängigen quantitativen Variablen ermittelt:

Geschlecht A1	-0.0018
Geschlecht A2	0.0023
Note	0.0598
Alter	0.0402
Konstante	0.0027

Für jede Untersuchungseinheit wird nun die Wahrscheinlichkeit  $p$  des Erfolgs durch folgende lineare Wahrscheinlichkeitsfunktion prognostiziert:

$$p = -0.0018 \cdot A1 + 0.0023 \cdot A2 + 0.0598 \cdot \text{Note} + 0.0402 \cdot \text{Alter} + 0.0027$$

Für eine männliche Person in der Altersgruppe 8 und der Note 7 ergibt sich etwa:

$$p = -0.0018 \cdot 1 + 0.0023 \cdot 0 + 0.0598 \cdot 7 + 0.0402 \cdot 8 + 0.0027 = 0.7411$$

also eine Erfolgswahrscheinlichkeit  $p$  von 0.7411.

Wäre dieses  $p$  größer als 1.0, dann würde es von Almo auf 0.999 gesetzt werden. Wäre es kleiner 0 dann würde es auf 0.001 gesetzt werden. Unsere Erfahrung ist, dass bei empirischen Daten die Grenzen 0 und 1 selten überschritten werden.

Die Standardabweichung für diese Person ist dann:

$$s = \sqrt{(p * (1 - p))} = \sqrt{(0.7411 * (1 - 0.7411))} = 0.438$$

Sämtliche Variablenwerte (inklusive der Konstanten) dieser Person werden nun mit  $s$  dividiert. Der neue Datensatz wird zwischengespeichert. So wird mit jeder Person verfahren. Almo gibt diese zwischengespeicherten Daten zum Anschauen aus, wenn in der Optionsbox im 2. Eingabefeld „1“ eingesetzt wird. Die zwischengespeicherten Daten werden dann einer 2. Kovarianzanalyse unterworfen. Diese liefert die Ergebnisse der "gewichteten" Analyse:

```

Geschlecht A1  0.0062
Geschlecht A2 -0.0078
Note           0.0318
Alter          0.0139
Konstante     -1.3556
    
```

## Anhang: Kovarianzanalyse mit SPSS und Almo - ein Vergleich

Wir rechnen eine Standard-Kovarianzanalyse mit folgenden Variablen:

Abhängige quantitative Variable: V7 Einkommen (*gemessen in 9 Einkommensklassen*)

Unabhängige nominale Variable: V1 Geschlecht (*mit nur 2 Geschlechtern*)

V3 Beruf (*mit 3 Berufsgruppen*)

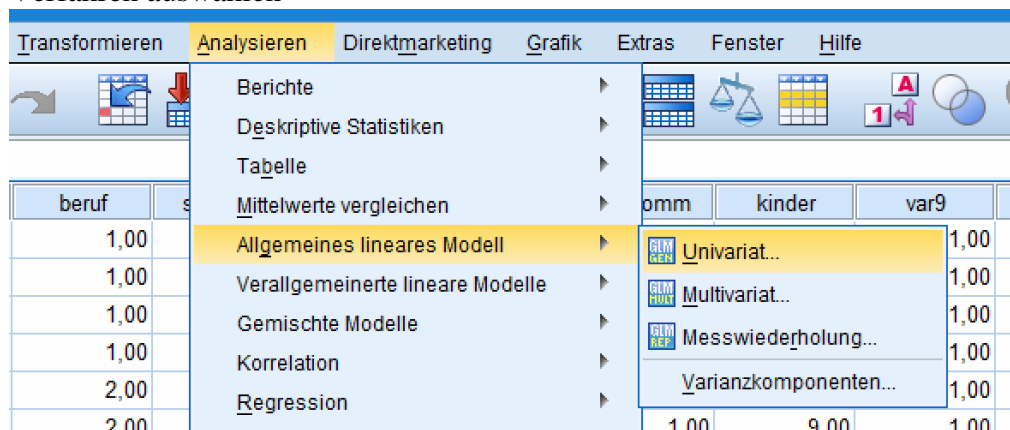
Unabhängige quantitative Variable: V5 Leistung (*in irgend einem Test*)

V6 Alter (*gemessen in 9 Altersklassen*)

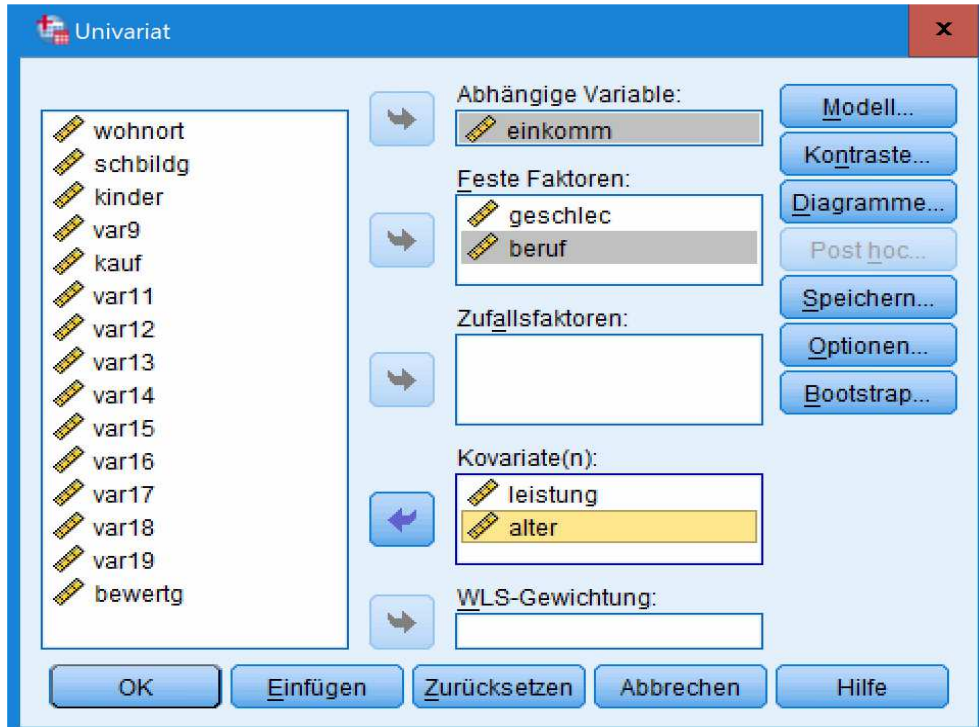
Verwendet werden die Daten "Testdat.fre" aus dem Almo-Ordner "TESTDAT". Dies sind konstruierte, nicht-empirische Daten.

Wir zeigen die **Eingabe in SPSS und die Ergebnis-Ausgabe** anhand der Version 22. SPSS verändert manchmal von einer Version zur nächsten diese beiden ohne dass sich substantiell etwas verändert. So sind z.B. die Randmittel in Version 25 als eigenständige Option über einen Schaltknopf aktivierbar.

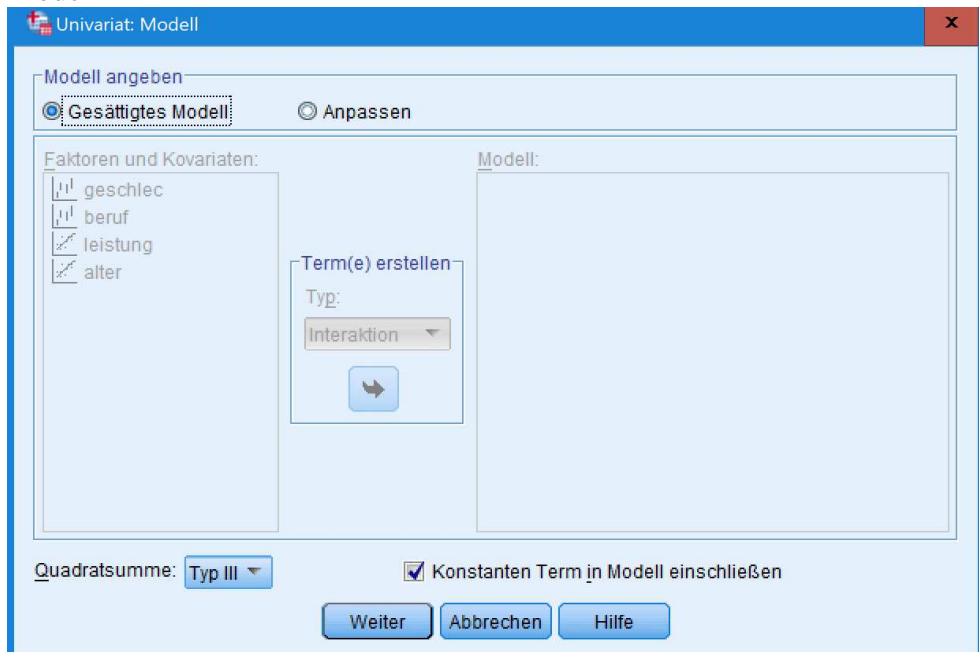
Verfahren auswählen



## Variable auswählen

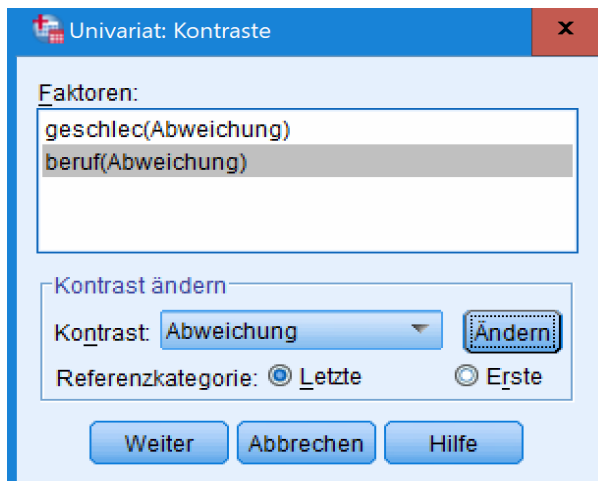


## Modell



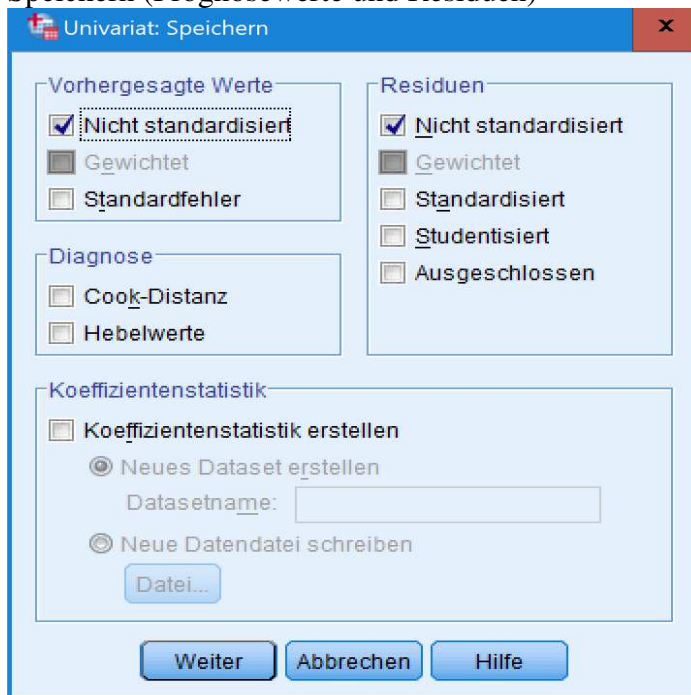
Es wird ein volles Modell mit der 2-er Interaktion Geschlecht \* Beruf gerechnet. Die Schätzung erfolgt nach dem SPSS-Typ III. In Almo entspricht dem das Verfahren der "weighted squares of means".

## Kontraste



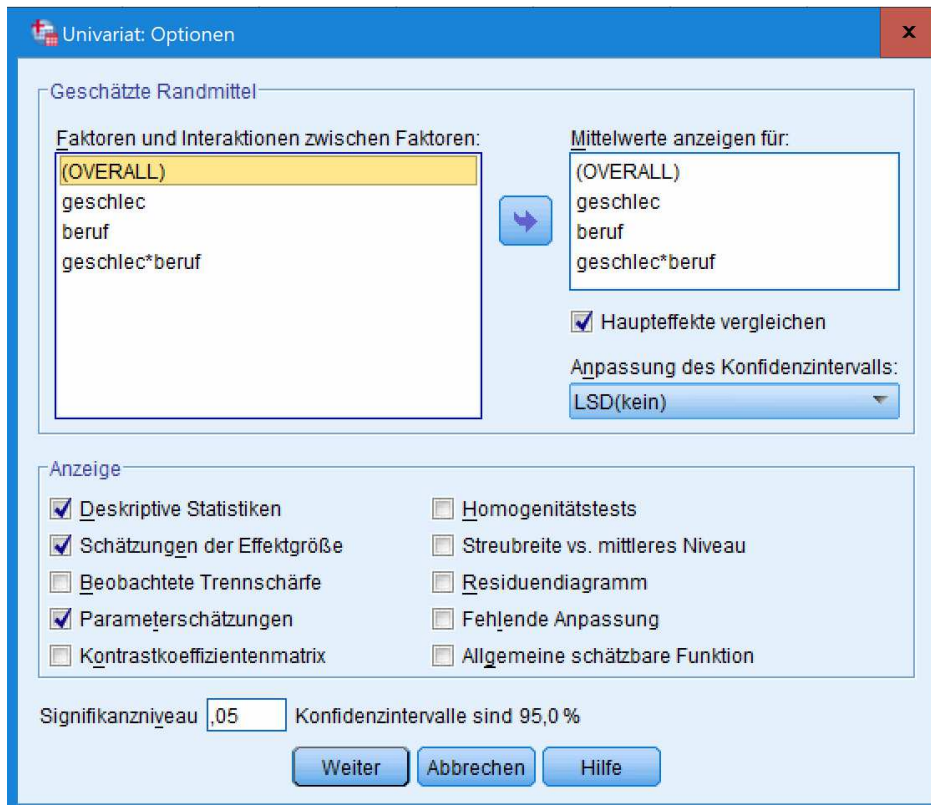
Für die nominalen Variablen werden die Abweichungskontraste angefordert. Wie wir noch zeigen werden, sind daraus die Haupteffekte der nominalen Variablen Geschlecht und Beruf ableitbar.

### Speichern (Prognosewerte und Residuen)



Die vom Modell prognostizierten Werte in der abhängigen Variablen "Einkommen" und die Residuen sollen für jeden Probanden ausgegeben werden. Kann unterbleiben.

Optionen



Verschiedene Koeffizienten der unabhängigen Variablen werden angefordert:

- (1) Empirische Mittelwerte der abhängigen Variablen Einkommen je Merkmalskombination der nominalen Variablen Geschlecht mit Beruf
- (2) Vom Modell reproduzierte Mittelwerte (Gesamtmittel, Randmittel, Zellenmittel)
- (3) geschätzte Parameter

## Eingabe in Almo

In Almo wird die Programm-Maske Prog20mo verwendet. Die entsprechend ausgefüllte Maske ist als Programm unter dem Namen "Vergleich\_mit\_SPSS.Alm" abgespeichert. Das Programm findet man durch Klick auf den Knopf "alle Progs" am Oberrand des Almo-Fensters. Wir bilden hier nur zwei Eingabeboxen ab.

Eingabebox für unabhängige Variable

Analyse-Variable: Unabhängige Variable Hilfe

nominale unabhängige Variable Hilfe

**Geschlecht, Beruf**

**2**

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

**Geschlecht, Beruf** paarweise Vergleiche (Kontraste) für die nominalen unabhängige Variablen rechnen

---

quantitative unabhängige Variable Hilfe

**Leistung, Alter**

---

ordinale unabhängige Variable Hilfe

Eingabebox für Prognosewerte

Loesche wieder diese Box (dann Voreinstellungen wieder gultig)

**Prognosewerte und Residuen ermitteln**

**1**

1=ermitteln und in Ergebnisliste ausgeben  
 2=ermitteln aber nicht in Ergebnisliste geben  
 Wenn abhängige Variable nominal dann  
 wird Prognoseerfolg ausgeben  
 0=keine Prognosewerte und Residuen ermitteln

**Identifikationsvariable**

**Prognosewerte und Residuen nur ermitteln für eine Stichprobe von ca x %**

---

seitherige Daten plus Prognosewerte und Residuen  
 in eine neue Datei speichern  
 Eingabefeld leer = nicht in neue Datei speichern

Prognosewerte und Residuen werden angefordert.

## Ergebnisse aus SPSS und Almo

Wir zeigen das SPSS-Ergebnis in etwas verkürzter Form und in der Reihenfolge der Ausgabe etwas umgestellt. Zum Vergleich werden die entsprechenden Ergebnisse aus Almo dazugestellt.

### SPSS: Quadratsummen je Variable

#### Tests der Zwischensubjekteffekte

Abhängige Variable: einkomm

Quelle	Typ III Quadratsumme	df	Quadratischer Mittelwert	F	Sig.	Partielles Eta hoch zwei
Korrigiertes Modell	53,526 <sup>a</sup>	7	7,647	1,254	,291	,142
Konstanter Term	148,489	1	148,489	24,353	,000	,315
leistung	,020	1	,020	,003	,955	,000
alter	16,018	1	16,018	2,627	,111	,047
geschlec	6,768	1	6,768	1,110	,297	,021
beruf	19,571	2	9,786	1,605	,211	,057
geschlec * beruf	10,741	2	5,370	,881	,420	,032
Fehler	323,162	53	6,097			
Gesamtsumme	1214,000	61				
Korrigierter Gesamtwert	376,689	60				

a. R-Quadrat = ,142 (Angepasstes R-Quadrat = ,029)

### ALMO: Quadratsummen je Variable

## Zusammenfassung

hinsichtlich der abhaengigen Variablen V7 Einkommen

Streuungsquelle	Streuung	Korrel Koeff.	F-Wert	df	Signifikanz p	(1-p)100	Test- staerke
Gesamtstreuung	376.6885						
Fehlerstreuung	323.1622			53			
alle unabh. Var. zusamm	53.5263	0.3770	1.2541	7	0.2903	70.9694	0.4852
quant./ordin. Var. zusamm	16.0196	0.2173	1.3136	2	0.2768	72.3225	0.2719
nominale Variable u. ihre Interaktionen zusammen	45.2669	0.3505	1.4848	5	0.2096	79.0384	0.4808
V5 Leistung	0.0196	0.0078	0.0032	1	0.9528	4.7187	0.0504
V6 Alter	16.0181	-0.2173	2.6270	1	0.1112	88.8812	0.3564
V1 Geschlecht	6.7677	0.1432	1.1099	1	0.2968	70.3227	0.1787
V3 Beruf	19.5715	0.2390	1.6049	2	0.2090	79.1049	0.3247
V1*V3	10.7408	0.1794	0.8808	2	0.4234	57.6625	0.1938

Die Ergebnisse stimmen exakt überein. Die verwendeten Bezeichnungen sind teilweise verschieden.

Die zu erklärende gesamte Abweichungsquadratsumme in der abhängigen Variablen "Einkommen" beträgt 376.6885. Bei Almo wird sie *Gesamtstreuung* genannt, bei SPSS *korrigierter Gesamtwert*.

Die bei SPSS bezeichnete *Gesamtsumme* von 1214 ist die Summe der quadrierten Rohwerte der abhängigen Variablen. In Almo kann der Benutzer in der Optionsbox "Streuungsmatrix" anfordern, dass als Streuung die "Kreuzprodukte" verwendet werden sollen. Dann werden einige Berechnungen auf Basis der quadrierten Rohwerte und Kreuzprodukte gerechnet und als Gesamtstreuung 1214 gemeldet. Almo meldet in diesem Falle auch für die Konstante deren Streuung, F-Wert und Signifikanz. Üblicherweise wird das ALM mit den Abweichungsquadratsummen (kurz: Quadratsummen) gerechnet.

Die in der abhängigen Variablen "Einkommen" durch *alle unabhängigen Variablen zusammen* erklärte Streuung in Almo wird bei SPSS *korrigiertes Modell* genannt. Ihre Abweichungsquadratsumme beträgt 53.5263. Zusammen mit der Fehlerstreuung von 323.1622 ergibt sich die Gesamtstreuung von 376.6885.

Wären die unabhängigen Variablen untereinander exakt unkorreliert, dann wäre die Summe aus Fehlerstreuung und der durch die einzelnen unabhängigen Variablen erklärten Streuungen gleich der Gesamtstreuung. In unserem Rechenbeispiel ist dies nicht der Fall.

SPSS gibt die Korrelation quadriert aus und bezeichnet sie als partielle Eta-Korrelation.

## SPSS: Parameterschätzung

### Parameterschätzungen

Abhängige Variable: einkomm

Parameter	B	Standardfehler	t	Sig.	95 % Konfidenzintervall		Partielles Eta hoch zwei
					Untergrenze	Obergrenze	
Konstanter Term	5,427	1,322	4,104	,000	2,774	8,079	,241
leistung	,010	,184	,057	,955	-,359	,380	,000
alter	-,243	,150	-1,621	,111	-,544	,058	,047
[geschlec=1,00]	,551	1,289	,428	,671	-2,034	3,136	,003
[geschlec=2,00]	0 <sup>a</sup>	.	.	.	.	.	.
[beruf=1,00]	-,419	1,249	-,336	,739	-2,925	2,086	,002
[beruf=2,00]	-,382	1,162	-,329	,744	-2,713	1,948	,002
[beruf=3,00]	0 <sup>a</sup>	.	.	.	.	.	.
[geschlec=1,00] * [beruf=1,00]	-2,022	1,778	-1,137	,261	-5,588	1,544	,024
[geschlec=1,00] * [beruf=2,00]	-1,883	1,574	-1,196	,237	-5,040	1,274	,026
[geschlec=1,00] * [beruf=3,00]	0 <sup>a</sup>	.	.	.	.	.	.
[geschlec=2,00] * [beruf=1,00]	0 <sup>a</sup>	.	.	.	.	.	.
[geschlec=2,00] * [beruf=2,00]	0 <sup>a</sup>	.	.	.	.	.	.
[geschlec=2,00] * [beruf=3,00]	0 <sup>a</sup>	.	.	.	.	.	.

a. Dieser Parameter wurde auf den Wert null gesetzt, da er redundant ist.

## ALMO: Regressionskoeffizienten und Effekte

Koeffizienten fuer quantitative/ordinale Variable aus univariater Analyse hinsichtlich der abhaeng. Var. V7 Einkommen

Variable	standard. Regress- koeff.	Regress- koeff.	Standard fehler	95% Konfidenz- bereich nach oben u. unten
V5 Leistung	0.0080	0.0104	0.1841	0.3692
V6 Alter	-0.2163	0.2432	0.1501	0.3010

Variable	erklärte Streuung	part. Korrel.	F-Wert	Signifikanz p	(1-p) 100	df1	df2	Test- staerke
V5 Leistung	0.0196	0.008	0.003	0.953	4.72	1	53	0.0504
V6 Alter	16.0181	-0.217	2.627	0.111	88.88	1	53	0.3569

Koeffizienten fuer Konstante  
Effekt (Regressionskoeffizient) 4.784218

## Effekte von A Geschlecht

	Effekte	Standard- fehler	erklärte Streuung	partielle Korrelat.	t-Wert	Signifikanz p	(1-p) 100	Test- staerke
A1 männlic	-0.3752	0.3561	6.7677	-0.1432	1.0535	0.2968	70.32%	0.1788

A2 weiblich 0.3752 0.3561 6.7677 0.1432 1.0535 0.2968 70.32% 0.1788  
 =====

**Effekte von B Beruf**

	Effekte	Standard fehler	erklärte Streuung	partielle Korrelat.	t-Wert	Signifikanz p	Signifikanz (1-p)100	Test-stärke
B1 Arbeite	-0.5123	0.4956	6.5160	-0.1406	1.0338	0.3061	69.39%	0.1739
B2 Angeste	-0.4056	0.4363	5.2687	-0.1267	0.9296	0.3567	64.33%	0.1495
B3 Selbstä	0.9179	0.5133	19.4966	0.2385	1.7882	0.0796	92.04%	0.4198

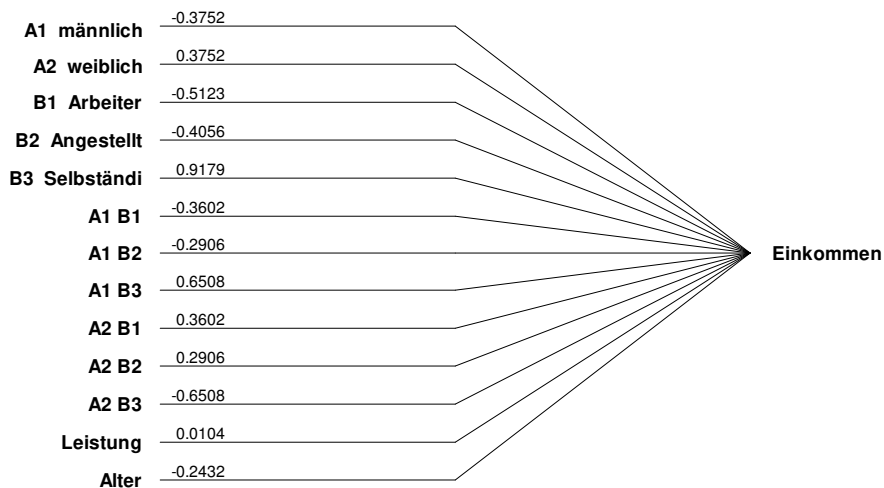
**Effekte von AB**

	Effekte	Standard- fehler	erklärte Streuung	partielle Korrelat.	t-Wert	Signifikanz p	Signifikanz (1-p)100	Test- stärke
A1 B1	-0.3602	0.4994	3.1723	-0.0986	0.7213	0.4738	52.62%	0.1091
A1 B2	-0.2906	0.4387	2.6761	-0.0906	0.6625	0.5100	49.00%	0.0997
A1 B3	0.6508	0.4919	10.6729	0.1788	1.3230	0.1916	80.84%	0.2551
A2 B1	0.3602	0.4994	3.1723	0.0986	0.7213	0.4738	52.62%	0.1091
A2 B2	0.2906	0.4387	2.6761	0.0906	0.6625	0.5100	49.00%	0.0997
A2 B3	-0.6508	0.4919	10.6729	-0.1788	1.3230	0.1916	80.84%	0.2551

Die Regressionskoeffizienten der beiden unabhängigen quantitativen Variablen (Kovariaten), *Leistung* und *Alter* die SPSS und *Almo* ausgeben, stimmen überein, die Konstante jedoch nicht. Auch die Effekte in *Almo* und die Parameter B in SPSS sind verschieden. Wir kommen anschließend darauf zurück. Für die Berechnung der **Konstanten** müssen die Kovariaten berücksichtigt werden. Deren Mittelwert berechnet SPSS wie üblich aus den vorhandenen Daten. *Almo* rechnet beim Verfahren der "fitting constant" genau so, beim Verfahren der "weighted squares of means" berechnet es jedoch den Mittelwert aus den Zellenmittelwerten der Kovariaten. Letzendlich sind diese Differenzen - wie wir anschließend zeigen - ohne Bedeutung.

*Almo* unterstützt seine Ausgabe noch durch folgendes Flussdiagramm

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



Die **Effekte** aus *Almo* und die **Parameter B** der nominalen Dummies aus SPSS sind verschieden. In Abschnitt P20.7.5, Unterabschnitt "Vergleich mit SAS und SPSS" haben wir gezeigt, wie in SPSS die Parameter berechnet werden und dass sie inhaltlich kaum zu interpretieren sind. Unsere Darstellung dort bezog sich dabei auf den einfacheren Fall der Varianzanalyse, bei der keine Kovariaten vorhanden sind. Bei der Kovarianzanalyse in

unserem Beispiel ist zu berücksichtigen, dass die Parameter der nominalen Dumies an die Kovariaten angepasst sind.

In Almo lautet die Gleichung für diese Kovarianzanalyse so

$$y = a(i) + b(j) + ab(ij) + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + c$$

- a(i) = Haupteffekt der Variablen A für die Ausprägung i
- b(j) = Haupteffekt der Variablen B für die Ausprägung j
- ab(ij) = Interaktionseffekt der Ausprägungskombination ij der Variablen A und B
- x1 = Kovariate 1
- x2 = Kovariate 2
- $\beta_1$  = Regressionskoeffizient für Kovariate 1
- $\beta_2$  = Regressionskoeffizient für Kovariate 2
- c = Konstante
- y = abhängige Variable

In SPSS lautet die Gleichung im Prinzip genauso, allerdings sind a(i), b(j) und ab(ij) nicht Effekte sondern "Parameter" B.

Betrachten wir den 1. Probanden aus unseren Testdaten. Seine Werte sind folgende

Geschlecht a(i)	Beruf b(j)	Leistung x1	Alter x2	Einkommen y
1	1	4	4	2
.	.	.	.	.
.	.	.	.	.

In obige Gleichung für Almo eingesetzt entsteht für den 1. Probanden

$$y = \underbrace{-0.3752}_{a(1)} + \underbrace{-0.5123}_{b(1)} + \underbrace{-0.3602}_{ab(11)} + \underbrace{0.0104}_{\beta_1} \cdot \underbrace{4}_{x_1} + \underbrace{-0.2432}_{\beta_2} \cdot \underbrace{4}_{x_2} + \underbrace{4.784218}_{c} = 2.605$$

-0.3752 ist der 1. Effekt der Variablen Geschlecht, -0.5123 ist der 1. Effekt des Berufs

In obige Gleichung für SPSS eingesetzt entsteht

$$y = \underbrace{0.551}_{a(1)} + \underbrace{-0.419}_{b(1)} + \underbrace{-2.022}_{ab(11)} + \underbrace{0.010}_{\beta_1} \cdot \underbrace{4}_{x_1} + \underbrace{-0.243}_{\beta_2} \cdot \underbrace{4}_{x_2} + \underbrace{5.427}_{c} = 2.605$$

Obwohl Almo-Effekte und SPSS-Parameter verschieden sind, liefern sie die gleichen Prognosewerte. SPSS fügt die Prognosewerte und Residuen in der Datentabelle als letzte Variable an die Analysevariable an. In Almo werden sie in der Ergebnisliste ausgegeben. Optional kann eine neue Datei aus den seitherigen Variablen plus den Prognosewerten erzeugt und gespeichert werden. Wir zeigen nur die ersten 10 Datensätze. Man vergleiche den 1. Wert in der Tabelle mit obigem Rechenergebnis.

### SPSS: Prognosewerte und Residuen

PRE_1	RES_1
2,61	-,61
2,86	-1,86
3,09	-,09
3,31	,69
3,02	-2,02
2,54	-1,54
3,25	-,25
2,78	1,22
5,28	-3,28
4,33	-2,33

**ALMO: Prognosewerte und Residuen**

Datensatz	Wert	Prognosewert	Residuen
1	2.00000	2.60527	-0.6053
2	1.00000	2.85895	-1.8589
3	3.00000	3.09175	-0.0918
4	4.00000	3.31412	0.68588
5	1.00000	3.02484	-2.0248
6	1.00000	2.53837	-1.5384
7	3.00000	3.24721	-0.2472
8	4.00000	2.78160	1.21840
9	2.00000	5.27940	-3.2794
10	2.00000	4.32732	-2.3273
.	.	.	.
.	.	.	.

**SPSS: Abweichungskontrast, Quadratsumme und Signifikanz von "Geschlecht"**

### Kontrastergebnisse (K-Matrix)

		Abhängige Variable
geschlec Abweichungskontrast <sup>a</sup>		einkomm
Stufe 1 vs. Mittelwert	Kontrastschätzung	-,375
	Hypothetischer Wert	0
	Differenz (Schätzung - Hypothetischer Wert)	-,375
	Standardfehler	,356
	Sig.	,297
	95 % Konfidenzintervall für Differenz	Untergrenze Obergrenze

a. Ausgelassene Kategorie = 2

### Testergebnisse

Abhängige Variable: einkomm

Quelle	Quadratsumme	df	Quadratischer Mittelwert	F	Sig.	Partielles Eta hoch zwei
Kontrast	6,768	1	6,768	1,110	,297	,021
Fehler	323,162	53	6,097			

### ALMO: Effekt, erklärte Streuung und Signifikanz von "Geschlecht"

Koeffizienten fuer Variable	V1 Geschlecht
Korrelat.koeff.	0.143222
quadriert	0.020512
erklarte Streuung	6.767666
F-Wert f. erklarte Streuung	1.109927
Freiheitsgrade Nenner=1 Zaehler=53	
Signifikanz: p	0.296773
Signifikanz: (1-p)*100	70.322688 %
Teststaerke von F	0.178741

Unter der Bezeichnung "Testergebnisse" wird bei SPSS der Erklärungswert der Variablen Geschlecht mitgeteilt. Erklärte Streuung, Korrelation, F-Wert und Signifikanz stimmen mit Almo überein

### Effekte von A Geschlecht

	Effekte	Standardfehler	erklarte Streuung	partielle Korrelat.	t-Wert	Signifikanz p	Signifikanz (1-p) 100	Teststaerke
A1 männlic	-0.3752	0.3561	6.7677	-0.1432	1.0535	0.2968	70.32%	0.1788
A2 weiblic	0.3752	0.3561	6.7677	0.1432	1.0535	0.2968	70.32%	0.1788

Unter der Bezeichnung "Kontrastergebnisse (K-Matrix)" werden in SPSS die *Abweichungskontraste* der Variablen des Geschlechts ausgegeben. Sie sind identisch mit den *Effekten* in Almo (siehe Abschnitt P20.6.5). Da sich die Effekte zu 0 summieren, ergibt sich der Effekt der letzten Ausprägung durch Subtraktion der Summe der vorderen Effekte von 0. In Abschnitt P20.7.5 haben wir gezeigt, dass sich im einfachen Fall der Varianzanalyse mit gleichen Zellenhäufigkeiten der Effekt ergibt als *Abweichungskontrast* des Ausprägungs-

mittelwerts vom Gesamtmittelwert. Bei der Kovarianzanalyse ist dieser Sachverhalt nicht ganz so übersichtlich. Die Effekte in Almo bzw. die Abweichungskontraste bei SPSS werden noch an die Kovariaten angepasst.

**SPSS: Abweichungskontrast, Quadratsumme und Signifikanz von "Beruf"**

### Kontrastergebnisse (K-Matrix)

beruf Abweichungskontrast <sup>a</sup>		Abhängige Variable
		einkomm
Stufe 1 vs. Mittelwert	Kontrastschätzung	-,512
	Hypothetischer Wert	0
	Differenz (Schätzung - Hypothetischer Wert)	-,512
	Standardfehler	,496
	Sig.	,306
	95 % Konfidenzintervall für Differenz	Untergrenze Obergrenze
Stufe 2 vs. Mittelwert	Kontrastschätzung	-,406
	Hypothetischer Wert	0
	Differenz (Schätzung - Hypothetischer Wert)	-,406
	Standardfehler	,436
	Sig.	,357
	95 % Konfidenzintervall für Differenz	Untergrenze Obergrenze

a. Ausgelassene Kategorie = 3

### Testergebnisse

Abhängige Variable: einkomm

Quelle	Quadratsumme	df	Quadratische Mittelwert	F	Sig.	Partielles Eta hoch zwei
Kontrast	19,571	2	9,786	1,605	,211	,057
Fehler	323,162	53	6,097			

Der 3. Kontrast (für Selbständige), den SPSS nicht ausgibt, ist sehr einfach

$$0 - (-.512 + -.406) = 0.918$$

Er ist identisch mit dem 3. Effekt aus Almo

### ALMO: Effekt, erklärte Streuung und Signifikanz von "Beruf"

Koeffizienten fuer Variable	V3 Beruf
Korrelat.koeff.	0.238965
quadriert	0.057104
erklärte Streuung	19.571475
F-Wert f. erklärte Streuung	1.604903
Freiheitsgrade Nenner=2 Zaehler=53	
Signifikanz: p	0.208951
Signifikanz: (1-p)*100	79.104938 %
Teststaerke von F	0.324725

### Effekte von B Beruf

	Effekte	Standard fehler	erklärte Streuung	partielle Korrelat.	t-Wert	Signifikanz p	Signifikanz (1-p) 100	Test-staerke
B1 Arbeite	-0.5123	0.4956	6.5160	-0.1406	1.0338	0.3061	69.39%	0.1739
B2 Angeste	-0.4056	0.4363	5.2687	-0.1267	0.9296	0.3567	64.33%	0.1495
B3 Selbstä	0.9179	0.5133	19.4966	0.2385	1.7882	0.0796	92.04%	0.4198

**SPSS: Paarweise Vergleiche der Haupteffekte von "Geschlecht"**

**Paarweise Vergleiche**

Abhängige Variable: einkomm

(I) geschlec	(J) geschlec	Mittelwertdifferenz (I-J)	Standardfehler	Sig. <sup>a</sup>	95 % Konfidenzintervall für Differenz <sup>a</sup>	
					Untergrenze	Obergrenze
1,00	2,00	-,750	,712	,297	-2,179	,678
2,00	1,00	,750	,712	,297	-,678	2,179

Basierend auf geschätzten Randmitteln

a. Anpassung für Mehrfachvergleiche: geringste signifikante Differenz (entspricht keinen Anpassungen).

**ALMO: Paarweise Vergleiche der Haupteffekte**

**Paarweise Vergleiche (Kontraste) von A Geschlecht**

	Differenz	Standardfehler	erklärte Streuung	t-Wert (LSD)	Signifikanz p	Signifikanz (1-p) 100	Teststärke
A1 - A2	-0.7504	0.7123	6.7677	1.0535	0.2968	70.32%	0.1788

Freiheitsgrade fuer t-Wert: 53

**SPSS: Paarweise Vergleiche der Haupteffekte von "Beruf"**

**Paarweise Vergleiche**

Abhängige Variable: einkomm

(I) beruf	(J) beruf	Mittelwertdifferenz (I-J)	Standardfehler	Sig. <sup>a</sup>	95 % Konfidenzintervall für Differenz <sup>a</sup>	
					Untergrenze	Obergrenze
1,00	2,00	-,107	,780	,892	-1,671	1,458
	3,00	-1,430	,910	,122	-3,255	,395
2,00	1,00	,107	,780	,892	-1,458	1,671
	3,00	-1,324	,814	,110	-2,956	,309
3,00	1,00	1,430	,910	,122	-,395	3,255
	2,00	1,324	,814	,110	-,309	2,956

Basierend auf geschätzten Randmitteln

a. Anpassung für Mehrfachvergleiche: geringste signifikante Differenz (entspricht keinen Anpassungen).

**ALMO: Paarweise Vergleiche der Haupteffekte von "Beruf"**

**Paarweise Vergleiche (Kontraste) von B Beruf**

	Differenz	Standardfehler	erklärte Streuung	t-Wert (LSD)	Signifikanz p	Signifikanz (1-p) 100	Teststärke
B1 - B2	-0.1067	0.7801	0.1141	0.1368	0.8910	10.90%	0.0521
B1 - B3	-1.4303	0.9099	15.0665	1.5719	0.1221	87.79%	0.3392
B2 - B3	-1.3236	0.8138	16.1297	1.6264	0.1100	89.00%	0.3590

Freiheitsgrade fuer t-Wert: 53

## SPSS: Randmittel

### 1. Gesamtmittelwert

Abhängige Variable: einkomm

Mittelwert	Standardfehler r	95 % Konfidenzintervall	
		Untergrenze	Obergrenze
3,868 <sup>a</sup>	,331	3,204	4,532

a. Kovariate im Modell werden für die folgenden Werte ausgewertet: leistung = 3,8852, alter = 3,9344.

### Schätzungen

Abhängige Variable: einkomm

geschlec	Mittelwert	Standardfehler r	95 % Konfidenzintervall	
			Untergrenze	Obergrenze
1,00	3,493 <sup>a</sup>	,474	2,542	4,443
2,00	4,243 <sup>a</sup>	,498	3,243	5,243

a. Kovariate im Modell werden für die folgenden Werte ausgewertet:  
leistung = 3,8852, alter = 3,9344.

### Schätzungen

Abhängige Variable: einkomm

beruf	Mittelwert	Standardfehler r	95 % Konfidenzintervall	
			Untergrenze	Obergrenze
1,00	3,355 <sup>a</sup>	,624	2,104	4,607
2,00	3,462 <sup>a</sup>	,478	2,504	4,420
3,00	4,786 <sup>a</sup>	,641	3,500	6,071

a. Kovariate im Modell werden für die folgenden Werte ausgewertet:  
leistung = 3,8852, alter = 3,9344.

### 4. geschlec \* beruf


Abhängige Variable: einkomm

geschlec	beruf	Mittelwert	Standardfehler r	95 % Konfidenzintervall	
				Untergrenze	Obergrenze
1,00	1,00	2,620 <sup>a</sup>	,888	,838	4,402
	2,00	2,796 <sup>a</sup>	,583	1,627	3,965
	3,00	5,061 <sup>a</sup>	,939	3,179	6,944
2,00	1,00	4,091 <sup>a</sup>	,912	2,262	5,920
	2,00	4,128 <sup>a</sup>	,757	2,609	5,647
	3,00	4,510 <sup>a</sup>	,878	2,749	6,271

a. Kovariate im Modell werden für die folgenden Werte ausgewertet: leistung = 3,8852, alter = 3,9344.

## ALMO: Randmittel und Zellenmittel

In Almo muss der Benutzer in der Ergebnisliste auf den Knopf klicken

 **Zeige Ausgabe:      Geschaetzte Randmittel und Zellenmittel**

Erst dann zeigt Almo seine Ergebnisse.

<b>modellreproduzierter Gesamtmittelwert</b>	<b>3.867763</b>
<hr/> <hr/>	
<b>Randmittel</b>	
<b>V1 Geschlecht</b>	
<b>A1 männlich</b>	<b>3.492573</b>
<b>A2 weiblich</b>	<b>4.242953</b>
<b>V3 Beruf</b>	
<b>B1 Arbeiter</b>	<b>3.355434</b>
<b>B2 Angestellt</b>	<b>3.462150</b>
<b>B3 Selbständi</b>	<b>4.785706</b>
<hr/> <hr/>	
<b>2-er Randmittel AB Geschlecht*Beruf</b>	
<b>A1 B1</b>	<b>2.620023</b>
<b>A1 B2</b>	<b>2.796352</b>
<b>A1 B3</b>	<b>5.061345</b>
<b>A2 B1</b>	<b>4.090845</b>
<b>A2 B2</b>	<b>4.127948</b>
<b>A2 B3</b>	<b>4.510067</b>
<hr/> <hr/>	

SPSS liefert folgende vom Modell reproduzierte Mittelwerte:

- (1) den modellreproduzierten Gesamtmittelwert
- (2) die modellreproduzierten "Randmittel" der beiden nominalen Variablen
- (3) die modellreproduzierten Zellen-Mittelwerte

Almo liefert die Randmittel und modellreproduzierten Zellenmittelwerte und den an die Kovariaten angepassten Gesamtmittelwert. Sie stimmen mit denen von SPSS überein. Die modellreproduzierten Zellenmittelwerte werden von Almo bis zu 4 kombinierten nominalen Variablen berechnet.

Aus den von SPSS ausgegebenen reproduzierten Mittelwerten lassen sich unschwer die Effekte, wie Almo sie ausgibt, errechnen. Die Haupteffekte und Interaktionseffekte ergeben sich aus

Haupteffekt      = Randmittel - Gesamtmittel  
Interakt.effekt=Zellenmittel (AB)-Haupteffekt (A)-Haupteffekt (B)-Gesamtmittel

Gemeint sind jeweils die modellreproduzierten Mittelwerte

Beispiele:

Haupteffekt A1            = 3.493 - 3.868 = -0.375  
Haupteffekt B1            = 3.355 - 3.868 = -0.513  
Interaktionseffekt A1B1 = 2.62 - -0.375 - -0.513 - 3.868 = -0.360

Die so berechneten Haupt- und Interaktionseffekte stimmen mit denen von Almo überein

## Terminologie in Almo und SPSS

SPSS	Almo
partielltes Eta hoch zwei	partielle Korrelation (quadriert oder nicht quadriert)
dezentraler Parameter	$F \cdot df$
beobachtete Trennschärfe	Teststärke
Abweichungskontrast für Variable i Stufe j vs Mittelwert	Effekt für Variable i Ausprägung j
Testergebnis für abhängige Variable i Kontrast/Quadratsumme	durch Variable i erklärte Streuung

# Schlagwortverzeichnis

- 0,1,-1 -Kodierung 6
- 0,1-Kodierung 6
- abhängige nominale Variable 23
- allgemeines lineares Modell 5
- Alternativhypothese 127
- Anpassung 28
- aposteriori t-Test 18
- apriori t-Test 18
- balancierten Zellenhäufigkeiten 45
- Cholesky-Matrix 126
- d\_Kreuzprodukt 100
- Diskriminanzanalyse 23, 26
- Effekte 13, 14, 18, 24, 49, 130, 168
- einmalige Anpassung 31
- erklärte Streuung 9, 44, 120
- Erwartungswerte 89
- Eta-Koeffizienten 12
- Fehlerstreuung 9, 67
- fitting constants 29
- fitting constants I 29
- fitting constants II 33
- Fitting constants II 33
- fitting\_constants\_I 101
- fitting\_constants\_II 102
- F-Test 128
- Gauss-Jordan 63
- gewichtete Kleinste-Quadrate 26, 107, 168
- Gewichtung 99
- Gruppierungsvariable 116, 152
- harmonische Mittel 124
- Haupteffekte 13, 44, 120
- Heteroskedastizität 25, 106
- hierarchische Anpassung 31
- Interaktionseffekte 13, 17, 120, 144
- Kein-Wert-Behandlung 86
- Kendall's tau-b 12
- Konfidenzniveau 148
- Konstante 149
- Konstanteneffekt 14, 44, 120
- Kontraste 54, 82
- Korrelationsmatrix 145
- Kovariante 14, 29, 157
- Kovarianzanalyse 6, 29, 157
- Kreuzprodukt 100
- leere Zellen 49
- Leere Zellen 56
- lineare Abhängigkeit 56, 126
- lineare Wahrscheinlichkeitsanalyse 23
- listenweises Ausscheiden 89
- Logit-Modell 6
- LSD 18, 56
- Multikollinearität 126
- multipler Korrelationskoeffizient 121
- multipler Korrelationskoeffizient 147
- multivariate Analyse 7
- Normalverteilungstest 165
- Nullhypothese 127
- ordinale Variable 27
- Ordinale Variable 5
- paarweise Vergleiche 54
- Paarweise Vergleiche 82
- Paarweises Ausscheiden 89
- PARTIAL-Dummies 31
- partielle erklärte Streuung 129
- partieller Korrelationskoeffizient 129
- Phi-Koeffizient 12
- post hoc Test 18
- PRE-Korrelation 10
- Probit-Modell 6
- Produkt-Moment-Korrelation 12
- Prognosewerte 107, 163
- proportional reduction of error 10
- proportionale Fehlerreduktion 11
- punktbiseriale Korrelationskoeffizient 12
- Quadratsumme 100
- Quadratsummenmatrix 145
- Quartilsabstand 90
- Randmitte 134
- Randmittel 18, 130, 135, 159
- Regressionsanalyse 6
- Regressionsebene 155
- Regressionsgerade 149, 153
- Regressionskoeffizienten 10, 16, 105, 148
- Residuen 107, 163
- Risiko 127
- SAS 48, 52
- Scheffé-Test 18
- sequentiell 101
- Signifikanz 12, 127, 128, 129
- Simple Effects 144
- Sonderprogramm 40
- SPSS 48, 52, 170
- SS Typ I 101
- SS Typ II 33, 102
- SS Typ III 42, 101
- SS Type 48
- Standardisierung 23
- Streudiagramm 149
- Streuungsmatrizen 100
- taub-b 27
- Teststärke 128
- ungleiche Zellenhäufigkeiten 49, 53
- univariat 5
- Variablenhierarchie 78
- variablenweise hierarchische Auspartiellierung 38
- Varianzanalyse 6, 118
- Verfahren 100
- Vertrauensintervall 148
- w\_squares\_of\_means 101
- Wahrscheinlichkeit 24
- Wahrscheinlichkeitsfunktion 168
- wechselnde Werte 48, 167
- Wechselnde Werte 144
- weighted squares of means 42, 49, 51, 100
- Wertemuster 108
- within-group-error 67
- Zellenbesetzung 14
- Zellenmittel 137
- Zufallsgenerator 92
- Zufallsüberlagerung 90, 91
- Zufallswert 90

## Literatur zum Allgemeinen Linearen Modell

Veröffentlichungen zum Allgemeinen Linearen Modell sind so zahlreich und dazu noch über diverse Wissenschaftsbereiche verstreut, dass der Verfasser nicht einmal den Versuch unternommen hat, einen kleinen Literatur-Überblick zu präsentieren. Die nachfolgenden Veröffentlichungen sind deswegen nur solche, die im Dokument angesprochen wurden.

- 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** Statistical Principles in Experimental Design, New York  
**Michels, K.M:** 1991