



Kurt Holm

# Statistische Datenanalyse II

## Teil 2

**(Almo-Data-Mining)**  
**Ein Standard-Auswertungssystem**

## Zum Begriff "Data-Mining"

Inzwischen sind wir mit dem Begriff "Data-Mining" nicht mehr sehr glücklich. Deswegen haben wir als 1. Titel für dieses Dokument den Begriff "Statistische Datenanalyse" verwendet. Wir werden den Begriff "Data-Mining" jedoch beibehalten. Der Almo-Benutzer muss dabei akzeptieren, dass der Begriff in Almo eingengt wird auf die Auswertung von Daten, die die Form einer Datenmatrix besitzen. Siehe dazu Teil I.

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 [www.almo-statistik.de](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
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)
14. Ereignisanalyse: Sterbetafel-Methode, Kaplan-Meier-Schätzer, Cox-Regression.PDF (1,5 MB)
15. Faktorenanalyse.PDF (1,6 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 (0.8 MB)
28. Metrische multidimensionale Skalierung (MDS) (0,4 MB)
29. Metrisches multidimensionales Unfolding (MDU) (0,6 MB)
30. Nicht-metrische multidimensionale Skalierung (MDS) (0,5 MB)
31. Pfadanalyse als wiederholte Regressionsanalyse (0,7 MB)
32. Datei-Operationen mit Almo (1,1 MB)
33. Wählerstromanalyse und Wahlhochrechnung (1,6 MB)

## ÜBERSICHT ZU TEIL II

<b>KAPITEL 6: ZUSAMMENHÄNGE „BLIND“ SUCHEN .....</b>	<b>7</b>
<i>Schritt 8: Variable miteinander korrelieren.....</i>	<i>7</i>
<b>KAPITEL 7: EINZELNE ZUSAMMENHÄNGE GENAUER UNTERSUCHEN .....</b>	<b>33</b>
<i>Schritt 9a: Variable zwei- und beliebig-dimensional tabellieren.....</i>	<i>37</i>
<i>Schritt 9b: Streudiagramm für 2 oder 3 Variable .....</i>	<i>74</i>
<i>Schritt 10: ... gelöscht ...</i>	
<b>KAPITEL 8: MEHRFACH-ZUSAMMENHÄNGE UNTERSUCHEN.....</b>	<b>83</b>
<i>Schritt 11a: Ursachen für die Zielvariable: Allgemeines Lineares Modell (ALM).....</i>	<i>83</i>
<i>Ursachen für die Zielvariable: Zielvariable ist dichotom</i>	
<i>Ursachen für die Zielvariable: Zielvariable ist nominal-polytom</i>	
<i>Schritt 11b: Gewichtete Kleinste-Quadrate-Schätzung für nominal- polytome Zielvariable....</i>	<i>148</i>
<i>Ursachen für die Zielvariable: Zielvariable ist quantitativ</i>	
<i>Schritt 11c: Alternative wenn Zielvariable nominal (dichotom oder polytom): Die Logit-</i>	
<i>Analyse .....</i>	<i>171</i>
<i>Logitanalyse mit dichotomer Zielvariablen</i>	
<i>Logit-Analyse mit polytomer Zielvariablen</i>	
<b>KAPITEL 9: PROGNOSE LEISTEN.....</b>	<b>218</b>
<i>Schritt 12a: Werte für Zielvariable mit ALM prognostizieren .....</i>	<i>218</i>
<i>Schritt 12b: Werte für Zielvariable mit Logitanalyse prognostizieren.....</i>	<i>230</i>
<b>KAPITEL 10: ZUSAMMENGEHÖRIGKEITEN ZWISCHEN OBJEKTEN SUCHEN.....</b>	<b>236</b>
<i>Schritt 13: Cluster von Objekten bilden: Clusteranalyse .....</i>	<i>236</i>
<b>KAPITEL 11: URSACHEN FÜR DIE CLUSTERZUGEHÖRIGKEIT .....</b>	<b>255</b>
<i>Schritt 14a: Ursachen für die Clusterzugehörigkeit: Analyse mit ALM.....</i>	<i>244</i>
<i>Schritt 14b: Ursachen der Clusterzugehörigkeit: Analyse mit Logitmodell.....</i>	<i>267</i>
<b>LITERATUR.....</b>	<b>273</b>

# Inhaltsverzeichnis

## Teil II

<b>KAPITEL 6: ZUSAMMENHÄNGE „BLIND“ SUCHEN .....</b>	<b>7</b>
P45.12 Schritt 8: Variable miteinander korrelieren.....	7
P45.12.0 Meßniveaus von Variablen .....	7
P45.12.1 Eingabe in Programm P45m6 .....	8
P45.12.2 Erläuterungen zu den Eingabe-Boxen .....	11
P45.12.3 Ausgabe der Ergebnisse.....	16
P45.12.3.1 Der Groß-Gamma-Kalkül.....	16
P45.12.3.2 Korrelationskoeffizienten mit nominalen Variablen.....	18
P45.12.4 Das "paarweise Ausscheiden" zur Lösung des Kein-Wert-Problems .....	26
P45.12.4.1 Die Berechnung einer Korrelationsmatrix.....	27
P45.12.4.2 Die Berechnung einer "Quasi-Korrelationsmatrix" bei "paarweisem Ausscheiden" .....	29
P45.12.4.3 Die Berechnung einer Quadratsummen- oder einer Kovarianzmatrix bei "paarweisem Ausscheiden".....	30
P45.12.4.4. Das "paarweise Ausscheiden" beim Allgemeinen Linearen Modell (ALM) .....	34
P45.12.4.5 ALM auf Quasi-Korrelationsmatrix angewendet.....	35
P45.12.4.6 ALM auf Kovarianz- bzw. Quadratsummenmatrix angewendet.....	36
<b>KAPITEL 7: EINZELNE ZUSAMMENHÄNGE GENAUER UNTERSUCHEN .....</b>	<b>37</b>
P45.13 Schritt 9a: Variable zwei- und beliebig-dimensional tabellieren .....	37
P45.13.1 Erläuterungen zu den Eingabe-Boxen .....	43
P45.13.2 Ausgabe .....	58
P45.13.3 Weiterführende Hinweise.....	73
P45.14 Schritt 9b: Streudiagramm für 2 oder 3 Variable .....	74
P45.14.1 Erläuterungen zu den Eingabe-Boxen .....	77
P45.14.2 Ausgabe für 2 Variable.....	79
P45.14.3 Streudiagramm für 3 Variable .....	81
Schritt 10: ... gelöscht ...	
<b>KAPITEL 8: MEHRFACH-ZUSAMMENHÄNGE UNTERSUCHEN.....</b>	<b>83</b>
P45.15 Schritt 11a: Ursachen für die Zielvariable: Allgemeines Lineares Modell (ALM).....	83
P45.15.1 Ursachen für die Zielvariable: Zielvariable ist dichotom.....	83
P45.15.1.0 Eine theoretische Vorbemerkung .....	83
P45.15.1.1 Eingabe in Prog45mf.....	85
P45.15.1.2 Erläuterungen zu den Eingabe-Boxen.....	89
P45.15.1.3 Ausgabe bei stark verkürzten Ergebnissen.....	103
P45.15.1.3.1 Zum Begriff der „partiellen Korrelation“: .....	103
P45.15.1.4 Ausgabe bei etwas verkürzten Ergebnissen.....	105
P45.15.1.4.1 Gruppierungsvariable.....	116
P45.15.1.4.2 Kombinierte Gruppierungsvariable.....	119
P45.15.1.4.3 Mehrere Gruppierungsvariable .....	120
P45.15.1.5 Wertemuster.....	120
P45.15.1.6 Volle Ausgabe .....	121
P45.15.1.7 Prognosefähigkeit bzw. Reproduzierbarkeit des Modells .....	122
P45.15.1.8 Gewichtete Kleinste-Quadrate-Schätzung für dichotome Zielvariable .....	126
P45.15.2 Ursachen für die Zielvariable: Zielvariable ist nominal-polytom .....	130
P45.15.2.1 Ausgabe.....	134
P45.15.2.2 Schritt 11b: Gewichtete Kleinste-Quadrate-Schätzung für nominal- polytome Zielvariable .....	148
P45.15.2.2.1 Erläuterung zu den Eingabe-Boxen .....	152
P45.15.2.2.2 Ausgabe der Ergebnisse .....	155
P45.15.3 Ursachen für die Zielvariable: Zielvariable ist quantitativ .....	163
P45.15.3.1 Ausgabe.....	165
P45.15.4 Weiterführende Hinweise.....	170
P45.16 Schritt 11c: Alternative wenn Zielvariable nominal (dichotom oder polytom): Die <b>Logit-Analyse</b> ..	171
P45.16.0 Einführung .....	171
P45.16.1 Logitanalyse mit dichotomer Zielvariablen .....	174
P45.16.1.1 Eingabe in Programm.....	174
P45.16.1.2 Erläuterungen zu den Eingabe-Boxen.....	177

<i>P45.16.1.3 Ausgabe</i> .....	185
<i>P45.16.2 Logit-Analyse mit polytomer Zielvariabler</i> .....	206
<i>P45.16.2.1 Ausgabe (verkürzt)</i> .....	207
<i>P45.16.3 Weiterführende Hinweise</i> .....	217
<b>KAPITEL 9: PROGNOSE LEISTEN</b> .....	<b>218</b>
P45.17 Schritt 12a: Werte für Zielvariable mit ALM prognostizieren.....	218
<i>P45.17.1 Eingabe</i> .....	219
<i>P45.17.2 Erläuterungen zu den Eingabe-Boxen von Prog45mp</i> .....	224
<i>P45.17.3 Ausgabe aus Prog45mp</i> .....	227
<i>P45.17.4 Prognose im Anschluß an gewichtetes ALM mit nominal-polytomer Zielvariablen</i> .....	229
P45.18 Schritt 12b: Werte für Zielvariable mit Logitanalyse prognostizieren.....	230
<i>P45.18.1 Eingabe in Prog 45mt</i> .....	230
<i>P45.18.2 Erläuterungen zu den Eingabe-Boxen</i> .....	233
<i>P45.18.3 Ausgabe</i> .....	234
<b>KAPITEL 10: ZUSAMMENGEHÖRIGKEITEN ZWISCHEN OBJEKTEN SUCHEN</b> .....	<b>236</b>
P45.19 Schritt 13: Cluster von Objekten bilden: Clusteranalyse.....	236
<i>P45.19.1 Erläuterung zu den Eingabe-Boxen</i> .....	239
<i>P45.19.2 Ausgabe der Ergebnisse</i> .....	246
<i>P45.19.3 Technische Anmerkung zur Clusteranalyse im Almo-Data-Mining</i> .....	254
<i>P45.19.4 Weiterführende Hinweise</i> .....	254
<b>KAPITEL 11: URSACHEN FÜR DIE CLUSTERZUGEHÖRIGKEIT</b> .....	<b>255</b>
P45.20 Schritt 14a: Ursachen für die Clusterzugehörigkeit: Analyse mit ALM.....	255
<i>P45.20.1 Erläuterungen zu den Eingabe-Boxen</i> .....	259
<i>P45.20.2 Ausgabe der Ergebnisse</i> .....	262
P45.21 Schritt 14b: Ursachen der Clusterzugehörigkeit: Analyse mit Logitmodell.....	267
<i>P45.21.1 Erläuterungen zu den Eingabe-Boxen</i> .....	269
<i>P45.21.2 Ausgabe der Ergebnisse (gekürzt)</i> .....	269
P45.22 Clusterzugehörigkeit als unabhängige oder als abhängige Variablen.....	273
<b>SCHLAGWORTVERZEICHNIS</b> .....	<b>274</b>
<b>LITERATUR</b> .....	<b>274</b>

## Übersicht zu Teil I (enthalten in Almo-Dokument 24)

### Kapitel 1: Eine Almo-Arbeitsdatei erstellen

*Schritt 1a: „Tabulator-getrennte“ Daten aus Excel nach Almo übertragen*

*Schritt 1b: Daten im Format FREI oder FIX in eine Almo-Arbeitsdatei schreiben*

### Kapitel 2: Daten kennenlernen

*Schritt 2: Daten anschauen*

*Schritt 3: Kennwerte der Variablen anschauen*

*Schritt 4: Variable auszählen*

### Kapitel 3: Daten bereinigen

*Schritt 5a: Mittelwert für fehlende Werte einsetzen*

*Schritt 5b: Prognosewerte für fehlende Werte durch das Allgemeine Lineare Modell  
ermitteln*

*Schritt 5c: Prognosewerte für fehlende Werte durch Logitanalyse ermitteln*

*Schritt 5d: Multiple Imputation*

### Kapitel 4: Dateien vereinen

*Schritt 6a: Datenfusion mit dem Allgemeinen Linearen Modell*

*Schritt 6b: Datenfusion mit der Logitanalyse*

*Schritt 6c: Fusionierte Dateien vereinen*

### Kapitel 5: Mehrere Variable zu einer Messung kombinieren

*Schritt 7a: Aus mehreren Variablenwerten einen Gesamtpunktwert bilden*

*Schritt 7b: Mit Faktorenanalyse einen gewichteten Gesamtpunktwert (Faktorwert) bilden*

*Schritt 7c: Rasch-Skalierungsverfahren*

## Übersicht zu Teil III (enthalten in Almo-Dokument 26)

### Arbeiten mit dem Almo-Datenanalyse-System

# Kapitel 6: Zusammenhänge „blind“ suchen

## **P45.12 Schritt 8: Variable miteinander korrelieren**

Wir suchen „blind“ nach Zusammenhängen zwischen Variablen. Zu diesem Zweck korrelieren wir jede Variable mit jeder anderen.

Wenn sehr viele Variable vorhanden sind, kann eine sehr große Matrix von Korrelationen entstehen. In diesem Falle wird man nur „halbblind“ suchen. Man wird jene Variable ausschließen, die uninteressant sind.

Die Variablen dürfen dabei quantitativ, nominal und ordinal sein. Wir können dann einen bestimmten niedrigsten Korrelationskoeffizienten als Grenzwert festlegen. Liegt ein Koeffizient höher, dann betrachten wir dies als Hinweis auf einen Variablen-Zusammenhang.

## **P45.12.0 Meßniveaus von Variablen**

Almo unterscheidet die 3 Meßniveaus: quantitativ, ordinal und nominal.

**Nominale** (oder qualitative) Variable

Beispiel: Beruf:

- 1 Arbeiter
- 2 Angestellte
- 3 Beamte
- 4 Bauern
- 5 Selbständige

Den Ausprägungen werden Ziffern zugeordnet. Die Ziffern drücken keine Ordnungsrelation aus. Sie sind lediglich Kennziffern für die jeweilige Ausprägung. Die Zuordnung der Ziffern ist beliebig. So könnte etwa umgekehrt "Angestellter" mit 1 und "Arbeiter" mit 2 kodiert werden. Bei den nominalen Variablen unterscheidet Almo gelegentlich zwischen **dichotomen** Variablen (2 Ausprägungen) und **polytomen** Variablen (mehr als 2 Ausprägungen)

**Ordinale** Variable

Beispiel: Schulbildung:

- 1 Volksschulabschluß
- 2 Hauptschulsabschluß
- 3 Gymnasium
- 3 Fachschule
- 4 Universitätsabschluß

Die Ziffern 1 bis 4, die den Ausprägungen der Schulbildung zugeordnet werden, drücken eine Rangordnung im Bildungsniveau aus. 4 ist mehr als 3 und 3 ist mehr als 2 und 2 ist mehr als 1 - um wieviel mehr ist nicht bekannt. Die Differenzen zwischen den Ziffern sind nicht bekannt. Gymnasium und Fachschule wurden gleichrangig mit 3 eingestuft. Die Ziffern drücken also die Relation "mehr" oder "weniger" oder "gleich" aus.

**Rangwert-Variable** als Sonderform der ordinalen Variablen.

Gelegentlich werden die Untersuchungsobjekte entsprechend ihrer ordinalen Werte hintereinander gestellt. Dann werden ihnen fortlaufende Rangplatzziffern zugeordnet. Betrachten wir ein Beispiel:

Die Variable "sportliche Leistung" sei eine ordinale Variable. Sie wird in folgender Weise kodiert:

		<u>Code</u>
Leistung:	hervorragend	1
	gut	2
	mittel	3
	schwach	4

7 Personen wurden in ihrer Leistung gemessen. In nachstehender Tabelle ist angegeben welche Werte sie erzielen konnten und welchen Rangplatz sie damit einnahmen.

Person	Wert in der ordinalen Variablen Leistung	Code	Wert in der Rangvariablen
-----	-----	-----	-----
1	hervorragend	1	1
2	gut	2	2.5
3	gut	2	2.5
4	mittel	3	5
5	mittel	3	5
6	mittel	3	5
7	schwach	4	7

Der "Wert in der Rangvariablen" ist sehr einfach der Rangplatz der Person, wenn alle Personen nach ihrer Leistung hintereinander gestellt werden. Da manche Personen dieselbe Leistung erbringen, wie z.B. die Personen 2 und 3 wird eine "Rangteilung" vorgenommen. Person 2 und 3 teilen sich die Rangplätze 2 und 3. Der mittlere Wert ist 2.5. Die Personen 4, 5 und 6 teilen sich die Rangplätze 4, 5, 6. Der Wert in der Mitte ist 5. Siehe dazu Almo-Dokument Nr.3 "Nichtparametrische Verfahren", Abschnitt P8.2.7.

### **Quantitative Variable**

Beispiel: Lebensalter

Die zugeordneten Zahlen drücken nicht nur eine Rangordnung aus wie bei den ordinalen Zahlen, sie geben auch die Distanz zwischen den Meßobjekten an. Gernot ist 24 Jahre alt, Ariane 18 und Roland 16. Gernot ist also 6 Jahre älter als Ariane und 8 Jahre älter als Roland

In den verschiedenen Almo-Programmen werden für die 3 Meßniveaus teilweise unterschiedliche Koeffizienten berechnet.

Bei einigen Programmen, etwa Prog45ml (Kennwerte der Variablen, Basis-Statistiken) kann auch ein und dieselbe Variable als quantitativ und als nominal und als ordinal angegeben werden. Der Benutzer erhält dann für diese Variable die Koeffizienten, die Almo für diese 3 Meßniveaus ermittelt.

## **P45.12.1 Eingabe in Programm P45m6**



6 **Zu korrelierende Variable** Hilfe

quantitative Variable

**Einkommen, Rueckrate, Laufzeit**

---

6 **nominale Variable** Hilfe  
sie werden in Dummies aufgelöst

**Wohnort, Beruf, Produkt, Rueckzahl**

---

**ordinale Variable** Hilfe

7  **Option: Ein- und Ausschliessen von Untersuchungseinheiten**

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

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

10  **Option: Untersuchungseinheiten gewichten**

11  **Option: Partielle Korrelationsmatrix bilden**

12  **Option: "Aussehen" der auszugebenden Tabelle bzw. Matrix**

13     **Option: Nur Koeffizienten ausgeben, deren absolute Werte größer sind als ...** Hilfe

14  **Grafik-Optionen**

15 **Programmende**

## P45.12.2 Erläuterungen zu den Eingabe-Boxen

**Eingabe-Box 1:** Speicher für x Variable.

Siehe "Arbeiten mit Almo-Datenanalyse-System", Abschnitt P0.1.

**Eingabe-Box 2:** Weitere Vereinbarungen

Siehe "Arbeiten mit Almo-Datenanalyse-System", Abschnitt P0.2.

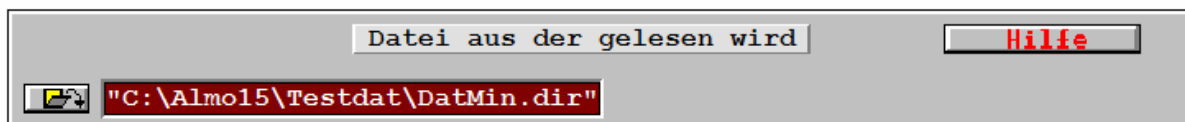
**Eingabe-Box 3:** Datei der Variablennamen

Siehe "Arbeiten mit Almo-Datenanalyse-System", Abschnitt P0.3.

**Eingabe-Box 4:** Freie Namensfelder

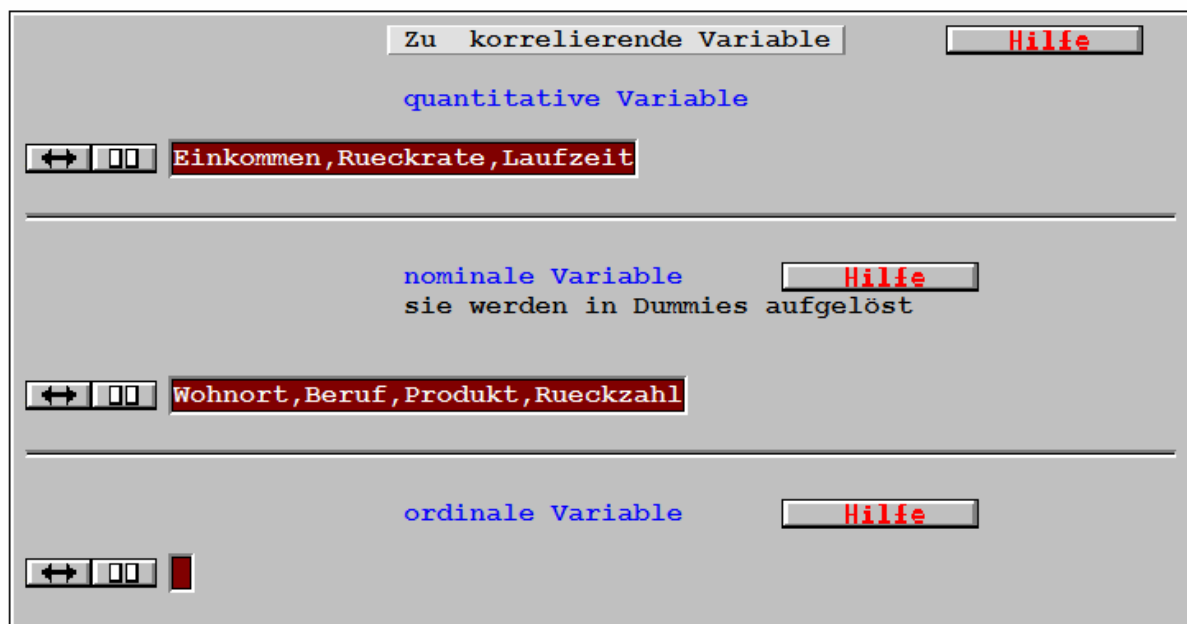
Siehe "Arbeiten mit Almo-Datenanalyse-System", Abschnitt P0.3.

**Eingabe-Box 5:** Datei aus der gelesen wird



Geben Sie hier die Almo-Arbeitsdatei ein (im Format direkt), die Sie mit Prog45md oder Prog45mh erzeugt haben. Siehe dazu auch "Arbeiten mit Almo-Datenanalyse-System", Abschnitt P0.4. Diese beiden Programm-Masken wurden im Teil 1 des "statistischen Datenanalyse-Systems" Abschnitt P45.1 und P45.2 ausführlich erläutert.

**Eingabe-Box 6:** Zu korrelierende Variable



*Eingabefeld 1:* Geben Sie hier die quantitativen Variablen an

*Eingabefeld 2:* Geben Sie hier die nominalen Variablen an

Almo löst die nominalen Variablen in so viele Dummy-Variable auf wie sie Ausprägungen besitzen. Zur Kodierung der nominalen Variablen siehe Almo-Handbuch „P20 Allgemeines Lineares Modell“, Abschnitt P20.3.

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

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

Im vorliegenden Programm P45m6 ist es jedoch zulässig, daß sie auch Dezimalwerte besitzen. Beispiel:

Beruf	Codeziffer
-----	-----
Arbeiter	10.5
Angestellter	12
Beamter	13.9
Bauer	8.1
.	.
.	.
.	.

Als Codeziffern sind hier etwa Qualifikationspunkte verwendet worden. Almo kodiert diese Werte zwangsweise um, und zwar in folgender Weise:

alte Codeziffer	neue Codeziffer
-----	-----
8.1	1
10.5	2
12	3
13.9	4
.	.
.	.
.	.

Almo stellt gewissermaßen den nominalen Charakter wieder her, den der Benutzer mit seiner Einordnung der Variablen als nominale festgelegt hat.

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 8 "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 (Code : 12) und Beamte (Code: 13.9) in der Art ihrer Tätigkeit als sehr ähnlich erachtet und sie deswegen in einer Kategorie zusammengefaßt.

### Eingabefeld 3: Ordinale Variable

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 Programm 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 8 "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 (Code : 13.1) und Fachschule (Code : 13.9) für gleichrangig erachtet und in einer Kategorie zusammengefaßt.

Wenn Sie ordinale Variable mit mehr als ca. 6 Ausprägungen haben, dann ist es sinnvoll diese als quantitative anzugeben. Ordinale Variable verlängern die Rechenzeit und benötigen sehr viel zusätzlichen Speicherplatz.

#### **Eingabe-Box 7:** Option Ein- und Ausschließen

Siehe "Arbeiten mit Almo-Datenanalyse-System", Abschnitt P0.7.

#### **Eingabe-Box 8:** Kein-Wert-Angabe und Umkodierungen.

Siehe "Arbeiten mit Almo-Datenanalyse-System", Abschnitt P0.5.

#### **Eingabe-Box 9:** Spezielle Kein-Wert-Behandlung



Wurden mit Prog45mo oder Prog45mm fehlende Werte in den zu korrelierenden Variablen bereits durch Schätzwerte ersetzt, dann wird diese Option nicht benötigt.

Diese Option wird auch nicht benötigt, wenn der Benutzer das "paarweise Ausscheiden" als Kein-Wert-Behandlung akzeptiert. Dies ist die Voreinstellung in Almo.

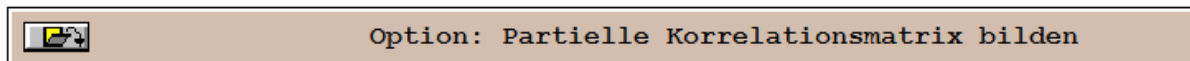
Wird die Optionsbox geöffnet, dann präsentiert Almo eine sehr große Eingabe-Box mit 7 Methoden der Kein-Wert-Behandlung. Wir haben die Methoden 1 bis 3 bereits ausführlich in P45.7.1.3, Eingabe-Box 9 bei den Erläuterungen zu Prog45mm und

die Methoden 4 bis 7 in P45.6.1, Eingabe-Box 7 bei den Erläuterungen zu Prog45mo beschrieben.

Die Methode 1, das „paarweise Ausscheiden“ haben wir besonders ausführlich im folgenden Abschnitt P45.12.4 dargestellt. Wählt der Benutzer diese Methode, dann kann er auch eine Entscheidung darüber treffen, welche Fallzahl *Almo* für Signifikanzberechnungen einsetzen soll. Das Eingabefeld dafür steht ganz unten in der großen Eingabe-Box. Siehe dazu die ausführliche Darstellung in P45.12.4. Die Voreinstellung in *Almo* ist das harmonische Mittel.

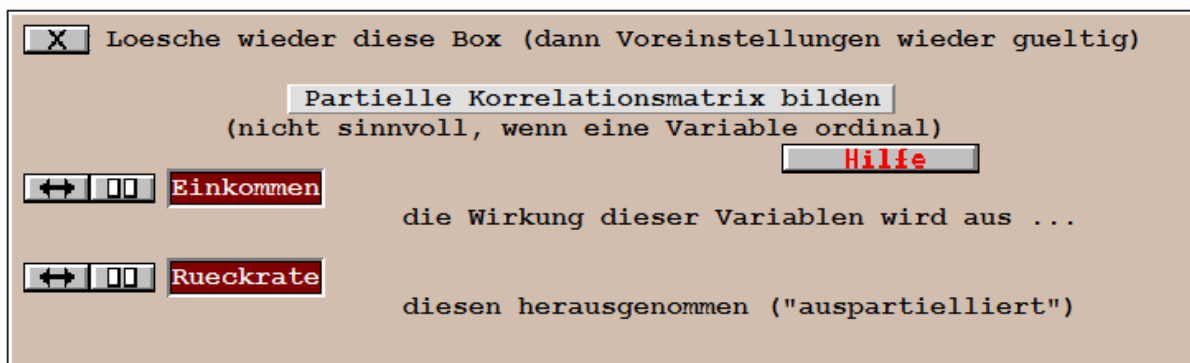
**Eingabe-Box 10:** Option : Untersuchungseinheiten gewichten  
Siehe "Arbeiten mit *Almo*-Datenanalyse-System", Abschnitt P0.8.

**Eingabe-Box 11:** Option: Partielle Korrelationsmatrix



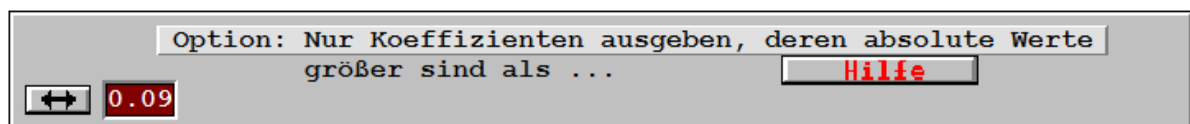
Wir empfehlen, nur dann eine partielle Korrelationsmatrix zu bilden, wenn man auch wirklich einen Grund dafür besitzt. In den ersten Phasen des Data-Mining-Prozesses wird man eine partielle Korrelationsmatrix nicht bilden.

Wenn Sie auf den Öffne-Knopf der Optionsbox klicken, dann wird folgende Eingabe-Box inkludiert:



Beispiel: Die Variablen V1 bis V10 werden korreliert. Nun ist es möglich, in der Matrix der Variablen V1 bis V10, die Variablen V1 bis V4 aus den anderen heraus zu "partiellieren". Aus der Korrelation z.B.  $r_{78}$  von V7 mit V8 ist dann der Einfluß, den die Variablen V1 bis V4 auf diese haben, "herausgenommen". Wir erhalten also die Korrelation  $r_{78.1234}$ . Siehe Handbuch, Teil 3, Abschnitt P19.

**Eingabe-Box 12:** Option: Nur Koeffizienten ausgeben, deren absolute Werte größer sind als ...



Werden viele Variable korreliert, dann ist die Korrelationsmatrix sehr unübersichtlich. In diesem Falle ist folgende Vorgehensweise sinnvoll: Man bildet zunächst – ohne diese Option zu nutzen – eine Korrelationsmatrix. *Almo* gibt unterhalb der Matrix eine Tabelle der Signifikanzen der Korrelationskoeffizienten aus. Der Benutzer sieht folgendes:

Mindestgroesse des Produkt-Moment-Korrelationskoeffizienten r	bei Signifikanz (1-p)*100	fuer df=n-2=1000-2=998
0.0405	80	
0.0455	85	
0.0520	90	
0.0563	92.5	
0.0619	95	
0.0709	97.5	
<b>0.0813</b>	<b>99</b>	
0.1051	99.9	

Bei einer Signifikanz von 99% müssen die Korrelationskoeffizienten grösser gleich 0.0813 sein.

Man kann nun in oben abgebildeter Eingabe-Box diesen Wert von 0.0813 eingeben und die Analyse ein 2. Mal rechnen. Dann bleiben in der Korrelationsmatrix alle Zellen leer, in denen ein Koeffizient von kleiner 0.0813 enthalten ist. Die Korrelationsmatrix wird dadurch wesentlich übersichtlicher. Man erkennt die bedeutsamen Variablen-zusammenhänge prägnanter.

Man sollte folgende 2 Punkte aber berücksichtigen:

- (1) Signifikanzen zu ermitteln ist nur sinnvoll, wenn die Daten als eine Zufallsstichprobe aus einer definierten Grundgesamtheit stammen.
- (2) Bei großen Datenmengen (Stichproben) sind schon relativ kleine Korrelationskoeffizienten signifikant.

Die Folgerung daraus: Uns geht es in dieser Phase des Data-Mining-Prozesses darum, Zusammenhänge zu explorieren. Die Korrelationskoeffizienten sind dafür Kennziffern. Es spricht also nichts dagegen, unabhängig von Überlegungen zur Signifikanz, einen Mindestwert festzulegen – z.B. 0.09. Koeffizienten, die größer sind, weisen uns auf einen *relevanten* Variablenzusammenhang hin.

### **Eingabe-Box 13:** Grafik-Optionen

Siehe "Arbeiten mit Almo-Datenanalyse-System", Abschnitt P0.10.

## P45.12.3 Ausgabe der Ergebnisse

### P45.12.3.1 Der Groß-Gamma-Kalkül

Almo ermöglicht es, die Korrelationen zwischen nominalen, ordinalen und quantitativen Variablen zu ermitteln. Dazu verwendet es den Groß-Gamma-Kalkül. Siehe dazu unsere ausführliche Darstellung im Almo-Dokument Nr. 5 "Korrelation", Abschnitt P19.0.3 und im Exkurs von Heinrich Potuschack, sowie Denz (1977, 1979) und Almo-Dokument Nr. 13a "Das Allgemeine Lineare Modell", Abschnitt P20.6.9).

Abhängig vom Meßniveau der Variablen berechnet Almo folgende Korrelationskoeffizienten:

	quantit.	ordinal	nominal-dichotom
quantit.	r	namenlos	punktbiser. r
ordinal		tau-b	biserial. tau_b
nominal-dichotom			Phi

Der in obiger Tabelle als „namenlos“ bezeichnete Korrelationskoeffizient zwischen einer quantitativen und einer ordinalen Variablen, ist in der statistischen Literatur nicht als spezifischer Koeffizient anzutreffen. Er ergibt sich im Rahmen des Groß-Gamma-Modells.

Alle (quadrierten) Korrelationskoeffizienten  $r_{ik}$  sind "proportional reduction of error"-Koeffizienten (abgekürzt: PRE-Koeffizienten). Sie drücken den Anteil aus, um den sich die Fehlerstreuung in der Variablen  $k$  reduziert, wenn  $i$  als erklärende Variable eingeführt wird. Da die von Almo errechneten Koeffizienten PRE-Korrelationskoeffizienten sind, können sie untereinander verglichen werden. Es ist also möglich beispielsweise zu sagen: Die Korrelation zwischen Wohnort und Rückzahlung ist (mit 0.13) im Absolutwert kleiner als die Korrelation zwischen Einkommen und Rückzahlung (mit 0.43).

Almo liefert folgende Ergebnisse (etwas verkürzt):

Korrelations-Matrix

		Wohnort Stadt A1	Wohnort Land A2	Beruf Selbst B1	Beruf Unselbst B2	Produkt Kleidung C1	Produkt Möbel C2	Produkt Technik C3
Wohnor	Stadt A1	1.00	-1.00	0.04	-0.04	0.02	-0.03	0.02
Wohnor	Land A2	-1.00	1.00	-0.04	0.04	-0.02	0.03	-0.02
Beruf	Selbst B1	0.04	-0.04	1.00	-1.00	0.00	0.01	-0.01
Beruf	Unselb B2	-0.04	0.04	-1.00	1.00	0.00	-0.01	0.01
Produk	Kleidu C1	0.02	-0.02	0.00	0.00	1.00	-0.40	-0.42
Produk	Möbel C2	-0.03	0.03	0.01	-0.01	-0.40	1.00	-0.66
Produk	Techni C3	0.02	-0.02	-0.01	0.01	-0.42	-0.66	1.00
Rueckz	nein D1	0.13	-0.13	0.07	-0.07	0.12	0.03	-0.13
Rueckz	ja D2	-0.13	0.13	-0.07	0.07	-0.12	-0.03	0.13
Einkom	V4	-0.01	0.01	0.00	0.00	-0.01	0.01	0.00
Rueckr	V7	0.01	-0.01	0.00	0.00	0.00	0.03	-0.03
Laufze	V8	-0.02	0.02	-0.04	0.04	0.02	0.01	-0.02

		Rueckzah nein D1	Rueckzah ja D2	Einkomme V4	Rueckrat V7	Laufzeit V8
Wohnor	Stadt A1	0.13	-0.13	-0.01	0.01	-0.02
Wohnor	Land A2	-0.13	0.13	0.01	-0.01	0.02
Beruf	Selbst B1	0.07	-0.07	0.00	0.00	-0.04
Beruf	Unselb B2	-0.07	0.07	0.00	0.00	0.04
Produk	Kleidu C1	0.12	-0.12	-0.01	0.00	0.02
Produk	Möbel C2	0.03	-0.03	0.01	0.03	0.01
Produk	Techni C3	-0.13	0.13	0.00	-0.03	-0.02
Rueckz	nein D1	1.00	-1.00	-0.43	0.43	0.09
Rueckz	ja D2	-1.00	1.00	0.43	-0.43	-0.09
Einkom	V4	-0.43	0.43	1.00	-0.48	-0.17
Rueckr	V7	0.43	-0.43	-0.48	1.00	0.37
Laufze	V8	0.09	-0.09	-0.17	0.37	1.00

**\*\*\*\*\*Erläuterung:**

Die nominalen Variablen werden von Almo in ihre Dummies aufgelöst. Jede Ausprägung bildet eine Dummy-Variable. Sofern eine nominale Variable nur 2 Ausprägungen besitzt treten keine Probleme auf. Die Korrelationen ihrer beiden Dummies mit irgend einer anderen Variablen sind absolut gleich groß und haben entgegengesetzte Vorzeichen.

Beispiel: Rückzahlung mit Einkommen

	Einkommen
Rueckz nein	-0.43
Rueckz ja	0.43

Dies gilt jedoch nicht mehr, wenn die nominale Variable 3 und mehr Ausprägungen besitzt.

Mindestgroesse des Produkt-Moment-Korrelationskoeffizienten r	bei Signifikanz (1-p)*100	fuer df=n-2=1000-2=998
0.0405	80	
0.0455	85	
0.0520	90	
0.0563	92.5	
0.0619	95	
0.0709	97.5	
<b>0.0813</b>	<b>99</b>	
0.1051	99.9	

**\*\*\*\*\*Erläuterung:**

Beispiel: Eine Signifikanz von mindestens 99 % besitzen alle Korrelationskoeffizienten, die größer als 0.0813 sind. Sind ordinale Variable an den Korrelationen beteiligt, dann gibt Almo zusätzlich die Signifikanz für Kendalls tau-b (ohne Bindungen) aus. Der Benutzer sollte diese Signifikanzwerte auch für die „Misch“-Korrelationskoeffizienten verwenden, bei denen die eine Variable ordinal und die andere quantitativ oder nominal-dichotom ist. Heinrich Potuschak hat die exakte Signifikanz aller dieser „Misch“-Koeffizienten abgeleitet.

Korrelations-Matrix

(Dummies der nominalen Variablen werden über eine kanonische Korrelationsanalyse zusammengefaßt)

	Wohnort V1	Beruf V3	Produkt V9	Rueckzah V10	Einkomme V4	Rueckrat V7	Laufzeit V8
Wohnort V1	1.00	0.04	0.04	0.13	0.01	0.01	0.02
Beruf V3	0.04	1.00	0.01	0.07	0.00	0.00	0.04
Produkt V9	0.04	0.01	1.00	0.15	0.01	0.03	0.02
Rueckz V10	0.13	0.07	0.15	1.00	0.43	0.43	0.09
Einkom V4	0.01	0.00	0.01	0.43	1.00	-0.48	-0.17
Rueckrat V7	0.01	0.00	0.03	0.43	-0.48	1.00	0.37
Laufzeit V8	0.02	0.04	0.02	0.09	-0.17	0.37	1.00

**P45.12.3.2 Korrelationskoeffizienten mit nominalen Variablen**

Wenn eine Variable nominal ist, dann wird sie von Almo in Dummies aufgelöst. In der Korrelationsmatrix stehen dann die Korrelationskoeffizienten zwischen den Dummies.

In obiger Korrelationsmatrix werden sie dann im Rahmen einer kanonischen Korrelationsanalyse wieder zusammengefasst.

Betrachten wir ein Beispiel mit zwei nominalen Variablen mit je 3 Ausprägungen. Da wir in unserem Beispiel diese Konstellation nicht antreffen, betrachten wir ein eigens konstruiertes Beispiel.

Die Variable Beruf (mit 3 Ausprägungen) und die Variable Schulbildung (mit 3 Ausprägungen) werden korreliert.

Almo ermittelt folgende Korrelationsmatrix der Dummies:

		Beruf Arbeit	Beruf Angest	Beruf Selbst	Schulbil Hauptsch	Schulbil Gymnasiu	Schulbil Uni
Beruf Arbeit		1.00	-0.73	-0.24	-0.27	0.11	0.21
Beruf Angest		-0.73	1.00	-0.48	0.06	-0.01	-0.06
Beruf Selbst		-0.24	-0.48	1.00	0.25	-0.12	-0.17
Schulb Haupts		-0.27	0.06	0.25	1.00	-0.71	-0.34
Schulb Gymnas		0.11	-0.01	-0.12	-0.71	1.00	-0.40
Schulb Uni		0.21	-0.06	-0.17	-0.34	-0.40	1.00

Wir benötigen nun einen Koeffizienten, der in einer einzigen Zahl ausdrückt, wie Beruf und Schulbildung miteinander korrelieren. Almo errechnet nun „Pillais Spur“ und ermittelt daraus einen Korrelationskoeffizienten.

Pillais Spur wird üblicherweise im Rahmen der multivariaten Version des Allgemeinen Linearen Modells errechnet. Siehe Almo-Dokument Nr. 13a "Allgemeines lineares Modell II", Abschnitt P20.9.4.1 und besonders P20.9.5.1. Sie ergibt sich auch im Rahmen der „kanonischen Korrelationsanalyse“ als Summe der Eigenwerte. Siehe dazu Almo-Dokument Nr. 4 "Kanonische Korrelation", Abschnitt P29.1.2.

Es entsteht die Korrelation = 0.14. Wir verfügen damit über eine einzige Zahl, die die Korrelation zwischen Beruf und Schulbildung ausdrückt.

Siehe dazu unsere ausführliche Darstellung in Almo-Dokument Nr.5 "Korrelation", Abschnitt P19.2.

Dieser Korrelationskoeffizient ist vorzeichenlos bzw. immer positiv. Er ist ein PRE-Koeffizient (PRE="proportional reduction of error"). Der quadrierte Korrelationskoeffizient  $r_{ik}^2$  drückt den Anteil aus, um den sich die Fehlerstreuung in der Variablen k reduziert, wenn Variable i als erklärende Variable eingeführt wird.

Haben beide nominale Variable nur 2 Ausprägungen, dann ist die so errechnete Korrelation identisch mit dem Phi-Korrelationskoeffizienten.

Sind beide Variable polytom, dann entsteht "Cramers V" - auch "Cramers Index" genannt - (siehe Almo-Dokument Nr. 1b "Zwei- und dreidimensionale Tabellierung", Abschnitt P10.4.2). In unserem obigen Beispiel haben wir also ein Cramers V von 0.14 errechnet.

Wird eine nominale Variable mit einer quantitativen Variablen x korreliert, dann geht Almo im Prinzip genau so vor. Der dabei entstehende Korrelationskoeffizient ist dann identisch mit dem Eta-Korrelationskoeffizient (wie er im Rahmen der Varianzanalyse berechnet wird). Er entspricht auch exakt der multiplen Korrelation  $R(x.123...)$  zwischen den Dummies 1,2,3,... der nominalen Variablen und der quantitativen Variablen x. Ist die nominale Variable dichotom, dann entsteht der punktbiseriale Korrelationskoeffizient.

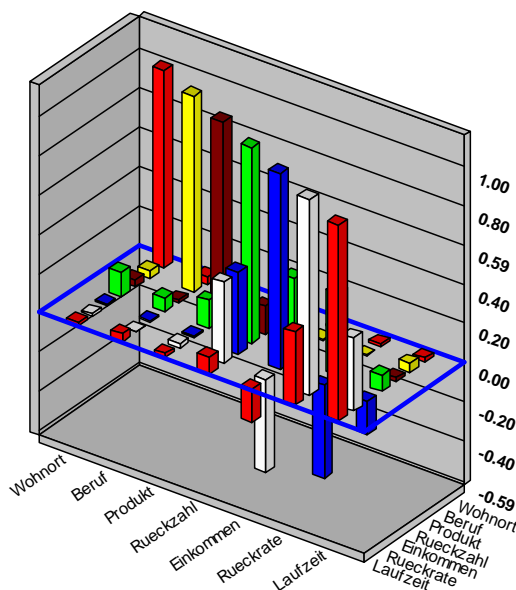
Almo errechnet also, abhängig vom Meßniveau der Variablen folgende Koeffizienten.

	quant.	ordinal	nomin-dichotom	nomin-polytom
quantitativ	r	namenlos	punktbiser.r	Eta
ordinal		tau-b	biser.tau-b	namenlos
nomin-dichotom			Phi	Phi'
nomin-polytom				Cramers V

Almo errechnet mit demselben Algorithmus auch einen Korrelationskoeffizient zwischen einer ordinalen Variablen und einer nominal-polytomen Variablen. In der statistischen Literatur gibt es dafür kein Äquivalent. Wir raten hier eher zur Vorsicht. Wenn es jedoch darum geht, die Stärke des Zusammenhangs zwischen einer Menge von Variablen relativ grob zu explorieren, dann darf er sicherlich als ein brauchbarer Indikator betrachtet werden. Siehe unsere Ausführungen zum "Groß-Gamma-Koeffizienten" in Almo-Dokument Nr. 5 "Korrelation", Abschnitt P19.0.3 und den ausführlichen Exkurs von Heinrich Potuschack

Betrachten wir jetzt noch die Grafiken die Almo liefert, dabei wollen wir uns beschränken auf die Grafik der 2. Korrelationsmatrix.

Korrelationsmatrix



Almo zeichnet ein Balkendiagramm. Ob dieses Diagramm anschaulich ist, überlassen wir dem Urteil des Benutzers (!).

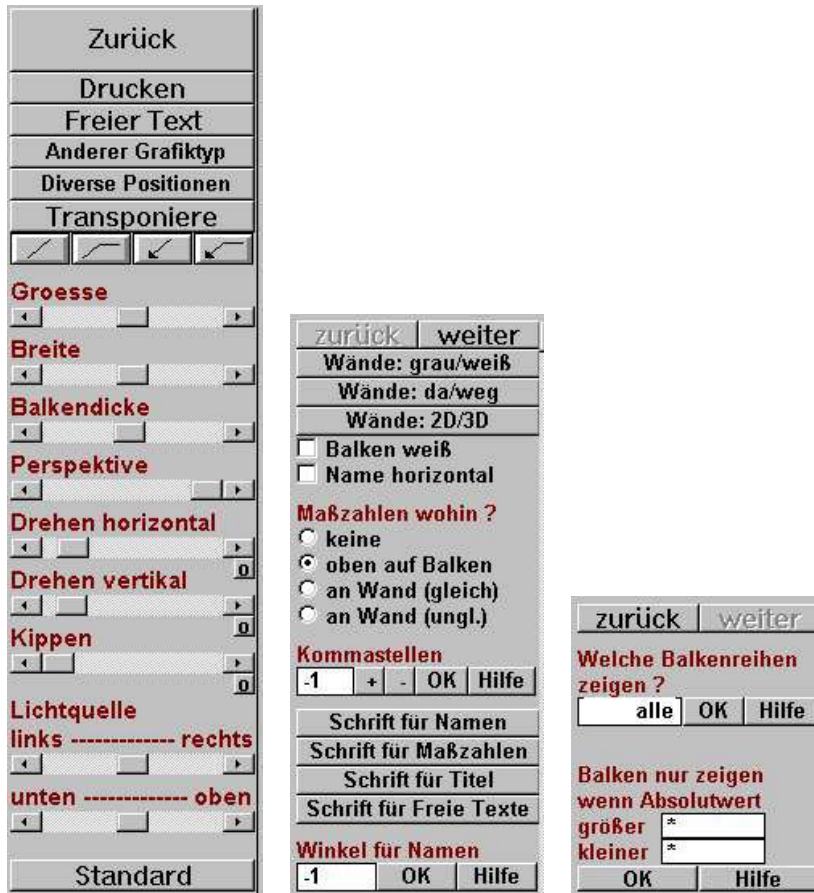
In das Diagramm ist eine „Nullebene“ eingezeichnet. Am rechten Rand werden die Korrelationswerte von 1.00 absteigend mit Schrittweite 0.2 bis 0.00 und dann weiter in den negativen Bereich angeschrieben. Negative Korrelationen, z.B. Einkommen mit Rückrate ( $r = -0.48$ ), werden als Balken dargestellt, die von der Nullebene nach unten gerichtet sind. Positive Korrelationen sind von der Nullebene nach oben weisende Balken.

Links und rechts der Grafik befinden sich je eine Leiste mit Knöpfen und Schiebern.

linke Leiste

rechte Leiste

dritte Leiste



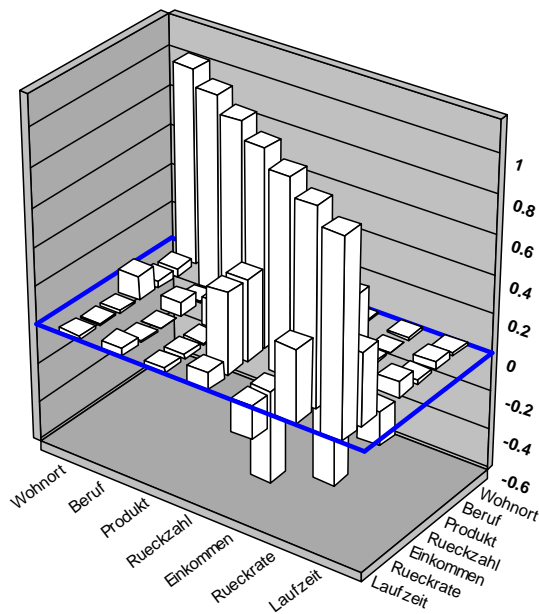
Wenn man in der rechten Leiste oben auf den Knopf „weiter“ klickt, dann wird noch eine 3. Leiste sichtbar, die wir oben auch abgebildet haben.

Die verschiedenen Elemente dieser Leisten werden in Teil 1 des Handbuchs (Bedienungsanleitung) in Abschnitt 10 erläutert.

Wir wollen hier nun einige vorteilhafte Manipulationen der Grafik vortragen:

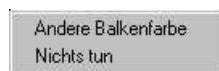
Wir verschönern das Balkendiagramm noch, indem wir in der linken Leiste 1 x in den Schieber „Perspektive“ klicken. Die Grafik erhält dadurch einen perspektivischen 3D-Eindruck. Da wir nicht farbig ausdrucken können (hier in diesem Text) klicken wir in der rechten Leiste noch auf „Balken weiß“. Wir sehen nun folgende Grafik.

## Korrelationsmatrix



Diese Grafik sollte man ausdrucken. Das erreichen wir durch Klick auf den Knopf „Drucken“ in der linken Leiste. Es entsteht ein Druckbild höchster Qualität. Es besteht auch die Möglichkeit, das Bild nach MS-Word (bzw. einer anderen kompatiblen Textverarbeitung) zu exportieren. Siehe dazu Handbuch, Teil 1, Bedienungsanleitung, Abschnitt 10.4.

Die Korrelationsmatrix ist symmetrisch. Es ist also kein Verlust, daß die Balken hinter den hohen Balken (mit 1.0) in der Diagonale unsichtbar bleiben. Wenn man will, dann kann man die höchsten Balken farblich hervorheben. Man klickt mit der linken Maustaste auf den Balken. Es erscheint dann eine kleine Auswahlbox mit dem Inhalt



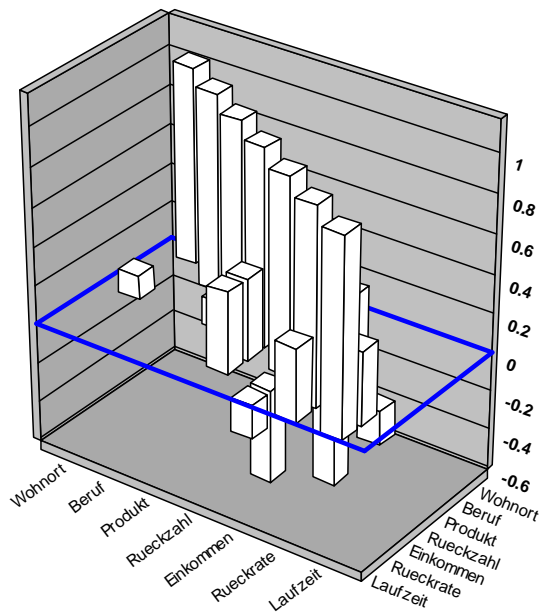
Erscheint diese Auswahlbox nicht, dann muß man an einer anderen Stelle auf den Balken klicken. Nach Klick auf „Andere Balkenfarbe“ wird die Farbauswahl-Box präsentiert, in der man sich dann für eine Farbe entscheidet.

In der 3. Grafikleiste (die man über Klick auf den Knopf „weiter“ in der rechten Grafikleiste sichtbar macht) findet man 2 Eingabefelder mit der Beschriftung „Balken nur zeigen, wenn Absolutwert größer“. Wir tragen den Wert 0.09 ein und klicken auf OK.



Es entsteht nun folgende Grafik

Korrelationsmatrix



Die Gestaltungsmöglichkeiten des Almo-Grafik-Editors sind noch keineswegs erschöpft. Klicken Sie in der linken Leiste auf den Knopf „Diverse Positionen“. Es erscheint dann folgende Auswahl.

			<p><b>Matrix</b> (=Balken von oben)</p> <table border="1"> <thead> <tr> <th></th> <th>Wohnort</th> <th>Beruf</th> <th>Produkt</th> <th>Rueckzahl</th> <th>Einkommen</th> <th>Rueckrate</th> <th>Laufzeit</th> </tr> </thead> <tbody> <tr> <th>Wohnort</th> <td>1</td> <td>-0.667</td> <td>0.679</td> <td>-0.500</td> <td></td> <td></td> <td></td> </tr> <tr> <th>Beruf</th> <td>-0.667</td> <td>1</td> <td>-0.500</td> <td>-0.667</td> <td></td> <td></td> <td></td> </tr> <tr> <th>Produkt</th> <td>0.679</td> <td>-0.500</td> <td>1</td> <td>0.679</td> <td></td> <td></td> <td></td> </tr> <tr> <th>Rueckzahl</th> <td>-0.500</td> <td>-0.667</td> <td>0.679</td> <td>1</td> <td></td> <td></td> <td></td> </tr> </tbody> </table>		Wohnort	Beruf	Produkt	Rueckzahl	Einkommen	Rueckrate	Laufzeit	Wohnort	1	-0.667	0.679	-0.500				Beruf	-0.667	1	-0.500	-0.667				Produkt	0.679	-0.500	1	0.679				Rueckzahl	-0.500	-0.667	0.679	1				Abbruch
	Wohnort	Beruf	Produkt	Rueckzahl	Einkommen	Rueckrate	Laufzeit																																					
Wohnort	1	-0.667	0.679	-0.500																																								
Beruf	-0.667	1	-0.500	-0.667																																								
Produkt	0.679	-0.500	1	0.679																																								
Rueckzahl	-0.500	-0.667	0.679	1																																								
				Weitere Auswahl																																								
				Zurück																																								
				Abbruch																																								

Klicken Sie auf das Bild „Matrix (= Balken von oben)“. Das ist das letzte Bild in der 2. Reihe. Almo erzeugt dann Balken, die (ohne Perspektive) von oben betrachtet werden. Es entsteht dabei der graphische Eindruck einer Matrix. Sie sehen folgendes:

Korrelationsmatrix

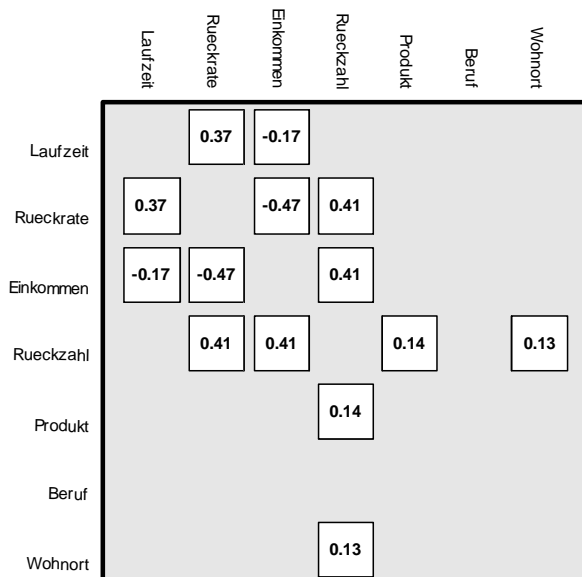
	Laufzeit	Rueckrate	Einkommen	Rueckzahl	Produkt	Beruf	Wohnort
Laufzeit	1	0.37	-0.17	0.08	0.02	0.04	0.02
Rueckrate	0.37	1	-0.47	0.41	0.02	0	0.01
Einkommen	-0.17	-0.47	1	0.41	0.01	0	0.01
Rueckzahl	0.08	0.41	0.41	1	0.14	0.07	0.13
Produkt	0.02	0.02	0.01	0.14	1	0.01	0.04
Beruf	0.04	0	0	0.07	0.01	1	0.04
Wohnort	0.02	0.01	0.01	0.13	0.04	0.04	1

Es ist sinnvoll mit dem Schieber „Größe“ und eventuell dem Schieber „Breite“ in der linken Leiste des Grafik-Bildschirms die Bildgröße anzupassen.

Wenn eine Korrelationsmatrix aus sehr vielen Variablen besteht, dann muß das Bild so stark vergrößert werden, daß es am linken und rechten Rand, aber auch oben und unten verschwindet. Das Bild kann jedoch gescrollt werden. Beim Ausdrucken werden die nicht sichtbaren Teile, sofern sie überhaupt noch auf das Papier passen, gedruckt. Auch der Titel „Korrelationsmatrix“ im linken oberen Eck wandert beim Drucken ganz nach oben auf das linke obere Eck des Papiers.

Wieder bestünde nun die Möglichkeit in der 3. Leiste, nur Balken zu zeigen mit einem Absolutwert größer 0.09 und kleiner 1.0. Es entsteht folgende Grafik:

Korrelationsmatrix



Beachten Sie, daß die Korrelationsmatrix symmetrisch ist. Die Werte unter- und oberhalb der Diagonalen entsprechen sich.

Wie ist unser Ergebnis nun inhaltlich zu interpretieren?

Wie gehen eine Zeile der Matrix nach der andern durch. Betrachten wir zuerst die Zeile mit der Variablen „Laufzeit“.

	Wohnort	Beruf	Produkt	Rueckzah	Einkomme	Rueckrat	Laufzeit
Laufzeit	-	-	-	-	-0.17	0.37	-

Die Laufzeit korreliert negativ mit  $-0.17$  mit dem Einkommen. Je höher das Einkommen umso kürzer die Laufzeit des Kredits. Mit der Rückzahlungsrate korreliert die Laufzeit relativ hoch mit  $0.37$ . Kunden, die hohe Rückzahlungsraten vereinbart haben, haben auch eine längere Kredit-Laufzeit.

Dann betrachten wir die nächste Zeile der Korrelationsmatrix usw.

Beim Betrachten und Interpretieren der Korrelationskoeffizienten werden sich uns mehrere Fragen aufdrängen, etwa: Wie beeinflusst die Laufzeit das Rückzahlungsverhalten, und überhaupt: Was sind die Ursachen dafür, daß der Kredit (zeitgerecht) zurückgezahlt bzw. nicht zurückgezahlt wird? Das führt uns nun zum nächsten Kapitel des Data-Mining-Prozesses, dem „gezielten“ Suchen nach Zusammenhängen.

## P45.12.4 Das "paarweise Ausscheiden" zur Lösung des Kein-Wert-Problems

Nicht selten sind Daten unvollständig. Für manche Untersuchungseinheiten fehlen Werte in der Variablen i für andere in der Variablen j und wieder für andere in der Variablen k etc. Wie sollen diese Variable miteinander korreliert werden. In Abschnitt P45.6 haben wir ausgeführt, daß es dafür in Almo 2 prinzipiell verschiedene Vorgehensweisen gibt.

1. Man sucht Ersatzwerte für die fehlenden Werte und erzeugt eine neue vollständige Datei. Almo stellt dafür 3 Programme zur Verfügung.
2. Datensätze, die auch nur in einer Analysevariablen keinen Wert besitzen, werden ausgeschlossen. Oder: Man führt das „paarweise Ausscheiden“ durch. Da dieses sehr gerne und häufig angewendet wird, wollen wir es im Folgenden genauer betrachten.

Das "paarweise Ausscheiden" kann bei allen Verfahren verwendet werden, bei denen eine Korrelationsmatrix oder allgemein: eine Streuungsmatrix (als Endergebnis oder als Zwischenschritt) errechnet werden muß.

In Almo sind dies:

Korrelationsmatrix:	Prog19
Allgemeines Lineares Modell (ALM):	Prog20
Pfadanalyse:	Prog25
Kanonische Korrelation:	Prog29
Kanonische Diskriminanzanalyse:	Prog29
Bivariate kanonische Korrespondenzanalyse:	Prog29
Faktorenanalyse:	Prog30
Multiple Korrespondenzanalyse:	Prog30

Für unsere Darstellung verwenden wir folgende Beispiel-Daten

*Tabelle 1: Beispieldaten*

Person	x1	x2	x3
1	kw	5	6
2	1	kw	3
3	9	9	kw
4	4	7	kw
5	4	7	5
6	1	1	1
7	5	6	5
8	5	4	4
9	3	5	4
10	2	5	3

Die Datei umfasst 10 Untersuchungseinheiten (Personen) und die 3 Variablen x1, x2, x3. Fehlende Werte wurden mit "kw" (=KeinWert) symbolisiert.

Die 1. Person besitzt in x1 keinen Wert.

Die 2. Person besitzt in x2 keinen Wert.

Die 3. und die 4. Person besitzen in x3 keinen Wert.

Diese Datei ist unter dem Namen "PaarwKW.fre" und "PaarwKW.dir" sowie als SPSS-File unter "PaarwKW.sav" im Almo-Ordner TESTDAT enthalten. Der Benutzer kann mit diesen Dateien unsere folgenden Berechnungen nachvollziehen bzw. selbst experimentieren.

**P45.12.4.1 Die Berechnung einer Korrelationsmatrix**

Wenn wir als Kein-Wert-Behandlung =3, das "vollständige Ausscheiden" wählen dann werden nur die Personen 5 bis 10 ausgewertet, also 6 Personen.

Wir erhalten folgendes Ergebnis:

*Tabelle 2: Korrelationsmatrix bei "vollständigem Ausscheiden"*

	x1	x2	x3
	v1	v2	v3
x1	1.0000	0.6325	0.8677
x2	0.6325	1.0000	0.9218
x3	0.8677	0.9218	1.0000

Wird als Kein-Wert-Behandlung =1, das "paarweise Ausscheiden" gewählt, dann erhalten wir folgendes Ergebnis:

*Tabelle 3: Zahl der Einheiten, die in die Analyse eingegangen sind je Zelle der Streuungsmatrix bei "paarweisem Ausscheiden"*

	x1	x2	x3
x1	9	8	7
x2	8	9	7
x3	7	7	8

Für das Variablenpaar x1 x2 werden nur die Personen, die in beiden Variablen valide Daten besitzen, ausgewertet. Das sind die Personen 3 bis 10, also 8 Personen. Für das Variablenpaar x1 x3 werden nur die Personen 2, 5 bis 10, also 7 Personen ausgewertet Für das Variablenpaar x2 x3 werden nur die Personen 1, 5 bis 10, also 7 Personen ausgewertet

*Tabelle 4: Korrelations-Matrix bei "paarweisem Ausscheiden"*

	x1	x2	x3
x1	1.0000	0.7790	0.8264
x2	0.7790	1.0000	0.8101
x3	0.8264	0.8101	1.0000

Jede einzelne Korrelation  $r_{ik}$  zwischen den Variablen i und k wird nur aus den Personen errechnet, die in beiden Variablen einen validen Wert besitzen.

Die Berechnungsformel ist

$$(1) \quad r_{ik} = \text{COV}_{ik} / (s_i * s_k)$$

$\text{COV}_{ik}$  = Kovarianz zwischen Variablen i und k

$s_i$  = Standardabweichung der Variablen i berechnet aus den Personen, die in i und k einen validen Wert besitzen

$s_k$  = Standardabweichung der Variablen k berechnet aus den Personen, die in i und k einen validen Wert besitzen

Almo errechnet folgende Standardabweichungen:

*Tabelle 5: Standardabweichungen der Variablen bei "paarweisem Ausscheiden"*

	x1	x2	x3
Standabwg. der Variablen x1	-	2.26039	1.60357
Standabwg. der Variablen x2	2.23607	-	1.74964
Standabwg. der Variablen x3	1.29363	1.51186	-

In der 1. Zeile stehen die Standardabweichungen der Variablen x1.

Im Variablenpaar x1 x2 beispielsweise ist die Standardabweichung von x1= 2.26039

In der 2. Zeile stehen die Standardabweichungen der Variablen x2.

Im Variablenpaar x1 x2 beispielsweise ist die Standardabweichung von x2= 2.23607

Die Standardabweichungen sind mit  $n_{ik}$ , nicht mit  $n_{ik}-1$  dividiert. Das kann aber über eine Option umgestellt werden.

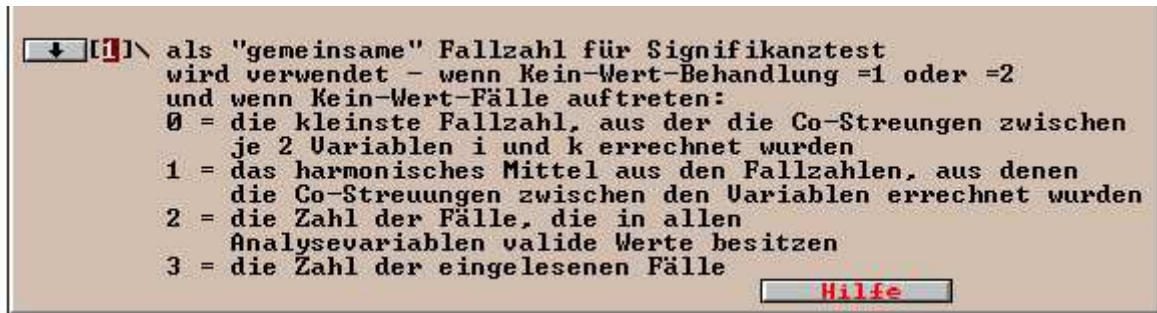
Für jeden Korrelationskoeffizient kann dann noch die Signifikanz ermittelt werden, bei deren Berechnung die jeweils verschiedenen Fallzahlen verwendet werden.

*Tabelle 6: Signifikanzen p bei "paarweisem Ausscheiden"*

	x1	x2	x3
x1	-	0.0223	0.0220
x2		-	0.0271
x3			-

Wir erkennen, dass die Korrelation  $r_{x1x2}=0.779$  mit  $p=0.0223$  eine bessere Signifikanz aufweist als die höhere Korrelation  $r_{x2x3}=0.8101$  mit  $p=0.0271$ . Im Prinzip ist es möglich, dass eine Variable i mit der Variablen j eine kleine Korrelation aufweist, die signifikant ist und mit der Variable k eine hohe Korrelation aufweist, die nicht signifikant ist - eben weil sie auf unterschiedlichen Häufigkeiten beruhen. Die Korrelationskoeffizienten sind also nicht vergleichbar. Das ist nicht sinnvoll und macht das "paarweise Ausscheiden" im Prinzip unbrauchbar.

Man wird so verfahren müssen, dass man doch eine Vergleichbarkeit unterstellt und für die Ermittlung der Signifikanz eine gemeinsame Häufigkeit einsetzt. Almo bietet 4 Möglichkeiten an. Um sie zu nutzen muß man im Prog45m6 die Optionsbox 9 „Spezielle Kein-Wert-Behandlung“ öffnen. Man sieht dann eine sehr große Eingabe-Box, die wir bereits in Abschnitt P45.7.3, Eingabe-Box 9 (für die Methode 1 bis 3) und in P45.6.1, Eingabe-Box 7 (für die Methode 4 bis 7) dargestellt haben. Im unteren Teil dieser großen Eingabe-Box ist folgendes zu sehen.



Als gemeinsame Fallzahl für Signifikanztests kann verwendet werden:

### Möglichkeit

- 0 das kleinste  $n_{ik}$  aus dem unteren oder oberen Dreieck der obigen Tabelle 3 der Zahl der Einheiten, die in die Analyse eingegangen sind (ohne Diagonale)
- 1 das harmonisches Mittel aus der obigen Tabelle 3 der Zahl der Einheiten
- 2 die Zahl der Fälle, die in allen Analysevariablen valide Werte besitzen. Diese Zahl ist identisch mit der Zahl der Fälle, wie sie beim "vollständigen Ausscheiden" vorhanden wären.
- 3 die Zahl der eingelesenen Fälle

Eine unseres Erachtens sichere Vorgehensweise ist es, die Möglichkeit 2 zu wählen. Dann werden die Korrelationen aus den paarweise vorhandenen Daten errechnet, die Daten also optimal ausgeschöpft. Dabei wird unterstellt, dass die Korrelationskoeffizienten alle aus dem  $n$  errechnet wurden, das beim "vollständigen Ausscheiden" entstehen würde. Damit sind sie vergleichbar.

Fehlen in einigen Variablen nur wenige in anderen, aber sehr viele, dann ist es eher sinnvoll die Möglichkeit 1, das harmonische Mittel zu wählen.

Mit diese Unterstellung ist das "paarweise Ausscheiden", wenn es um die Berechnung einer Korrelationsmatrix geht, eine plausible Vorgehensweise.

#### ***P45.12.4.2 Die Berechnung einer "Quasi-Korrelationsmatrix" bei "paarweisem Ausscheiden"***

Wird als Streuungsmatrix "Quasi\_Korrelation" eingestellt und als Kein-Wert-Behandlung =1, das "paarweise Ausscheiden", dann errechnet Almo die einzelnen Korrelationen ebenfalls nach der oben unter (1) angegebenen Formel. Die Kovarianz  $cov_{ik}$  wird aus den Personen ermittelt, die in beiden Variablen i und k einen validen Wert haben. Das geht nicht anders. Die Standardabweichungen  $s_i$  und  $s_k$  sind nun anders definiert:

- $s_i$  = Standardabweichung der Variablen i  
 berechnet aus den Personen, die in i (und nur in i) einen validen Wert besitzen
- $s_k$  = Standardabweichung der Variablen k  
 berechnet aus den Personen, die in k (und nur in k) einen validen Wert besitzen

Die Daten werden also besser ausgenützt. Die Standardabweichung der Variablen i wird aus allen Personen, die in der Variablen i einen validen Wert besitzen, ermittelt und nicht nur aus jenen, die in i und k einen validen Wert besitzen.

Almo ermittelt folgende Standardabweichungen:

Standardabweichung von Variable x1: 2.3465  
Standardabweichung von Variable x2: 2.1140  
Standardabweichung von Variable x3: 1.4524

Die Standardabweichungen sind mit  $n$ , nicht mit  $n-1$  dividiert. Das kann aber über eine Option umgestellt werden.

Tabelle 7: Quasi-Korrelationsmatrix

	x1	x2	x3
x1	1.0000	0.7938	0.5030
x2	0.7938	1.0000	0.6979
x3	0.5030	0.6979	1.0000

Die Unterschiede zur Korrelationsmatrix in Tabelle 4 sind teilweise erheblich. Das erklärt sich dadurch, dass wir nur 10 Untersuchungseinheiten haben und im Verhältnis zu diesen relativ viele Kein-Wert-Fälle. Bei größeren Datenmatrizen und prozentual weniger Kein-Wert-Fällen werden die Unterschiede minimal.

Bei der Quasi-Korrelationsmatrix werden die Daten besser ausgenützt. Die Standardabweichung der Variablen  $i$  wird aus allen Personen, die in der Variablen  $i$  einen validen Wert besitzen, ermittelt und nicht nur aus jenen, die in  $i$  und  $k$  einen validen Wert besitzen.

Dafür muß allerdings ein Preis bezahlt werden: Es können Korrelationskoeffizienten entstehen, die außerhalb des Bereichs 0 - 1 liegen. Deswegen sprechen wir auch von "Quasi-Korrelation".

Bei empirischen Daten wird dies sehr selten sein. Wir haben das noch nie erlebt.

Einige Autoren weisen auch darauf hin (so Little/Rubin 1990, S. 379), dass die Korrelationsmatrix mit "paarweisem Ausscheiden" und die Quasi-Korrelationsmatrix so gestaltet sein können, dass sie nicht invertierbar sind, mit der Folge, daß der Kalkül des ALM nicht anwendbar ist. Auch das haben wir noch nie erlebt. Bei empirischen Daten wird dieser Fall nicht auftreten.

Wir können folgendes Fazit ziehen:

Für das "paarweise Ausscheiden" gibt es Varianten. Es ist durchaus eine plausible Vorgehensweise, wenn Kein-Wert-Fälle vorliegen. Es ist aber keinesfalls frei von Problemen.

#### ***P45.12.4.3 Die Berechnung einer Quadratsummen- oder einer Kovarianzmatrix bei "paarweisem Ausscheiden"***

Wir rechnen nun mit denselben Daten eine Quadratsummenmatrix, zuerst wieder mit Kein-Wert-Behandlung =3, dem "vollständige Ausscheiden". Dabei werden nur die Personen 5 bis 10 ausgewertet, also 6 Personen.

Wir erhalten folgendes Ergebnis:

Tabelle 8: Quadratsummen-Matrix bei "vollständigem Ausscheiden"

	x1	x2	x3
	v1	v2	v3
x1	13.3333	10.6667	10.6667
x2	10.6667	21.3333	14.3333
x3	10.6667	14.3333	11.3333

Wird als Kein-Wert-Behandlung =1, das "paarweise Ausscheiden" gewählt, dann erhalten wir zunächst wieder die Matrix der

Zahl der Einheiten, die in die Analyse eingegangen sind  
je Zelle der Streuungsmatrix

die wir bereits oben in Tabelle 3 abgebildet haben. Also benötigt nun für den weiteren Rechengang ein einziges gemeinsames  $n$ . Wir bezeichnen es mit  $n_g$ . Dieses  $n_g$  wird für die Ermittlung der Signifikanzen in Programmen gebraucht, in die die Streuungsmatrix als Eingabe eingeht - z.B. das ALM. Also bietet hier 4 Alternativen an. Als gemeinsame Fallzahl wird verwendet:

### Möglichkeit

- 0 das kleinste  $n_{ik}$  aus dem unteren oder oberen Dreieck der obigen Tabelle 3 der Zahl der Einheiten, die in die Analyse eingegangen sind (ohne Diagonale)
- 1 das harmonisches Mittel aus der obigen Tabelle 3 der Zahl der Einheiten
- 2 die Zahl der Fälle, die in allen Analysevariablen valide Werte besitzen. Diese Zahl ist identisch mit der Zahl der Fälle, wie sie beim "vollständigen Ausscheiden" vorhanden wären.
- 3 die Zahl der eingelesenen Fälle

Die vorsichtigste Alternative ist =2. Für die Signifikanztests (etwa im Rahmen des ALM) 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.

Wenn wir die Alternative	0 wählen, dann ist $n_g =$	7
	1	7
	2	6
	3	10

Beim Maskenprogramm kann diese Option in der entsprechend Eingabe-Box eingesetzt werden. Ist die Eingabe-Box nicht vorhanden, dann verwendet Almo in der Regel =1. Beim in der Almo-Programmiersprache "selbst geschriebenen" Programm kann eine dieser 4 Möglichkeiten über die Option 78 eingegeben werden.

Almo berechnet zuerst für jedes Variablenpaar die Kreuzprodukte.

Tabelle 9: Matrix der Kreuzprodukte bei "paarweisem Ausscheiden"

	x1	x2	x3
x1	178	213	87
x2	213	308	137
x3	87	147	-

Für das Variablenpaar x1 x2 beispielsweise werden nur die Personen, die in beiden Variablen valide Daten besitzen, ausgewertet. Das sind die Personen 3 bis 10, also 8 Personen. Die Matrix ist selbstverständlich symmetrisch. Entsprechend wird für die anderen Variablenpaare verfahren.

In der Diagonale der Matrix stehen die Kreuzprodukte der Variablen "mit sich selbst". Sie werden aus den Personen gebildet, die in der Variablen valide Werte besitzen, bei x1 und x2 sind das 9 Personen, bei x3 sind es 8.

Almo berechnet dann noch für jedes Variablenpaar die Wertesummen.

Tabelle 10: Matrix der Wertesummen bei "paarweisem Ausscheiden"

	x1	x2	x3
Wertesumme der Variablen x1	34	33	21
Wertesumme der Variablen x2	44	49	33
Wertesumme der Variablen x3	25	28	31

In der 1. Zeile stehen die Wertesummen der Variablen x1. In der Zelle x1 x2 steht 33. Das ist die Summe der Variablenwerte der Variablen x1, wenn sie mit x2 gepaart wird. Für diese Summe werden nur die x1-Werte verwendet, für die x1 und x2 je einen validen Wert besitzen. Das sind die Personen 3 bis 10.

In der Zelle x1 x3 steht 21. Das ist die Wertesumme der Variablen x1, wenn sie mit x3 gepaart wird.

Die Variable x1 hat also ungleiche Wertesummen - je danach mit welcher anderen Variablen sie gepaart wird. Entsprechendes gilt für die anderen Variablen x2 und x3.

In der Zeile i der Matrix steht also die Variable i, deren Wertesummen wir betrachten. In der Spalte k der Matrix steht die Variable k, mit der i gepaart wird.

In der Diagonale der Matrix stehen die Wertesummen der Variablen.

Die Quadratsumme für die Zelle ik ergibt sich dann gemäß folgendem Ausdruck:

$$(2) \quad Q_{ik} = K_{ik} - W_{ik} \cdot W_{ki} / n_{ik}$$

Für die Zelle x1 x2 beispielweise

$$Q_{12} = 213 - 33 \cdot 44 / 8 = 31.5$$

- $Q_{ik}$  = Quadratsumme für das Variablenpaar  $x_i x_k$
- $K_{ik}$  = Kreuzprodukt für das Variablenpaar  $x_i x_k$
- $W_{ik}$  = Wertesumme der Variablen  $x_i$ , wenn sie mit  $x_k$  gepaart wird
- $W_{ki}$  = Wertesumme der Variablen  $x_k$ , wenn sie mit  $x_i$  gepaart wird
- $n_{ik}$  = valide Häufigkeit für das Variablenpaar  $x_i x_k$

Für das Diagonalglied  $ii$  ergibt sich

$$(2a) \quad Q_{ii} = K_{ii} - W_{ii} \cdot W_{ii} / n_{ii}$$

So erhalten wir die

*Tabelle 11: Quadratsummenmatrix bei "paarweisem Ausscheiden"*

	x1	x2	x3
x1	49.56	31.50	12
x2	31.50	40.22	15
x3	12	15	16.88

Die Matrix ist symmetrisch. Die Quadratsumme  $Q_{ik}$  in der Zelle  $ik$  der Matrix ist nur aus den Personen errechnet worden, die für beide Variable  $i$  und  $k$  valide Werte besitzen. Die Quadratsummen beruhen also auf unterschiedlichen Häufigkeiten. Die Matrix ist in dieser Form nicht zu gebrauchen.

Wir dividieren nun die einzelnen Quadratsummen durch die entsprechenden Häufigkeiten. So erhalten wir die "durchschnittliche" Quadratsummen-Matrix. Das ist die Kovarianzmatrix. Für die Zelle x1 x2 rechnen wir also  $31.5/8=3.9375$

*Tabelle 12: Kovarianz-Matrix bei "paarweisem Ausscheiden"*  
(Kovarianz ist mit  $n$  dividiert)

	x1	x2	x3
x1	5.5062	3.9375	1.7143
x2	3.9375	4.4691	2.1429
x3	1.7143	2.1429	2.1094

Die "durchschnittlichen" Quadratsummen wird mit  $ng$  multipliziert. " $ng$ " ist die gemeinsame Fallzahl. Wir haben oben gezeigt, dass es 4 Möglichkeiten gibt,  $ng$  zu bestimmen. Wir wählen die Möglichkeit "1", das harmonische Mittel der unterschiedlichen Fallzahlen. Die gemeinsame Fallzahl  $ng$  ist dann 7.

Aus der Multiplikation erhalten wir die auf  $ng$  Personen hochgerechnete Quadratsummen-Matrix, die Almo schließlich als Ergebnis seiner Berechnungen ausgibt. Wird die Quadratsummen-Matrix als Zwischenschritt für Verfahren wie z.B. das ALM verwendet, dann ist es vollkommen gleichgültig, mit welcher konstanten Zahl multipliziert wird. Man kann auch immer das Multiplizieren unterlassen und die Kovarianz-Matrix verwenden.

Tabelle 13: "Hochgerechnete" Quadratsummen-Matrix bei "paarweisem Ausscheiden" mit dem harmonischen Mittel als gemeinsamer Fallzahl

	x1	x2	x3
	V1	V2	V3
x1	38.5432	27.5625	12.0000
x2	27.5625	31.2840	15.0000
x3	12.0000	15.0000	14.7656

**P45.12.4.4. Das "paarweise Ausscheiden" beim Allgemeinen Linearen Modell (ALM)**

Wir rechnen mit unseren Beispieldaten ein ALM mit "vollständigem Ausscheiden" und der Korrelationsmatrix aus Tabelle 2.  $X_1$  und  $x_2$  sind die unabhängigen Variablen,  $x_3$  ist die abhängige Variable. Es entsteht folgendes Ergebnis

Variable	Regr. koeff.	Standard fehler	95% Konfidenzbereich nach		erklärte Streuung	part. Korrel.	F-Wert	Signifikanz p	df1	df2	Test-stärke
			oben u. unten								
x1	0.4745	0.0918	0.2921	0.1351	0.948	26.727	0.012	98.76	1	3	0.9230
x2	0.6217	0.0918	0.2921	0.2319	0.969	45.873	0.005	99.45	1	3	0.9924

Die Regressionskoeffizienten in der 1. Spalte sind standardisiert.

Nun rechnen wir ein ALM mit "paarweisem Ausscheiden" und der Korrelationsmatrix aus Tabelle 4. Es entsteht folgendes Ergebnis.

Variable	Regr. koeff.	Standard fehler	95% Konfidenzbereich nach		erklärte Streuung	part. Korrel.	F-Wert	Signifikanz p	df1	df2	Test-stärke
			oben u. unten								
x1	0.4968	0.3961	1.0998	0.0970	0.531	1.573	0.278	72.19	1	4	0.1637
x2	0.4231	0.3961	1.0998	0.0704	0.471	1.141	0.347	65.30	1	4	0.1322

Die Regressionskoeffizienten in der 1. Spalte sind standardisiert. Die nicht-standardisierten Regressionskoeffizienten erhält man gemäß folgender Formel

$$(3) \quad b = \beta * s_y / s_x$$

- b = nicht-standardisierter Regressionskoeffizient
- $\beta$  = standardisierter Regressionskoeffizient
- $s_y$  = Standardabweichung der abhängigen Variablen
- $s_x$  = Standardabweichung der unabhängigen Variablen

Welche Standardabweichungen soll aber nun verwendet werden? Zwei Möglichkeiten bestehen hier

**Möglichkeit 1: "Vorhandene Standardabweichungen"**

Man verwendet die Standardabweichung, wie sie sich aus der Zahl der Fälle ergibt, für die valide Werte für die Variable vorhanden sind. So wird beispielsweise in der Regressionsfunktion von SPSS verfahren (Stand: Version 10).

**Möglichkeit 2: "Paarweise Standardabweichungen"**

Naheliegender ist es, die Standardabweichungen zu verwenden, die sich aus der Zahl der Fälle ergibt, für die valide Werte für das Variablenpaar xy vorhanden sind. Mit diesen Standardabweichungen wurde ja die Korrelationsmatrix gebildet. Siehe oben Tabelle 5.

Wir rechnen zuerst gemäß Möglichkeit 1.

Die Standardabweichungen gemäß 1 sind folgende (Standardabweichung ist mit n, nicht mit n-1 dividiert)

x1	2.3465
x2	2.1140
x3	1.4524

Wir erhalten also folgende nicht-standardisierte Regressionskoeffizienten.

$$\begin{aligned}
 & \beta \quad \quad \quad s_y \quad \quad \quad s_x \\
 & \text{-----} \quad \text{-----} \quad \text{-----} \\
 b1 &= 0.4968 * 1.4524 / 2.3465 = 0.3075 \\
 b2 &= 0.4231 * 1.4524 / 2.1140 = 0.2907
 \end{aligned}$$

SPSS (Version 10) rechnet die Standardabweichungen mit n-1 im Nenner und erhält dadurch etwas andere Ergebnisse. In Almo kann über eine Option der Nenner ebenfalls auf n-1 gesetzt werden. Die Ergebnisse sind dann voll identisch. Das Allgemeine Lineare Modell in SPSS (Version 10), dort GLM genannt, beherrscht nur das "vollständige Ausscheiden", kann also zu Vergleichszwecken nicht herangezogen werden.

Wir rechnen nun gemäß oben angegebener Möglichkeit 2 ("paarweise Standardabweichungen").

Die Standardabweichungen sind nun für jedes Variablenpaar verschieden. Wir haben sie oben in Tabelle 5 angegeben. Wir erhalten folgende nicht-standardisierte Regressionskoeffizienten.

$$\begin{aligned}
 & \beta \quad \quad \quad s_y \quad \quad \quad s_x \\
 & \text{-----} \quad \text{-----} \quad \text{-----} \\
 b1 &= 0.4968 * 1.29363 / 1.60357 = 0.4008 \\
 b2 &= 0.4231 * 1.51186 / 1.74964 = 0.3656
 \end{aligned}$$

#### **P45.12.4.5 ALM auf Quasi-Korrelationsmatrix angewendet**

Nun rechnen wir ein ALM mit "paarweisem Ausscheiden" und der Quasi-Korrelationsmatrix aus Tabelle 7. Es entsteht folgendes Ergebnis

Variable	Regr. koeff.	Standard fehler	95% Konfidenzbereich nach		erklärte Streuung	part. Korrel.	F-Wert	Signifikanz p	df1	df2	Test-stärke
			oben	u.unten							
x1	-0.1377	0.5847	1.6233	0.0070	-0.117	0.055	0.808	19.23	1	4	0.0539
x2	0.8072	0.5847	1.6233	0.2411	0.568	1.906	0.239	76.09	1	4	0.1879

Das Ergebnis ist deutlich anders als die obigen beiden Ergebnisse. Unsere Beispieldaten umfassen nur 10 Datensätze, wobei in 4 Datensätze Kein-Wert-Fälle auftreten. D.h. der Anteil der Kein-Wert-Fälle an der Zahl aller Datensätze ist sehr hoch. Wäre er niedriger, dann würden sich das Ergebnis sehr viel mehr an die beiden obigen annähern.

Die Regressionskoeffizienten in der 1. Spalte sind standardisiert. Die nicht-standardisierten Regressionskoeffizienten erhält man gemäß obiger Formel 3. Jetzt ist es aber nur sinnvoll, die Standardabweichungen entsprechender der oben genannten Möglichkeit 1 ("vorhandene Standardabweichungen") einzusetzen, da ja im Nenner der Korrelationsformel (siehe oben Gleichung 1) auch diese stehen.

$$\begin{aligned}
 & \beta & s_y & s_x \\
 & \text{-----} & \text{-----} & \text{-----} \\
 b1 & = -0.1377 * 1.4524 / 2.3465 = -0.0853 \\
 b2 & = 0.8072 * 1.4524 / 2.1140 = 0.5546
 \end{aligned}$$

#### ***P45.12.4.6 ALM auf Kovarianz- bzw. Quadratsummenmatrix angewendet***

Nun rechnen wir ein ALM mit "paarweisem Ausscheiden" und der Quadratsummenmatrix aus Tabelle 13. Mit Ausnahme der "erklärten Streuung" würde dasselbe Ergebnis entstehen, wenn wir die Kovarianzmatrix aus Tabelle 12 verwenden würden.

Es entsteht folgendes Ergebnis:

Variable	Regr. koeff.	Standard fehler	95% Konfidenzbereich nach		erklärte Streuung	part. Korrel.	F-Wert	Signifikanz p	df1	df2	Test-stärke
			oben	u.unten							
x1	-0.0853	0.3619	1.0048	0.1036	-0.117	0.055	0.808	19.23	1	4	0.0539
x2	0.5546	0.4017	1.1153	3.5598	0.568	1.906	0.239	76.09	1	4	0.1879

Es entsteht dasselbe Ergebnis, wie bei der Quasi-Korrelationsmatrix - nur dass unmittelbar die nicht-standardisierten Regressionskoeffizienten entstehen. Man erkennt, dass die partiellen Korrelationen, die F-Werte und die Signifikanzen identisch sind.

Das bedeutet: Das paarweise Ausscheiden bei der Kovarianz- bzw. Quadratsummenmatrix ist äquivalent dem bei der Quasi-Korrelationsmatrix, nicht jedoch dem bei der Korrelationsmatrix. Es ist beim paarweisen Ausscheiden nicht möglich, die Quadratsummenmatrix so zu modifizieren, daß sie die gleichen Ergebnisse erbringt wie die Korrelationsmatrix.

## **Kapitel 7: Einzelne Zusammenhänge genauer untersuchen**

Durch das Korrelieren mit Prog45m6 sind wir auf Variable gestoßen, die uns besonders interessieren. Sie werden für uns zu „Zielvariablen“, deren Zusammenhänge mit anderen Variablen wir untersuchen wollen.

Nun möchten wir einzelne Zusammenhänge zwischen der Zielvariablen und einer (oder auch mehreren) ursächlichen Variablen in Form von zwei- oder mehrdimensionalen Tabellen oder in Form von zwei- oder dreidimensionalen Streudiagrammen genauer betrachten.

### ***P45.13 Schritt 9a: Variable zwei- und beliebig-dimensional tabellieren***

Wir verwenden das nachstehend dargestellte Programm Prog45mb.

#### **Eingabe mit Maskenprogramm Prog45mb**

Prog45mb.Msk  
Tabellierung

1. 2-dimensionale Tabelle
2. 2-dimensionale Mehrfach-Tabelle
3. Partialtabellen (Interaktionstabelle)
4. multidimensionale Kontingenztabelle
5. Mittelwerts-Tabelle

Berechnet werden folgende Koeffizienten:  
Chi-Quadrat, Chi-Quadrat-Test und 2I-Test auf allseitige  
Abhängigkeit, Kontingenzkoeffizient C(cor), Cramers U

Zu 4. kann die Tabelle der Erwartungswerte zum Chi-Quadrat  
und die Tabelle der Chi-Quadrat-Beiträge ermittelt werden.

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

1    
Vereinbare Variable=  ;

2  Option: Weitere Vereinbarungen - nur wenn Almo dazu auffordert

3    
  "C:\Almo7\Testdat\DatMin.nam"  
  **zeige**                    zeige = Namensdatei in Output zeigen  
leer = nicht

4    
   
 **erzeuge zusätzliche Namensfelder**

5    
 "C:\Almo7\Testdat\DatMin.dir"

6

**2-dimensionale Tabelle** **Hilfe**

Beispiel:

		Rückzahlung		Summe
		nein	ja	
Einkommen	bis 10000	100	66	166
	10- 20000	86	123	209
	20- 30000	68	184	252
	30- 40000	23	188	211
	40- 50000	9	153	162
Summe		286	714	1000

**Einkommen mit Rueckzahl**

erzeuge zusätzliche Felder für Tabellen-Angaben

7

**2-dimensionale "Mehrfach-Tabelle"** **Hilfe**

Beispiel:

		Rueckzahlung		Summe
		nein	ja	
Wohnort	Stadt	126	215	341
	Land	160	499	659
Geschlecht	männlich	150	363	513
	weiblich	136	351	487
Beruf	Selbständig	52	93	145
	Unselbst.	234	621	855
Summe		286	714	1000

**Wohnort+Geschlecht+Beruf mit Rueckzahl**

erzeuge zusätzliche Felder für Tabellen-Angaben

8

**Partialtabellen**  
**Interaktionstabelle**

**Hilfe**

Beispiel:

Geschlecht	Produkt	Rückzahlung		Summe
		nein	ja	
männlich	Kleidung	55	45	100
	Möbel	29	71	100
	Technik	23	77	100
weiblich	Kleidung	35	65	100
	Möbel	32	68	100
	Technik	20	80	100
Summe		35	65	100

**Geschlecht\*Produkt mit Rueckzahl**

erzeuge zusätzliche Felder für Tabellen-Angaben

9

**Multidimensionale Kontingenztabelle**

**Hilfe**

Beispiel:

Wohnort	Beruf	Rueckzahl		Fälle
		nein	ja	
Stadt	Selbständ	27	29	99
	Unselbst.	186	135	
Land	Selbständ	25	64	135
	Unselbst.	435	1000	
Summe				1000

**Wohnort\*Beruf\*Rueckzahl**

erzeuge zusätzliche Felder für Tabellen-Angaben

10

**Mittelwerts-Tabelle** **Hilfe**

Beispiel:  
Mittelwerte der Kredit-Laufzeit  
je Geschlecht und Beruf

Geschl.	Beruf	Kredit-Laufzeit
männl	Selbstän	14.1481
	Unselbst	14.4954
weibl	Selbstän	14.4063
	Unselbst	15.2270
Gesamtmittel		14.7710

**Geschlecht, Beruf**      Zeilen-Variablen  
  **Laufzeit**      Spalten-Variablen: quantitativ  
        Spalten-Variablen: ordinal

11

**Variable aus Tabellenangaben**

**U1:4,8:10**

Almo ermittelt die Variablen, die in obige Tabellen eingegangen sind selbst. Sie können aber auch auf diesen Knopf klicken

12

Option: Nur für multidimensionale Kontingenztabelle

13

Option: Ein- und Ausschliessen von Untersuchungseinheiten

14

Loesche wieder diese Box

**Umkodierungen und Kein-Wert-Angaben**

Umkodierungen **Hilfe**  
Kein\_Wert-Angabe **Hilfe**

**Einkommen <0 Schritt 10000 bis 50000 = I>**

erzeuge zusätzliche Felder für Umkodierungen / Kein\_Wert-Angaben

---

Kontrollieren, ob Umkodierung so erfolgt wie gewünscht **Hilfe**

diese Variablen ...

**Einkommen**

**1:20**

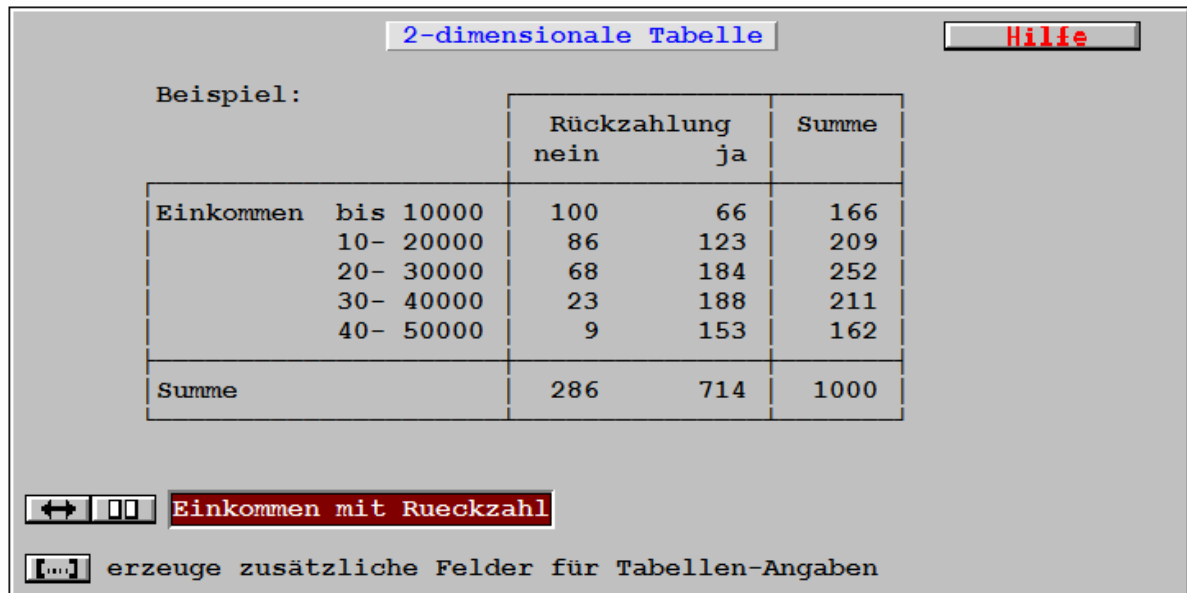
... aus diesen Datensätzen vor und nach der Umkodierung zur Kontrolle anzeigen

- 15  Option: Untersuchungseinheiten gewichten
- 16  Option: "Aussehen" der auszugehenden Tabelle bzw. Matrix
- 17  Grafik-Optionen
- 18 [Programmende](#)

## P45.13.1 Erläuterungen zu den Eingabe-Boxen

**Eingabe-Box 1** bis **Eingabe-Box 5**: Siehe "Arbeiten mit Almo-Datenanalyse-System", Abschnitt P0.1 bis P0.4.

**Eingabe-Box 6**: 2-dimensionale Tabelle



The screenshot shows a software interface titled "2-dimensionale Tabelle" with a "Hilfe" button. It displays a table example with the following data:

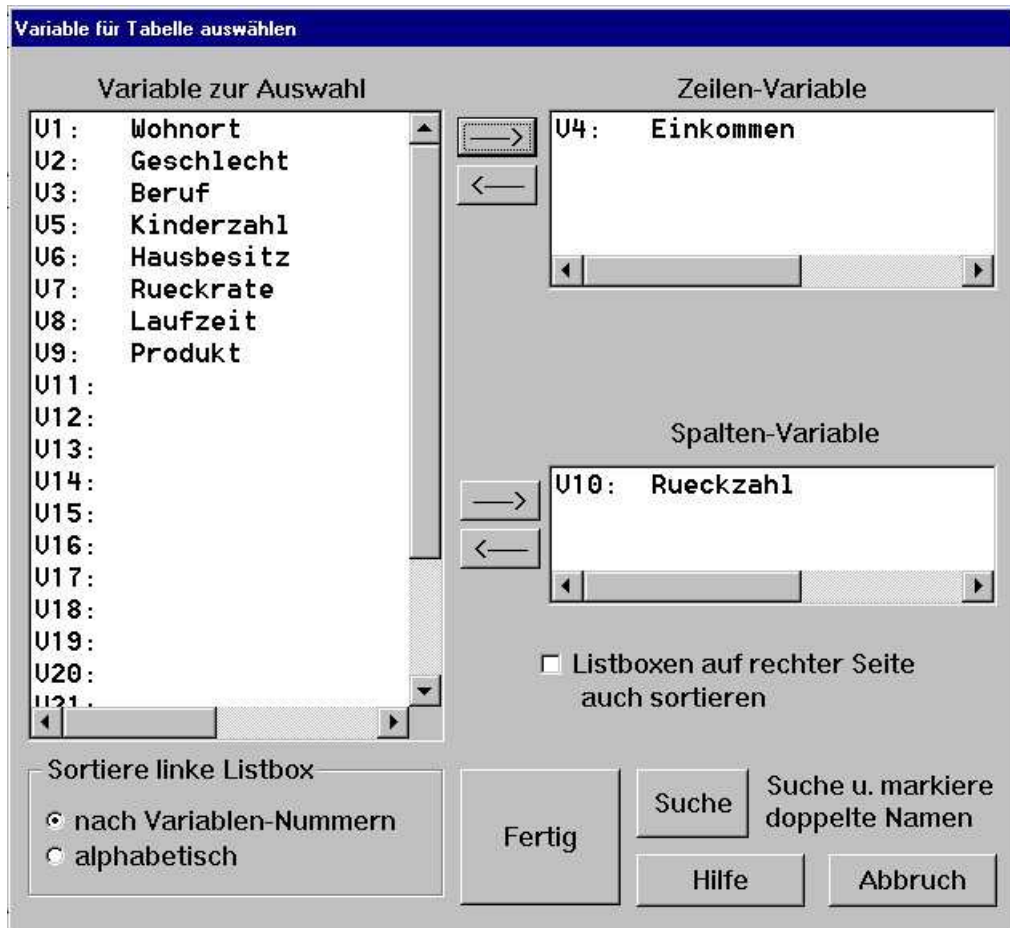
Beispiel:		Rückzahlung		Summe
		nein	ja	
Einkommen	bis 10000	100	66	166
	10- 20000	86	123	209
	20- 30000	68	184	252
	30- 40000	23	188	211
	40- 50000	9	153	162
Summe		286	714	1000

Below the table, there is a text input field containing "Einkommen mit Rueckzahl" and a button with a left-pointing arrow and two small squares. At the bottom, there is a button with three dots and the text "erzeuge zusätzliche Felder für Tabellen-Angaben".

In der Eingabe-Box ist ein Beispiel für eine 2-dimensionale Tabelle, wie sie Almo erzeugt, abgebildet. Die Zielvariable ist die „Rückzahlung“. Sie ist die „Spaltenvariable“ in der Tabelle. Die ursächliche Variable ist das Einkommen. Sie ist die „Zeilenvariable“ in der Tabelle.

Sie schreiben die beiden Variablen, die miteinander tabelliert werden sollen, in das Eingabefeld. Zwischen die beiden Variablennamen muß „mit“ geschrieben werden. Anstelle der Variablennamen können auch die Variablennummern geschrieben werden, z.B. „V5 mit V10“.

Die Eingabe kann bequemer mit Mausklick erfolgen. Wenn Sie auf den Knopf mit den 2 Fenstersymbolen klicken, dann wird die Eingabe-Box "Variablen für Tabelle auswählen" geöffnet. In ihr geben Sie an, welche Variable als Zeilenvariable und welche als Spaltenvariable die Tabelle bilden sollen.



Almo bildet bei dieser Eingabe die Tabelle V4 Einkommen mit V10 Rueckzahl.

*Wie man die Dialogbox "Variablen für Tabelle auswählen" bedient*

Klicken Sie auf eine Variable in der linken Listbox 'Variable zur Auswahl'. Dann klicken Sie auf den Pfeilknopf. Die Variable wird dann in die rechte Listbox "Zeilenvariable" oder "Spaltenvariable" transportiert. Der 'Transport' kann auch in der umgekehrten Richtung erfolgen.

Die Knöpfe am unteren Rand der Dialogbox haben folgende Bedeutung:

**SORTIERE** linke Listbox nach Variablennummern

Die Variablen in der linken Listbox werden nach aufsteigenden Nummern hintereinander gestellt.

**SORTIERE** linke Listbox alphabetisch

Die Variablen in der linken Listbox werden alphabetisch hintereinander gestellt. Variable, die keine Namen besitzen werden an das Ende gestellt.

**Knopf FERTIG**

Wenn Sie abschliessend auf den Knopf FERTIG klicken, dann werden die Variablennamen, die sich in den rechten Listboxen "Zeilenvariable" und "Spaltenvariable" befinden, in das Eingabefeld des Maskenprogramms eingesetzt. Wenn die hintereinander gestellten Variablennamen zu lang würden, dann verwendet Almo automatisch Variablennummern.

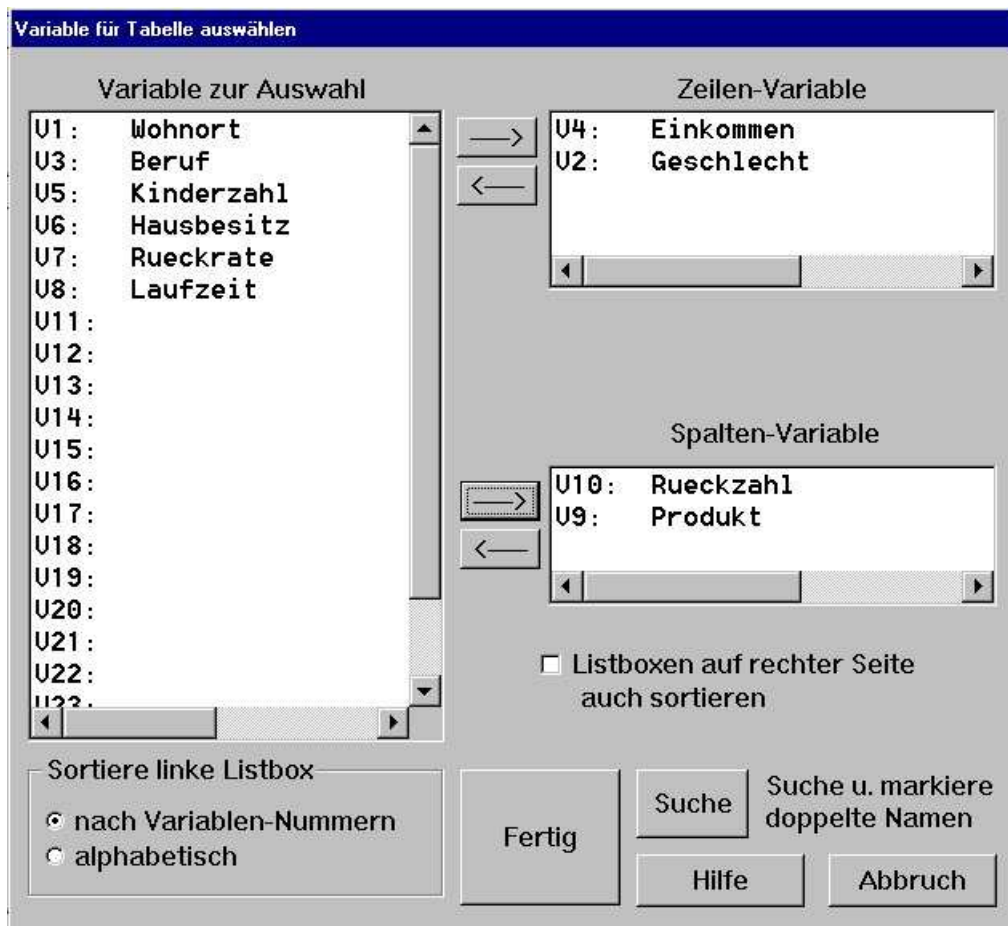
**Knopf SUCHE**

Variablenamen müssen eindeutig sein. Sie dürfen nicht doppelt vorhanden sein. Mit Klick auf den Knopf SUCHE prüft ALMO, ob Namen doppelt oder sogar mehrfach vorkommen. Diese Variablenamen werden dann durch 2 vorausgehende Unterstriche markiert, z.B. so:

V25: \_\_Geschlecht

Diese Variablenamen dürfen dann nicht für die Analyse ausgewählt werden.

Betrachten wir ein weiteres Beispiel für eine Tabellenangabe:



Nach Klick auf den Knopf FERTIG sehen Sie im Eingabefeld des Maskenprogramms folgende Eingabe:

Einkommen,Geschlecht mit Rueckzahl,Produkt

Alle Variable in der Eingabe-Box "Zeilenvariable" werden mit allen Variablen aus der Eingabe-Box "Spaltenvariable" gepaart. Für jedes Paar wird eine 2-dimensionale Tabelle gebildet. Also:

Einkommen mit Rueckzahl  
 Einkommen mit Produkt  
 Geschlecht mit Rueckzahl  
 Geschlecht mit Produkt

*Direkt in das Eingabefeld des Maskenprogramms hineinschreiben*

Sie können auch direkt in das Eingabefeld des Maskenprogramms hineinschreiben. Wenn Sie z.B. schreiben:

V1 mit V2

dann ist V1 die Zeilen- und V2 die Spalten-Variable. V1 steht vorne in der Tabelle und V2 oben rüber in der Tabelle

**BEACHTEN:**

Sie müssen dann auf den im Maskenprogramm weiter unten folgenden Knopf in der Eingabe-Box "Variable aus Tabellenangaben" klicken. Also registriert dann die Variablen, die zur Tabellenbildung verwendet werden.

Sie können in das Eingabefeld auch 2 oder mehrere Tabellenangaben schreiben, z.B. so

V1 mit V2 / V3 mit V4 / V5 mit V6

Zwischen die Tabellenangaben muss also ein Schrägstrich geschrieben werden. Zum Schluß kein Schrägstrich.

Vor und hinter "mit" können mehrere (durch Beistrich getrennte) Variable stehen. Folgende Angaben für 2-dimensionale Tabellen sind beispielsweise möglich:

V1 mit V3	tabelliert wird:	V1 mitV3
V1,2 mit V3	tabelliert wird:	V1 mit V3, V2 mit V3
V1 mit V3,4	tabelliert wird:	V1 mit V3, V1 mit V4
V1,2 mit V3,4	tabelliert wird:	V1 mit V3, V1 mit V4 V2 mit V3, V2 mit V4

## Eingabe-Box 7: 2-dimensionale Mehrfach-Tabelle

2-dimensionale "Mehrfach-Tabelle"Hilfe

Beispiel:

		Rueckzahlung		Summe
		nein	ja	
Wohnort	Stadt	126	215	341
	Land	160	499	659
Geschlecht	männlich	150	363	513
	weiblich	136	351	487
Beruf	Selbständig	52	93	145
	Unselbst.	234	621	855
Summe		286	714	1000

↔ ☐☐ Wohnort+Geschlecht+Beruf mit Rueckzahl

⋮ erzeuge zusätzliche Felder für Tabellen-Angaben

In der Eingabe-Box ist wieder ein Beispiel für eine derartige Tabelle abgebildet. Die Mehrfachtablette besteht aus beliebig vielen 2-dimensionalen Tabellen – mit der Besonderheit, daß die Spaltenvariable (in der Regel die Zielvariable) immer dieselbe ist und die Zeilenvariablen (die ursächlichen Variablen) jeweils andere sind.

Die Spaltenvariable kann im Prinzip beliebig viele Ausprägungen besitzen. Die Übersichtlichkeit geht jedoch verloren, wenn so viele Ausprägungen vorhanden sind, dass Almo die Tabelle umbrechen muss. Bei den Optionen, die das "Aussehen" der auszugebenden Tabelle steuern (siehe Eingabe-Box 16), besteht die Möglichkeit, die Breite der Ausgabe einzustellen. Wenn Sie hier beispielsweise 120 eingeben, dann wird die Tabelle mit einer Breite von 120 Zeichen dargestellt. Ist sie breiter, dann muss Almo umbrechen.

Wenn Sie auf den Knopf mit den 2 Fenstersymbolen klicken, dann wird die "Variablen-Auswahl-Box" geöffnet. In Ihr geben Sie an, welche Variable als Zeilenvariable und welche als Spaltenvariable die Tabelle bilden sollen.

Beispiel:

Wir haben die "Variablen-Auswahl-Box" bereits bei der vorausgehenden Eingabe-Box „2-dimensionale Tabelle“ abgebildet - so dass hier ein Ausschnitt genügt.



Almo bildet die Mehrfachtablette  $V1+V2+V3$  mit  $V8$ .

Die Zahl der Zeilenvariablen muss mindestens zwei sein. Die Spaltenvariable darf nur eine sein.

Sie können auch direkt in das Eingabefeld im Maskenprogramm hineinschreiben. Wenn Sie z.B. schreiben:

$V1+V2+V3$  mit  $V8$

dann sind  $V1, V2, V3$  die Zeilen- und  $V8$  die Spalten-Variable.

**BEACHTEN:**

Sie müssen dann auf den Knopf in der Eingabe-Box "Variable aus Tabellenangaben" klicken. Almo registriert dann die Variablen, die zur Tabellenbildung verwendet werden.

Sie können in das Eingabefeld auch 2 oder mehrere Tabellenangaben schreiben, z.B. so

$V1+V2+V3$  mit  $V8$  /  $V10+V11$  mit  $V15$  /  $V20+V21+V22$  mit  $V25$

Zwischen die Tabellenangaben muss also ein Schrägstrich geschrieben werden.

## Eingabe-Box 8: Partialtabelle, Interaktionstabelle

Partialtabellen  
Interaktionstabelle

Hilfe

Beispiel:

Geschlecht Produkt		Rückzahlung		Summe
		nein	ja	
männlich	Kleidung	55	45	100
	Möbel	29	71	100
	Technik	23	77	100
weiblich	Kleidung	35	65	100
	Möbel	32	68	100
	Technik	20	80	100
Summe		35	65	100

← → □ □ **Geschlecht\*Produkt mit Rueckzahl**

[...] erzeuge zusätzliche Felder für Tabellen-Angaben

Betrachten wir die Beispieltabelle, die in der Eingabe-Box abgebildet ist.

In dieser Tabelle wird untersucht, wie das "Produkt", das auf Kredit im Versandgeschäft gekauft wurde, die Bereitschaft zur "Rückzahlung" (des Kredits) bestimmt - dabei betrachten wir Männer und Frauen getrennt.

Die Tabelle besteht also aus 2 "Partialtabellen", aus einer Partialtabelle für die Männer und einer für die Frauen.

Die Variable des Geschlechts, für die die Partialtabellen gebildet werden, wird häufig als "Kontrollvariable" bezeichnet.

Wir bezeichnen die Variablen, die die Zeilen der Tabelle bilden auch als "interagierende" Zeilenvariable.

Es können beliebig viele "interagierende" Zeilenvariable eingesetzt werden. D.h. vor dem Alto-Schlüsselwort "mit" können beliebig viele Variable angegeben werden. Hinter "mit" darf nur eine stehen.

Wir erkennen, daß die Rückzahlungsbereitschaft bei Kleidung am schlechtesten und bei technischen Produkten am besten ist. Möbel liegen in der Mitte. Dabei ist bei weiblichen Kunden die Rückzahlungsbereitschaft insgesamt besser als bei Männern.

Auffallend ist, daß die Rückzahlungsbereitschaft bei Männern besonders schlecht ist, wenn das (auf Kredit) gekaufte Produkt Kleidung ist.

Wir können auch sagen, daß wir untersuchen, wie die "Interaktion" von Geschlecht und Produkt die Rückzahlungsbereitschaft bestimmt.

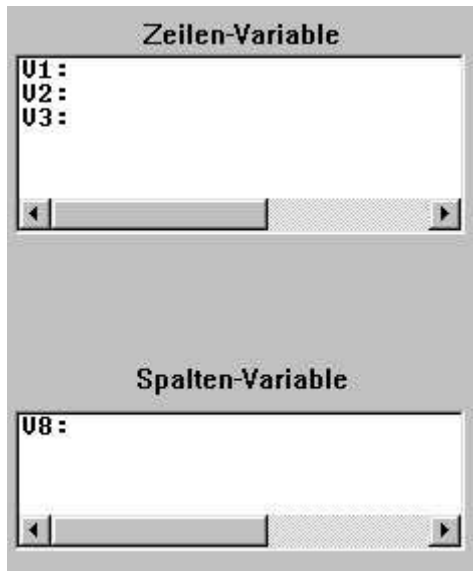
Deswegen bezeichnen wir diese Tabelle auch als Interaktionstabelle.

Die Spaltenvariable in der Partial- bzw. Interaktionstabelle kann im Prinzip beliebig viele Ausprägungen besitzen. Die Übersichtlichkeit geht jedoch verloren, wenn so viele Ausprägungen vorhanden sind, dass Almo die Tabelle umbrechen muss. Bei den Optionen, die das "Aussehen" der auszugebenden Tabelle steuern (siehe Eingabe-Box 16), besteht die Möglichkeit, die Breite der Ausgabe einzustellen. Wenn Sie hier beispielsweise 120 eingeben, dann wird die Tabelle mit einer Breite von 120 Zeichen dargestellt. Ist sie breiter, dann muss Almo umbrechen.

Wenn Sie auf den Knopf mit den 2 Fenstersymbolen klicken, dann wird die "Variablen-Auswahl-Box" geöffnet. In Ihr geben Sie an, welche Variable als (interagierende) Zeilenvariable und welche als Spaltenvariable die Tabelle bilden sollen.

Beispiel:

Wir haben die "Variablen-Auswahl-Box" bereits oben abgebildet - so dass hier ein Ausschnitt aus dieser Box genügt.



Almo bildet die Partial- bzw. Interaktionstabelle  $V1*V2*V3$  mit  $V8$ .

Die Zahl der Zeilenvariablen muss mindestens zwei sein. Die Spaltenvariable darf nur eine sein.

Sie können auch direkt in das Eingabefeld im Maskenprogramm hineinschreiben. Wenn Sie z.B. schreiben:

$V1*V2*V3$  mit  $V8$

dann sind  $V1, V2, V3$  die interagierenden Zeilenvariablen, die durch ein Multiplikationszeichen zu verbinden sind und  $V8$  die Spaltenvariable.

**BEACHTEN:**

Sie müssen dann auf den Knopf in der Eingabe-Box "Variable aus Tabellenangaben" klicken. Almo registriert dann die Variablen, die zur Tabellenbildung verwendet werden.

Sie können in das Eingabefeld auch 2 oder mehrere Tabellenangaben schreiben, z.B. so

V1\*V2\*V3 mit V8 / V10\*V11 mit V15 / V20\*V21\*V22 mit V25

Zwischen die Tabellenangaben muss also ein Schrägstrich geschrieben werden.

### Eingabe-Box 9: Multidimensionale Kontingenztabelle

Multidimensionale Kontingenztabelle Hilfe

Beispiel:

Wohnort	Beruf	Rueckzahl	Fälle
Stadt	Selbständ	nein	27
		ja	29
	Unselbst.	nein	99
		ja	186
Land	Selbständ	nein	25
		ja	64
	Unselbst.	nein	135
		ja	435
Summe			1000

Wohnort\*Beruf\*Rueckzahl

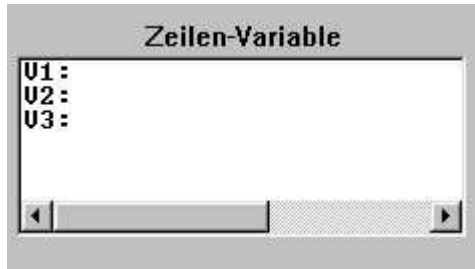
erzeuge zusätzliche Felder für Tabellen-Angaben

An der Beispieltabelle, die in der Eingabe-Box abgebildet wird, ist ersichtlich, was wir unter einer "multidimensionale Kontingenztabelle" verstehen: Zwei oder mehrere Variable werden kombiniert. Dann zählt Almo aus, wieviel Fälle es in den Daten für die jeweilige Kombination gefunden hat.

Wenn Sie auf den Knopf mit den 2 Fenstersymbolen klicken, dann wird die "Variablen-Auswahl-Box" geöffnet. In Ihr geben Sie an, welche Variable kombiniert werden sollen.

Beispiel:

Wir haben die "Variablen-Auswahl-Box" bereits oben abgebildet - so dass hier ein Ausschnitt aus dieser Box genügt. Einen Unterschied gibt es jedoch. Es ist nur eine Eingabebox für die Zeilenvariable vorhanden, keine für die Spaltenvariable - da es eben solche nicht gibt.



Almo bildet die multidimensionale Kontingenztabelle  $V1*V2*V3$ .

Die Zahl der Zeilenvariablen muss mindestens zwei sein. Spaltenvariable gibt es keine

Sie können auch direkt in das Eingabefeld im Maskenprogramm hineinschreiben. Wenn Sie z.B. schreiben:

$V1*V2*V3$

dann sind  $V1, V2, V3$  die zu kombinierenden Variablen.

BEACHTTE: Es gibt kein "mit".

Sie können in das Eingabefeld auch 2 oder mehrere Tabellenangaben schreiben, z.B. so

$V1*V2*V3 / V10*V11 / V20*V21*V22$

Zwischen die Tabellenangaben muss also ein Schrägstrich geschrieben werden.

## Box 10: Mittelwerts-Tabelle

Mittelwerts-Tabelle Hilfe

Beispiel:  
Mittelwerte der Kredit-Laufzeit  
je Geschlecht und Beruf

Geschl.	Beruf	Kredit-Laufzeit
männl	Selbstän	14.1481
	Unselbst	14.4954
weibl	Selbstän	14.4063
	Unselbst	15.2270
Gesamtmittel		14.7710

Zeilen-Variable

Spalten-Variable: quantitativ

Spalten-Variable: ordinal

Mit dieser Tabelle können Sie für die Ausprägungskombinationen der Zeilenvariablen Mittelwerte bzw. Mediane für die Spaltenvariablen berechnen. Zusätzlich werden auch noch Standardabweichungen bzw. mittlere Quartilsabstände ermittelt.

*Eingabefeld 1:* Geben Sie hier die Zeilenvariablen an

Wir haben hier Geschlecht und Beruf als Zeilenvariable eingegeben. Es werden also folgende Ausprägungskombinationen gebildet:

```
männlich - Selbständig
männlich - Unselbständig
.
.
```

Für diese Ausprägungskombinationen werden in unserem Beispiel die Mittelwerte der (im 2. Eingabefeld angegebenen) Spaltenvariablen der Laufzeit (des Kredits) ermittelt.

Auch die Standardabweichungen der Laufzeit je Ausprägungskombination von Geschlecht und Beruf werden errechnet.

Sie können nur eine oder auch beliebig viele Zeilenvariable angeben.

Die Zeilenvariablen dürfen beliebig viele Ausprägungen besitzen. Wenn sie sehr viele besitzen, dann entsteht eine sehr große, nicht mehr überschaubare Tabelle.

Die Zeilenvariablen sollten normalerweise ganzzahlig mit Schrittweite 1 kodiert sein. Sind sie das nicht, besitzen sie beispielweise Dezimalzahlen, dann werden sie von Almo automatisch auf Ganzzahligkeit umkodiert. Siehe dazu P45.12.0.

Wenn der Benutzer diese automatische Umkodierung vermeiden will, weil er anders als Almo umkodieren würde, dann muß er in der Box 14 "Kein-Wert-Angabe und Umkodierungen" selbst umkodieren.

*Eingabefeld 2: Spaltenvariable: quantitativ*

Wir haben hier "Laufzeit" (des Kredits) als Spaltenvariable angegeben. Diese Variable wird von Almo als quantitativ erachtet und demzufolge der Mittelwert je Ausprägungskombination errechnet.

Sie können auch mehrere quantitative Variable angeben, z.B.

Laufzeit, Einkommen

Almo rechnet dann für jede Spaltenvariable eine separate Analyse, d.h. es berechnet den Mittelwert (und die Standardabweichung) aus der Laufzeit und aus dem Einkommen je Ausprägungskombination der beiden Zeilenvariablen Geschlecht und Beruf.

*Eingabefeld 3: Spaltenvariable: ordinal*

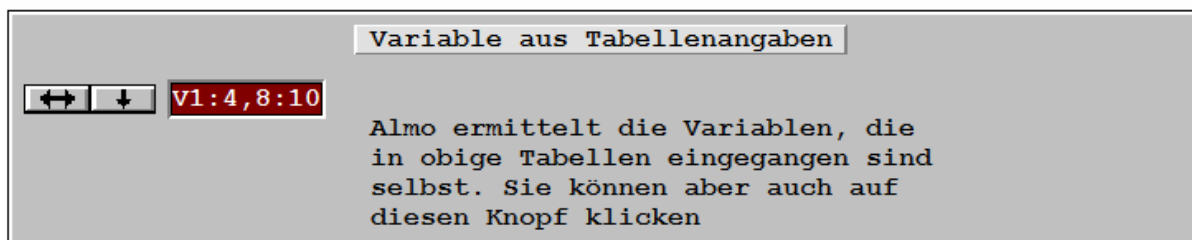
Anstelle der quantitativen Variablen oder zusätzlich zu ihnen können nun auch ordinale Variable eingesetzt werden. Almo errechnet für sie die Mediane und als Streuungsmaße die mittlere Quartilsdifferenzen je Ausprägungskombination der Zeilenvariablen.

Die Zahl der ordinalen Spaltenvariablen ist beliebig. Almo rechnet für jede eine separate Analyse.

Ordinale Variable werden normalerweise ganzzahlig mit Schrittweite 1 kodiert sein. Es ist es jedoch zulässig, daß sie auch Dezimalwerte besitzen. Almo kodiert diese Werte dann automatisch um, so dass sie ganzzahlig und mit Schrittweite 1 aufsteigend kodiert sind.

Wenn der Benutzer diese zwangsweise Umkodierung vermeiden will, weil er anders als Almo umkodieren würde, dann muß er in der Box 14 "Kein-Wert-Angabe und Umkodierungen" selbst umkodieren.

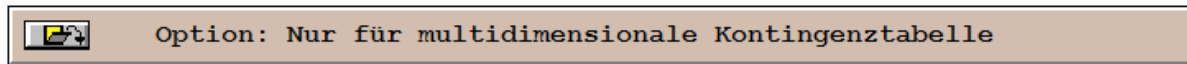
#### **Box 11:** Variable aus Tabellenangabe



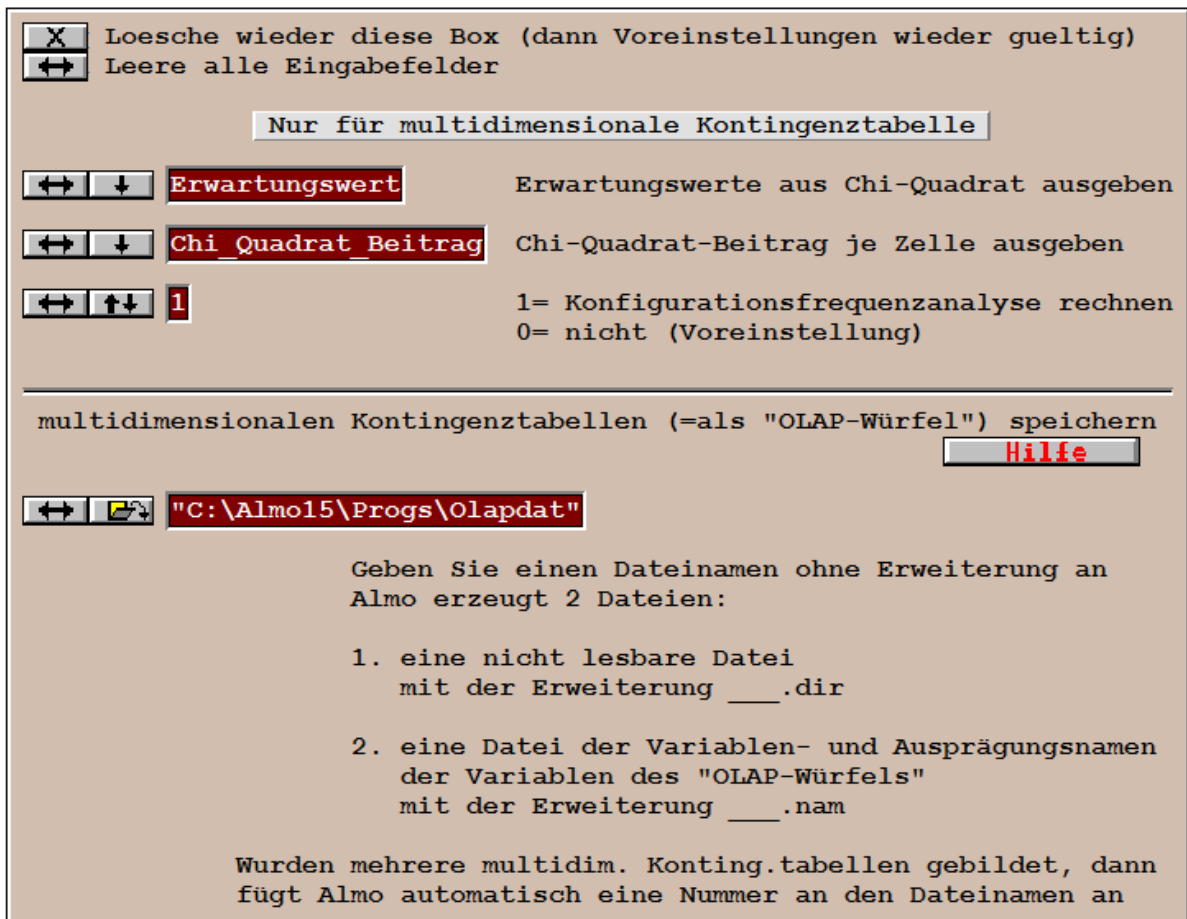
Almo ermittelt die Variablen, die in die Tabellen eingegangen sind automatisch. Wenn der Benutzer jedoch Eingaben mehrfach ändert, dann kann diese Automatik versagen.

Es ist deswegen sinnvoll zum Abschluss der Tabellenangaben auf den Knopf in dieser Box zu klicken. Almo ermittelt dann noch einmal abschliessend die Variablen, die in die Tabellen eingegangen sind.

**Box 12:** Option: Nur für multidimensionale Kontingenztabelle



Wird die Optionsbox geöffnet, dann sieht man die nachfolgend zweite Eingabe-Box.



In der Optionsbox kann man wahlweise bis zu 4 Optionen aktivieren. In unserem Beispiel werden alle aktiviert.

Es wird dann eine Konfigurationsfrequenzanalyse gerechnet und Erwartungswerte und Chi-Quadrat-Beiträge ermittelt.

Wir werden bei der Besprechung der Ergebnis-Ausgabe von Prog45mb im nächsten Abschnitt darauf zurückkommen.

#### 4. Eingabefeld

Wird eine multidimensionalen Kontingenztabelle als OLAP-Würfel in eine Datei gespeichert, so kann sie mit Prog11m3 wieder geladen werden. Dabei können dann aus ihr Sub-Tabellen jeglicher Art herausgelöst werden. Solche Sub-Tabellen können sein:

1. 2-dimensionale Tabelle
2. 2-dimensionale Mehrfach-Tabelle
3. Partial- bzw. Interaktions-Tabelle

#### 4. multidimensionale Kontingenztabelle

Wird im Eingabefeld ein Dateiname angegeben, dann erzeugt Almo 2 Dateien:

1. eine nicht lesbare Datei  
mit der Erweiterung `__.dir`  
in unserem Beispiel: `"C:\Almo\Progs\Olapdat.dir"`
2. eine Datei der Variablen- und Ausprägungsnamen  
der Variablen des "OLAP-Würfels"  
mit der Erweiterung `__.nam`  
in unserem Beispiel: `"C:\Almo\Progs\Olapdat.nam"`

Beachte: Damit Almo eine Datei der Variablen- und Ausprägungsnamen erstellen und speichern kann, sollten Sie zumindest den Variablen, die für die multidimensionalen Kontingenztabelle verwendet werden, Variablen- und möglichst auch Ausprägungsnamen geben.

Werden zu einem späteren Zeitpunkt mit Prog11m3 Sub-Tabellen gebildet, so müssen diese beiden Dateien in Prog11m3 in der Box

`Datei aus der der "OLAP-Würfel" eingelesen wird`

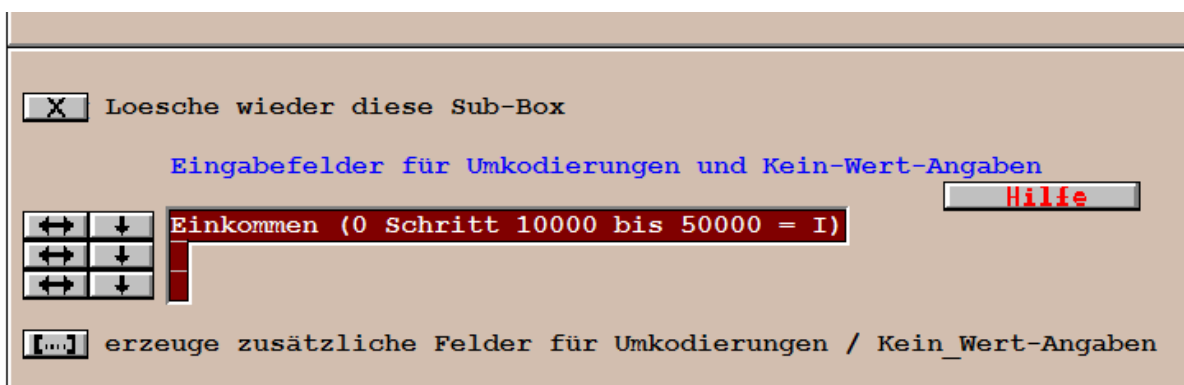
und in der Box

`Datei der Variablennamen`

eingesetzt werden.

**Eingabe-Box 13:** Ein- und Ausschliessen von Untersuchungseinheiten  
Siehe dazu "Arbeiten mit Almo-Datenanalyse-System", Abschnitt P0.7.

**Eingabe-Box 14:** Kein-Wert-Angabe und Umkodierung



Einkommen ist eine quantitative Variable, die über die 1 000 Personen hinweg unzählige Ausprägungen besitzt. Um sie als Variable für eine Tabelle verwendbar zu machen, muß sie so umkodiert werden, daß sie nur einige wenige Ausprägungen besitzt. Obige Anweisung bewirkt, daß sie von ihrer Werteuntergrenze (= 0) bis zu ihrer Werteobergrenze (= 50 000) mit einer Schrittweite von 10 000 zusammengefaßt wird.

Aus	0	bis	10 000	entsteht	1
	10 001		20 000		2
	20 001		30 000		3
	30 001		40 000		4
	40 001		50 000		5

Der Buchstabe „I“ in der Umkodierungsanweisung bedeutet „Intervall-Kodierung“.

Welche Möglichkeiten der Umkodierung und der Kein-Wert-Angabe bestehen ist ausführlich im "Arbeiten mit Almo-Datenanalyse-System", Abschnitt P0.5 dargestellt.

**Eingabe-Box 15:** Untersuchungseinheiten gewichten

Siehe dazu "Arbeiten mit Almo-Datenanalyse-System", Abschnitt P0.8.

**Eingabe-Box 16:** Optionen, die das Aussehen der auszugebenden Tabelle steuern

Siehe dazu "Arbeiten mit Almo-Datenanalyse-System", Abschnitt P0.9.

## P45.13.2 Ausgabe

Almo liefert folgende Ausgabe (gekürzt):

```
Zahl der eingelesenen Datensätze:      1000
#####
```

### Zwei-dimensionale Tabellen

Tabelle 1: V4 Einkommen  
mit  
V10 Rueckzahl

Bezeichnung der Variablen auf Stirnseite der Tabelle (Zeilenvariable)

```
V4    Einkommen
      1 bis 10
      2 10-20
      3 20 bis 30
      4 30 bis 40
      5 40-50
```

Bezeichnung der Variablen im Kopf der Tabelle (Spaltenvariable)

```
V10   Rueckzahl
      1 nein
      2 ja
```

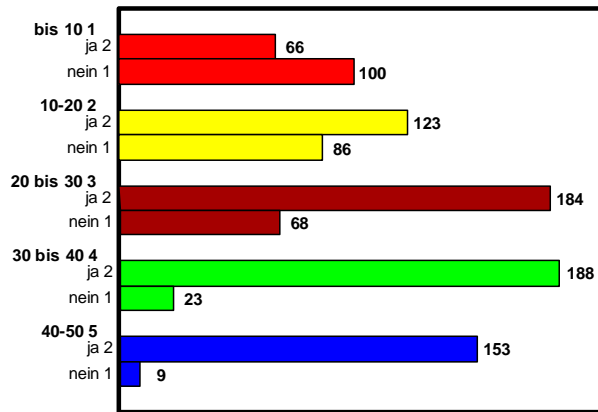
	Rueckzahl		Summe
	nein 1	ja 2	
Einkom bis 10	100	66	166
10-20	86	123	209
20 bis 3	68	184	252
30 bis 4	23	188	211
40-50	9	153	162
Summe	286	714	1000

#### \*\*\*\*\* Erläuterung:

Betrachten wir die hintere Randsumme. Von den insgesamt 1000 Personen haben 166 ein Einkommen bis 10000. Von ihnen haben 100 den Kredit nicht zurückgezahlt; 66 haben ihn zurückbezahlt, .... etc

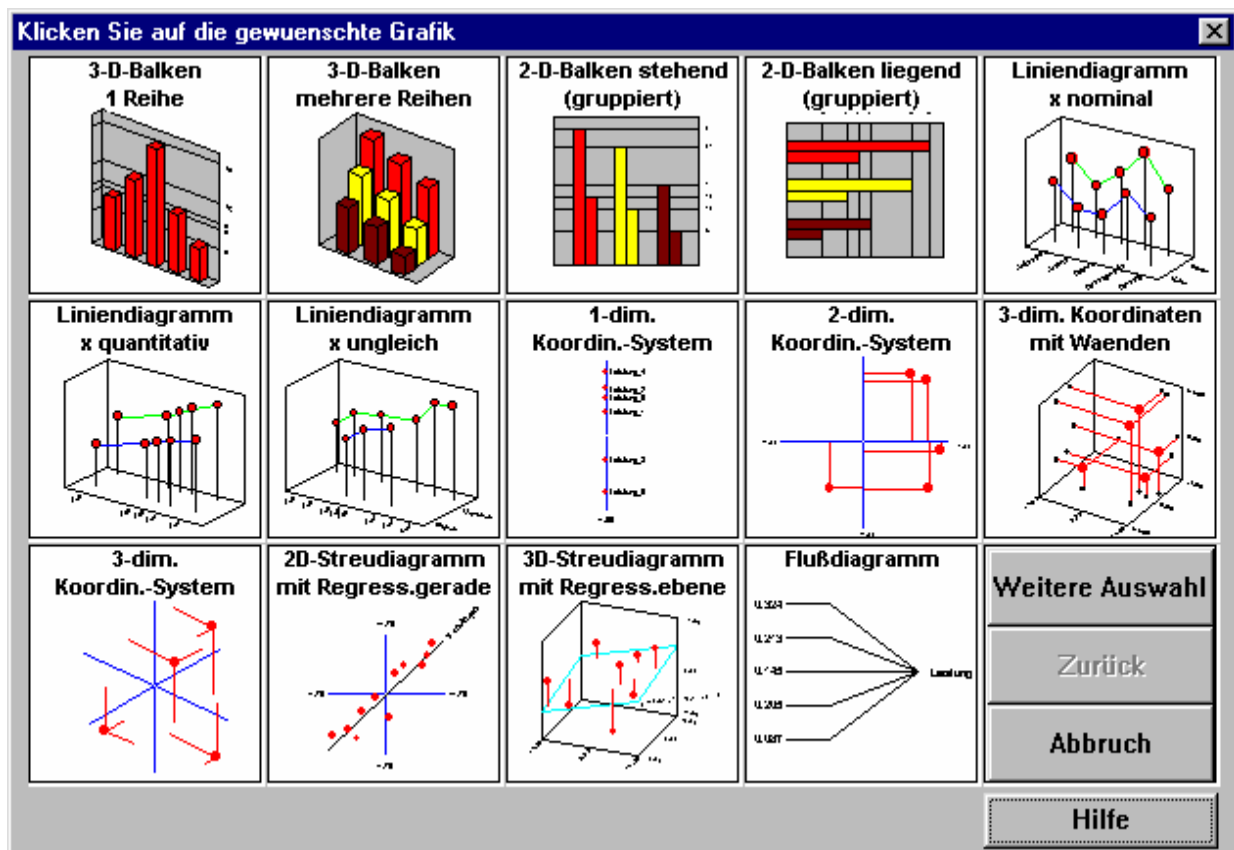
Almo liefert für diese Tabelle folgende Grafik

2-dimensionale Verteilung  
Einkommen  
und  
Rueckzahl



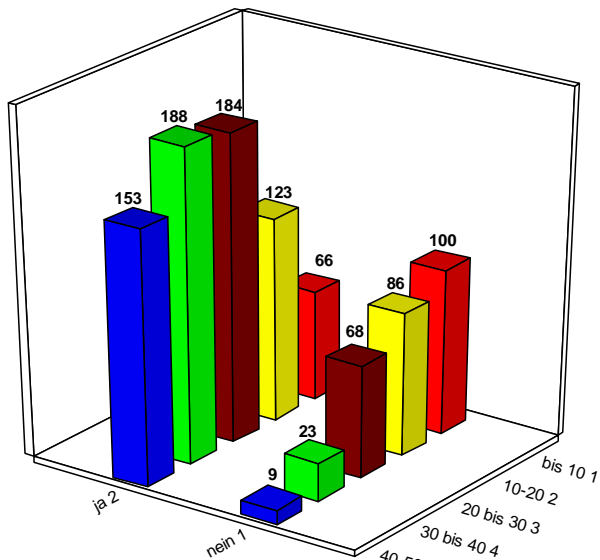
Im Grafik-Editor kann diese Grafik in vielfältiger Weise verändert werden. Siehe dazu Teil 1, Bedienungsanleitung, Abschnitt 10.2.

Wir wollen hier nur zeigen, wie obige Grafik in ein 3D-Balkendiagramm gewandelt werden kann. Nach Klick auf "Anderer Grafiktyp" auf der linken Seite des Grafikfensters zeigt Almo die folgende Auswahl:



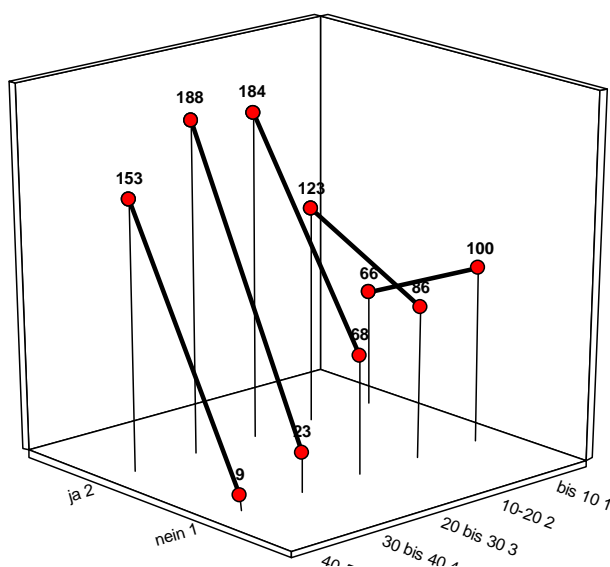
Wir klicken auf "3-D -Balken, mehrere Reihen" (das 2. Bild in der 1. Reihe) Also erzeugt dann ein Balkendiagramm, das wir noch durch einige Mausklicks auf "Transponiere", "Perspektive", "Balkendicke" und durch Verschieben der Masszahlen verschönern.

2-dimensionale Verteilung  
Einkommen  
und  
Rueckzahl



Sinnvoll wäre es auch ein Liniendiagramm zu erzeugen – vor allem dann, wenn die beiden Variablen mehrere Ausprägungen besitzen. Wir klicken zu diesem Zweck auf „Liniendiagramm, x nominal“. Das ist das letzte Bild in der ersten Reihe.

2-dimensionale Verteilung  
Einkommen  
und  
Rueckzahl



zeilenweise prozentuiert

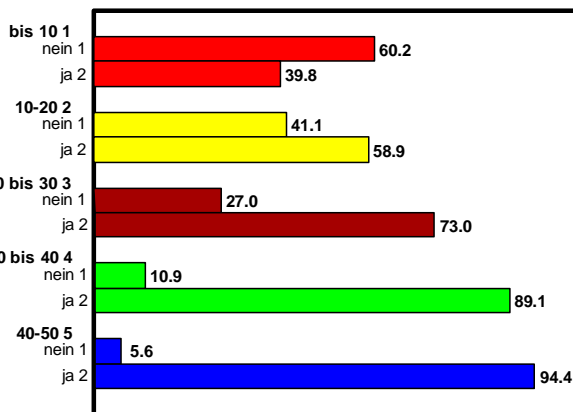
	Rueckzahl		Summe
	nein 1	ja 2	
Einkom bis 10	60.24	39.76	100.00
10-20	41.15	58.85	100.00
20 bis 3	26.98	73.02	100.00
30 bis 4	10.90	89.10	100.00
40-50	5.56	94.44	100.00
Summe	28.60	71.40	100.00

\*\*\*\*\* Erläuterung:

Das zeilenweise Prozentuieren ist die sinnvollste Art eine Tabelle zu prozentuieren. Die Prozentwerte in den Zellen der Tabelle sind auf die Zeilensumme bezogen. Die 5 Einkommensgruppen sind jetzt gleich groß, eben gleich 100, so dass wir sie miteinander vergleichen können. Wir sehen z.B., dass von den Personen der 1. Einkommensgruppe 60.24 % ihren Kredit nicht zurückzahlen; dies tun nur 39.76 %. Andererseits zahlen von den Personen in der höchsten Einkommensgruppe 94.44 % Ihren Kredit zurück und nur 5.56 % tun das nicht.

Almo liefert für diese Tabelle folgende Grafik

2-dimensionale Verteilung  
Einkommen  
und  
Rueckzahl



Koeffizienten

Chi-Quadrat = 172.3220

df = 4

Signifikanz (1-p)\*100 = 100.000

Ein Wert ueber ca. 95 bedeutet:  
Zwischen den Variablen besteht  
ein signifikanter Zusammenhang

Kontingenzkoeffizient C(cor) = 0.542

Cramers V = 0.415

\*\*\*\*\* Erläuterung:

Die Stärke des Zusammenhangs zwischen Einkommen und Kredit-Rückzahlung wird durch 2 Korrelationskoeffizienten ausgedrückt, den korrigierten

Kontingenzkoeffizient C(cor) und das Cramer'sche V. Die Werte stimmen selten überein. Wir würden Cramers V präferieren, da es sich als PRE-Koeffizient im Rahmen des Allgemeinen Linearen Modells interpretieren lässt. Siehe dazu P45.12.3, Erläuterungen zu den Korrelationskoeffizienten, sowie Handbuch zu P20, Abschnitt P20.9.5.1 und Bortz, Lienert, Boehnke, 1990, S. 357.

C(cor) und Cramers V sind Korrelationskoeffizienten für nominale Variable. Sind die Variablen quantitativ oder ordinal oder gemischt, dann wird die Korrelation durch diese beiden Koeffizienten in der Regel unterschätzt.

Mit dem Korrelationsprogramm Prog45m6 in Abschnitt P45.12 oder mit den Maskenprogrammen Prog10m1 oder Prog10m2 in Almo können die adäquaten Korrelationskoeffizienten ermittelt werden. Siehe auch Handbuch zu Teil 3: Grundlegende Verfahren, Abschnitt P10.4.2 ff.

#####  
####

### Mehrfach-Tabellen

Tabelle 2: V1 Wohnort + V2 Geschlecht + V3 Beruf  
mit  
V10 Rueckzahl

Bezeichnung der Variablen auf Stirnseite der Tabelle (Zeilenvariable)

V1 Wohnort  
1 Stadt  
2 Land  
V2 Geschlecht  
1 m  
2 w  
V3 Beruf  
1 Selbst  
2 Unselbst

Bezeichnung der Variablen im Kopf der Tabelle (Spaltenvariable)

V10 Rueckzahl  
1 nein  
2 ja

	Rueckzahl		Summe
	nein 1	ja 2	
Wohnort Stadt	126	215	341
Land	160	499	659
Summe	286	714	1000

Geschl m	150	363	513
w	136	351	487
Summe	286	714	1000

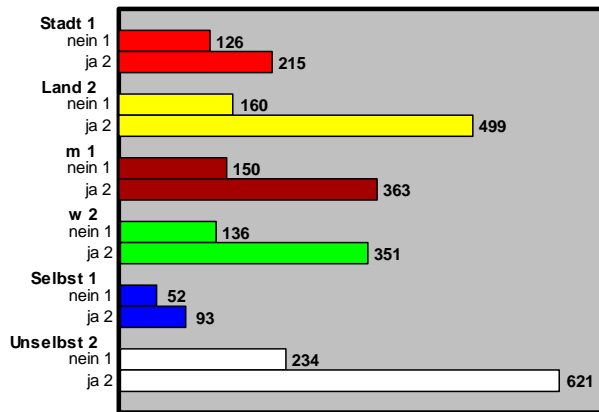
Beruf Selbst	52	93	145
Unselbst	234	621	855
Summe	286	714	1000

\*\*\*\*\* **Erläuterung:**

Es entstehen 3 zwei-dimensionale Teiltabellen, die zu einer einzigen Tabelle zusammengefasst sind. Dadurch entsteht eine sehr kompakte und übersichtliche Ausgabe.

Almo erzeugt für diese Tabelle folgendes Balkendiagramm:

2-dimensionale Verteilung



zeilenweise prozentuiert

		Rueckzahl		Summe
		nein 1	ja 2	
Wohnor	Stadt	36.95	63.05	100.00
	Land	24.28	75.72	100.00
Summe		28.60	71.40	100.00
Geschl	m	29.24	70.76	100.00
	w	27.93	72.07	100.00
Summe		28.60	71.40	100.00
Beruf	Selbst	35.86	64.14	100.00
	Unselbst	27.37	72.63	100.00
Summe		28.60	71.40	100.00

Auch für diese Tabelle erzeugt Almo ein Balkendiagramm, das wir hier nicht abbilden. Im folgenden gibt Almo für die 3 Teiltabellen die Signifikanz und die Korrelation aus. Siehe dazu oben bei zwei-dimensionaler Tabelle.



Bezeichnung der Variablen im Kopf der Tabelle (Spaltenvariable)

V10 Rueckzahl  
 1 nein  
 2 ja

Geschl Produk		Rueckzahl		Summe
		nein 1	ja 2	
m	Kleidu	41	50	91
	Möbel	60	149	209
	Techni	49	164	213
w	Kleidu	40	73	113
	Möbel	57	121	178
	Techni	39	157	196
Summe		286	714	1000

Partialtabelle für Männer

Partialtabelle für Frauen

\*\*\*\*\* **Erläuterung:**

Hier wird in 2 Tabellen "Produkt" gegen "Rückzahlung" tabelliert. Die 1. Tabelle gilt nur für die Männer, die 2. nur für die Frauen.

zeilenweise prozentuiert

Geschl Produk		Rueckzahl		Summe
		nein 1	ja 2	
m	Kleidu	45.05	54.95	100.00
	Möbel	28.71	71.29	100.00
	Techni	23.00	77.00	100.00
w	Kleidu	35.40	64.60	100.00
	Möbel	32.02	67.98	100.00
	Techni	19.90	80.10	100.00
Summe		28.60	71.40	100.00

Partialtabelle für Männer

Partialtabelle für Frauen

Koeffizienten fuer Partialtabelle 1

V9 Produkt  
 mit  
 V10 Rueckzahl  
 fuer  
 V2 Geschlecht: 1 m

Chi-Quadrat = 15.0316

df = 2

Signifikanz (1-p)\*100 = 99.913

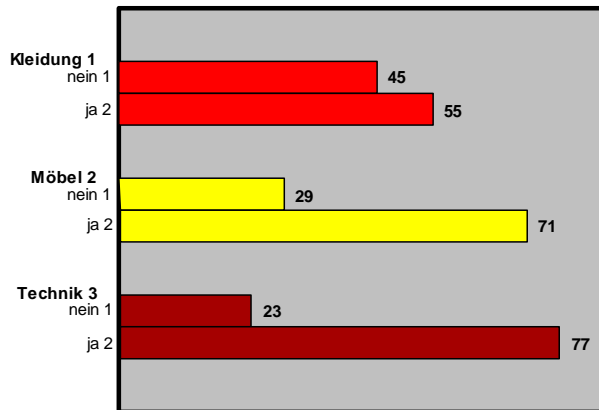
Ein Wert ueber ca. 95 bedeutet:  
 Zwischen den Variablen besteht  
 ein signifikanter Zusammenhang

Kontingenzkoeffizient C(cor) = 0.239

Cramers V = 0.171

Almo liefert für die Partialtabelle 1 zwei Grafiken, eine für die absoluten Häufigkeiten und eine für die zeilenweise prozentuierten Häufigkeiten. Wir zeigen hier nur die letztere.

2-dimensionale Verteilung  
 Produkt  
 mit  
 Rueckzahl  
 für  
 ....Geschlecht : m



```

=====
Koeffizienten fuer Partialtabelle  2      V9 Produkt
                                         mit
                                         V10 Rueckzahl
                                         fuer
                                         V2 Geschlecht:  2 w

Chi-Quadrat = 10.8948      df = 2      Signifikanz (1-p)*100 = 99.531

Ein Wert ueber ca. 95 bedeutet:
Zwischen den Variablen besteht
ein signifikanter Zusammenhang

Kontingenzkoeffizient C(cor)   = 0.209
Cramers V                      = 0.150
  
```

Almo liefert auch hier zwei Grafiken, eine für die absoluten Häufigkeiten und eine für die zeilenweise prozentuierten Häufigkeiten, die wir jedoch nicht abbilden.

```

=====
Koeffizienten fuer gesamte Interaktionstabelle

Beachte: Die Merkmalskombinationen der unabhängigen Variablen werden wie die
Ausprägungen einer unabhängigen Variablen betrachtet.

Chi-Quadrat = 26.1798      df = 5      Signifikanz (1-p)*100 = 99.979

Ein Wert ueber ca. 95 bedeutet:
Zwischen den Variablen besteht
ein signifikanter Zusammenhang

Kontingenzkoeffizient C(cor)   = 0.226
Cramers V                      = 0.162
  
```

**\*\*\*\*\* Erläuterung:**

Almo gibt hier einen pauschalen Signifikanz- und Korrelationskoeffizienten aus, bei dem die Merkmalskombinationen der ursächlichen Variablen betrachtet werden wie

die Ausprägungen einer einzigen ursächlichen Variablen. Diese Koeffizienten sind in der Regel nicht zu interpretieren.

### Multidimensionale Kontingenztabelle

Tabelle 4: V1 Wohnort \* V3 Beruf \* V10 Rueckzahl

Bezeichnung der Variablen auf Stirnseite der Tabelle (Zeilenvariable)

V1 Wohnort  
 1 Stadt  
 2 Land  
 V3 Beruf  
 1 Selbst  
 2 Unselbst  
 V10 Rueckzahl  
 1 nein  
 2 ja

			Fälle
Wohnort	Beruf	Rueckz	
Stadt	Selbst	nein	27
		ja	29
	Unselb	nein	99
		ja	186
Land	Selbst	nein	25
		ja	64
	Unselb	nein	135
		ja	435
Summe			1000

Matrix der Erwartungswerte

			Werte
Wohnort	Beruf	Rueckz	
Stadt	Selbst	nein	14.1413
		ja	35.3037
	Unselb	nein	83.3847
		ja	208.1703
Land	Selbst	nein	27.3287
		ja	68.2263
	Unselb	nein	161.1453
		ja	402.2997
Summe			1000.0000

\*\*\*\*\* **Erläuterung:**

Erwartungswerte sind diejenigen Häufigkeiten, die sich ergeben würden, wenn die Verteilung der 1000 Personen auf die Zellen zufällig erfolgen würde (bei gegebener Verteilung auf die 3 Variablen).

In der 1. Zelle der Tabelle wird ein Erwartungswert von 14.1413 (ganzzahlig auf 14 gerundet) angegeben. Die tatsächliche Häufigkeit ist jedoch 27, also deutlich überzufällig.

Matrix der Chi-Quadrat-Beitraege

Wohnor Beruf Rueckz			Werte
Stadt	Selbst	nein	11.6925
		ja	1.1256
	Unselb	nein	2.9242
		ja	2.3611
Land	Selbst	nein	0.1984
		ja	0.2618
	Unselb	nein	4.2420
		ja	2.6580
Summe			25.4637

\*\*\*\*\* **Erläuterung:**

Der gesamte Chi-Quadrat-Wert ist 25.4637. Mit 11.6925 trägt die 1. Zelle am meisten zu diesem Wert bei. Die Frage, die sich nun stellt, lautet: Welche der 8 Chi-Quadrat-Werte in den 8 Zellen sind signifikant, d.h. sind überzufällig weit von 0 entfernt. Diese Frage wird mit der nachfolgend behandelten Konfigurationsfrequenzanalyse (KFA) beantwortet.

Chi-Quadrat-Test auf allseitige Abhaengigkeit

-----  
 Chi-Quadrat = 25.4637            df = 4            Signifikanz (1-p)\*100 = 99.986

2I-Test auf allseitige Abhaengigkeit

-----  
 2I-Wert        = 23.1651            df = 4            Signifikanz (1-p)\*100 = 99.972

\*\*\*\*\* **Erläuterung:**

Es werden 2 Signifikanzkoeffizienten ausgegeben. In unserem Beispiel ist die aus 3 Variablen gebildete multidimensionale Kontingenztabelle hoch signifikant.

#### Konfigurationsfrequenzanalyse (KFA)

Schranke fuer	bei Signifikanz
Chi-Quadrat-Beitrag	(1-p)*100
4.324543	85.0
5.027256	90.0
<b>6.240185</b>	<b>95.0</b>
7.502315	97.5
9.266205	99.0
13.447144	99.9

#### \*\*\*\*\* Erläuterung:

Die Konfigurationsfrequenzanalyse (KFA) kann verwendet werden, um

1. zu überprüfen, ob die in der obigen Tabelle der Chi-Quadrat-Beiträge angegebenen einzelnen Werte signifikant sind
2. und ob ein "Typ" identifizierbar ist.

Ein Chi-Quadrat-Beitrag ist signifikant, wenn er den Schrankenwert (je gewählte Signifikanz) überschreitet. In unserem Beispiel ist ein Chi-Quadrat-Beitrag mit 95 % signifikant, wenn er gleich / größer 6.240185 ist. Dies gilt für die 1. Zelle in obiger Tabelle. Der betreffende Chi-Quadrat-Beitrag ist mit 11.6925 sogar mit über 99 % signifikant.

Ein "Typ" ist existent, wenn

- a. der Chi-Quadrat-Beitrag signifikant ist
- b. und wenn die tatsächliche Häufigkeit grösser ist als der Erwartungswert.

In unserem Beispiel trifft dies auf die 1. Zelle zu, d.h. Personen mit der Merkmalskombination

Stadtbewohner, Selbständiger, Nicht-Rückzahler

bilden einen signifikanten "Typ". In unserem Beispiel ist allerdings nicht viel damit gewonnen, dass wir diese Merkmalskombination als "Typ" bezeichnen dürfen.

Ist, bei signifikantem Chi-Quadrat-Beitrag, die tatsächliche Häufigkeit kleiner als der Erwartungswert, dann kann von einem "Antityp" gesprochen werden. Dieser ist inhaltlich nicht immer interpretierbar.

Siehe dazu: Krauth: Einführung in die Konfigurationsfrequenzanalyse (KFA), Beltz-Verlag, Weinheim, 1993, S. 23 ff.

Matrix der p-Werte aus KFA-Binomialtest

Wohnor Beruf Rueckz			Werte
Stadt	Selbst	nein	0.001384
		ja	0.159810
	Unselb	nein	0.044549
		ja	0.044092
Land	Selbst	nein	0.371617
		ja	0.325236
	Unselb	nein	0.012305
		ja	0.019067
Summe			-

Schranke fuer p-Wert in Tabelle	bei Signifikanz (1-p)*100
0.018750	85.0
0.012500	90.0
<b>0.006250</b>	<b>95.0</b>
0.003125	97.5
0.001250	99.0
0.000125	99.9

**\*\*\*\*\* Erläuterung:**

Hier wird der p-Wert aus dem KFA-Binomialtest dazu verwendet, um zu überprüfen, ob eine Zelle der Tabelle signifikant hervorsteht bzw. einen "Typ" bildet.

Ein p-Wert ist signifikant, wenn er den Schrankenwert (je gewählte Signifikanz) unterschreitet. In unserem Beispiel ist ein p-Wert mit 95 % signifikant, wenn er gleich / kleiner 0.006250 ist. Dies gilt für die 1. Zelle in obiger Tabelle. Der betreffende p-Wert ist 0.001384 und ist somit sogar mit über 97.5 % signifikant - aber nicht mehr, wie oben mit über 99 %.

Der KFA-Binomialtest gilt als der bessere Test, d.h. für die Identifizierung eines Typs, bzw. für das signifikante "Hervorstechen" einer Merkmalskombination ist der KFA-Binomialtest vorzuziehen.

## Mittelwerts-Tabelle

V2      Geschlecht: m w  
V3      Beruf: Selbst Unselbst  
V8      Laufzeit

V2 wird auch bezeichnet mit A  
die Ausprägungen (bzw.Dummies) mit  
A1 =m  
A2 =w

V3 wird auch bezeichnet mit B  
die Ausprägungen (bzw.Dummies) mit  
B1 =Selbst  
B2 =Unselbst

Zellenmittelwerte der  
quantitativen/ordinalen Variablen  
(arithmetisches Mittel bei quantitativen Variablen)

Geschlec Beruf		Laufzeit
m	Selbst	14.1481
	Unselbst	14.4954
w	Selbst	14.4063
	Unselbst	15.2270
Gesamtmittel		14.7710

### \*\*\*\*\* Erläuterung:

Mit dieser Tabelle werden für die Ausprägungskombinationen der Zeilenvariablen Mittelwerte bzw. Mediane für die Spaltenvariablen berechnen. Zusätzlich werden (in der nachfolgenden Tabelle) auch noch Standardabweichungen bzw. mittlere Quartilsabstände ermittelt.

Die Zeilenvariable müssen nominal sein. Sie dürfen polytom sein. Die Spaltenvariable muß quantitativ oder ordinal sein. Die Zahl der Zeilenvariablen und auch der Spaltenvariablen ist beliebig

In unserem Beispiel haben wir Geschlecht und Beruf als Zeilenvariable eingegeben. Es werden alle Ausprägungskombinationen gebildet:

männlich - Selbständig  
männlich - Unselbständig  
weiblich - Selbständig  
weiblich - Unselbständig

Für diese Ausprägungskombinationen werden in unserem Beispiel die Mittelwerte der quantitativen Variablen "Laufzeit (des Kredits)" ermittelt.

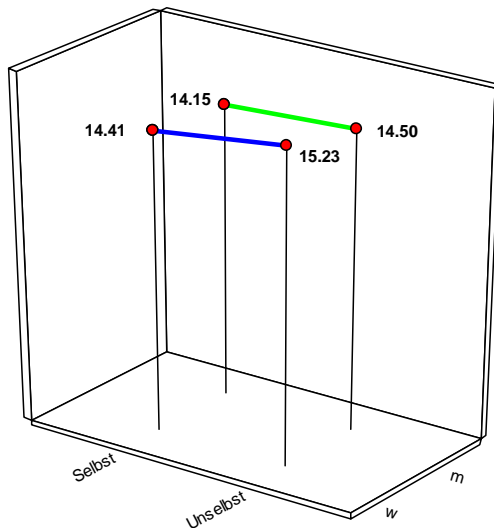
In unserem Beispiel ist zu erkennen, daß der Gesamt-Mittelwert der Kredit-Laufzeit 14.7710 Monate beträgt. Die männlichen Selbständigen weichen mit 14.1481 Monaten am stärksten nach unten ab und die weiblichen Unselbständigen mit 15.2270 Monaten am stärksten nach oben ab.

Anstelle der quantitativen Variablen oder zusätzlich zu ihnen können nun auch ordinale Variable als Spaltenvariable eingesetzt werden. Also errechnet für sie die Mediane und als Streuungsmaße die mittleren Quartilsdifferenzen je Ausprägungskombination der Zeilenvariablen.

**Hinweis:** Für Mittelwertstabellen mit nur einer unabhängigen nominalen Variablen und einer quantitativen abhängigen Variablen kann ein t-Test gerechnet werden. Ist die abhängige Variable ordinal, dann kann der Median-Test oder der U-Test oder der H-Test gerechnet werden. Der Benutzer muß dazu das Progl8 verwenden. Siehe auch Handbuch, Teil 3, Grundlegende Verfahren.

Also bildet aus obiger Tabelle folgendes Liniendiagramm, das nicht besonders spektakulär ist, da beide nominale Variable nur 2 Ausprägungen besitzen.

Mittelwerte von Laufzeit



Streuung der quantitativen/ordinalen Variablen je Zelle  
Standardabweichung bei quantitativen Variablen (Varianz bzw. Standardabweichung ist mit n, nicht mit n-1 dividiert)

Geschlec Beruf		Laufzeit
m	Selbst	5.5112
	Unselbst	5.4105
w	Selbst	5.2432
	Unselbst	5.6006
		-

=====

die Zellenmittelwerte und Streuungen der abhaengigen Variablen beruhen auf folgenden Besetzungszahlen (Zellenhaeufigkeiten)

Geschlec Beruf		Laufzeit
m	Selbst	81
	Unselbst	432
w	Selbst	64
	Unselbst	423
		-

=====

### P45.13.3 Weiterführende Hinweise

In Almo sind noch die Programme Prog10m1, Prog10m2 und Prog10m3 enthalten, die 2- und 3-dimensionale Tabellen erstellen und dabei eine Fülle von Koeffizienten berechnen. Will der Benutzer noch, neben Cramers V und dem Kontingenzkoeffizienten, die das Data-Mining-Programm Prog45mb liefert, noch weitere Korrelationskoeffizienten ausgegeben haben, dann sollte er Prog10m1 oder Prog10m2 rechnen. Bei Prog10m3 können fertige Tabellen eingegeben werden. Das Programm errechnet dann eine Fülle von Koeffizienten. Im Almo-Handbuch Teil 3 „Grundlegende Verfahren“ werden diese Programme und die Koeffizienten, die sie liefern, sehr ausführlich dargestellt.

### ***P45.14 Schritt 9b: Streudiagramm für 2 oder 3 Variable***

Betrachten wir in unserem Beispiel den Zusammenhang zwischen dem Einkommen als ursächlicher Variablen und der Rückzahlungsrate als Zielvariable. Dieser Zusammenhang soll grafisch als Streudiagramm dargestellt werden. Almo bietet dafür das Programm Prog45m7 an.

**Prog45m7.Msk**  
Streudiagramm und Regressionslinie bzw -ebene  
für eine Analyse mit 1 oder 2 unabhängigen Variablen

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

1    
Vereinbare Variable=  ;

2  Option: Weitere Vereinbarungen - nur wenn Almo dazu auffordert

3    
  "C:\Almo7\TESTDAT\DatMin.nam"  
        zeige = Namensdatei in Output zeigen  
leer = nicht

4    
    
 erzeuge zusätzliche Namensfelder

5    
 "C:\Almo7\TESTDAT\DatMin.dir"

6   
unabhängige quantitative Variable  
maximal 2 möglich  
    
-----  
abhängige quantitative Variable  
nur 1 möglich

7

↓ Loesche wieder diese Box

Ein- und Ausschliessen von Untersuchungseinheiten Hilfe  
Hilfe

---

Untersuchungseinheiten in Analyse nur EINSchließen wenn ..... Hilfe

↔

---

Untersuchungseinheiten aus Analyse AUSSchließen wenn ..... Hilfe

↔

---

nur diese Datensätzen verwenden Hilfe

↔

---

↔  10 nur eine Zufallsstichprobe von x % der Datensätze für Analyse verwenden Hilfe

↔  123457 Startwert für Zufallsgenerator Hilfe

8

↓ Option: Umkodierungen und Kein-Wert-Angaben

9

Was soll gezeichnet werden ?

↕  1 1 = Regressionsgerade bzw. -ebene zeichnen  
 0 = nicht, nur Punktwolke zeichnen

10

↓ Grafik-Optionen

11

Programmende

Die Datenpunkte können nach einer Gruppierungsvariablen verschieden markiert werden

12

↓ Option: Gruppierungsvariable

## P45.14.1 Erläuterungen zu den Eingabe-Boxen

### Eingabe-Box 1 bis Eingabe-Box 5:

Siehe "Arbeiten mit Almo-Datenanalyse-System", Abschnitt P0.1 bis P0.4.

### Eingabe-Box 6: Analyse-Variable

Analyse-Variable

unabhängige quantitative Variable  
maximal 2 möglich

↔   Einkommen, Kinderzahl

---

abhängige quantitative Variable  
nur 1 möglich

↔   Rueckrate

Die Analyse-Variablen müssen quantitativ sein. Die unabhängige Variable wird von Almo im Streudiagramm an der (horizontalen) X-Achse und die abhängige an der (vertikalen) Y-Achse dargestellt.

Es können maximal 2 unabhängige Variablen angegeben werden. Wir wollen aber zunächst den Fall mit einer unabhängigen Variablen betrachten.

### Eingabe-Box 7: Option: Ein- und Ausschliessen von Untersuchungseinheiten

Siehe "Arbeiten mit Almo-Datenanalyse-System", Abschnitt P0.7.

Wird diese Optionsbox geöffnet, dann sieht man mehrere Möglichkeiten Untersuchungseinheiten ein- oder auszuschliessen. In unserem Beispiel wird folgende Eingabe vorgenommen:

↔ 10 nur eine Zufallsstichprobe von x % der Datensätze für Analyse verwenden

↔ 123457 Startwert für Zufallsgenerator

In unserem Beispiel haben wir 1000 Datensätze. Es werden also 1000 Datenpunkte gezeichnet. Die Datenpunkte verschmelzen dabei zu einer Fläche. Deswegen ist es sinnvoll eine Zufallsauswahl von etwa 5 bis 20 % zu ziehen.

### Eingabe-Box 8: Umkodierungen und Kein\_Wert-Angaben

Siehe "Arbeiten mit Almo-Datenanalyse-System", Abschnitt P0.6.

### Eingabe-Box 9: Was soll gezeigt werden ?

Was soll gezeichnet werden ?

↑↓ 1

1 = Regressionsgerade bzw. -ebene zeichnen  
0 = nicht, nur Punktwolke zeichnen

Wird "0" eingesetzt, dann zeichnet Almo eine Punktwolke.

Wird "1" eingesetzt dann fügt Almo noch die Regressionsgerade hinzu – wurden 2 unabhängige Variable angegeben, dann die Regressionsebene.

Wurde in der Eingabe-Box 7 "Nur eine Zufallsauswahl der Datensätze verwenden" 5 (also 5 % der Datensätze) angegeben, dann errechnet Almo folgende Regressionsgleichung:

$$Y = -0.05 X + 5025$$

Werden alle Datensätze für das Streudiagramm verwendet, dann ist die Regressionsgleichung:

$$Y = -0.05 X + 4625$$

Trotz der sehr kleinen Stichprobe bleibt der Regressionskoeffizient von X mit -0.05 unverändert. Das muß aber nicht immer so sein. Nur die Konstante ändert sich von 5025 auf 4625.

Da der Regressionskoeffizient für X (das Einkommen) mit -0.05 sehr klein ist, empfiehlt es sich in der Eingabe-Box "Umkodierungen und Kein\_Wert-Angabe" das Einkommen mit 10000 zu dividieren.

Wir schreiben in diese Eingabe-Box

$$\text{Einkommen} = \text{Einkommen} / 10000 ;$$

Semikolon zum Schluß nicht vergessen !!

Der Regressionskoeffizient für X bleibt ziffernmäßig gleich, er wird jedoch um den Faktor 10000 größer: -535.99

Die Grafik bleibt unverändert nur die Maßzahlen für die X1-Achse ändern sich selbstverständlich.

### **Eingabe-Box 12:** Option "Gruppierungsvariable"

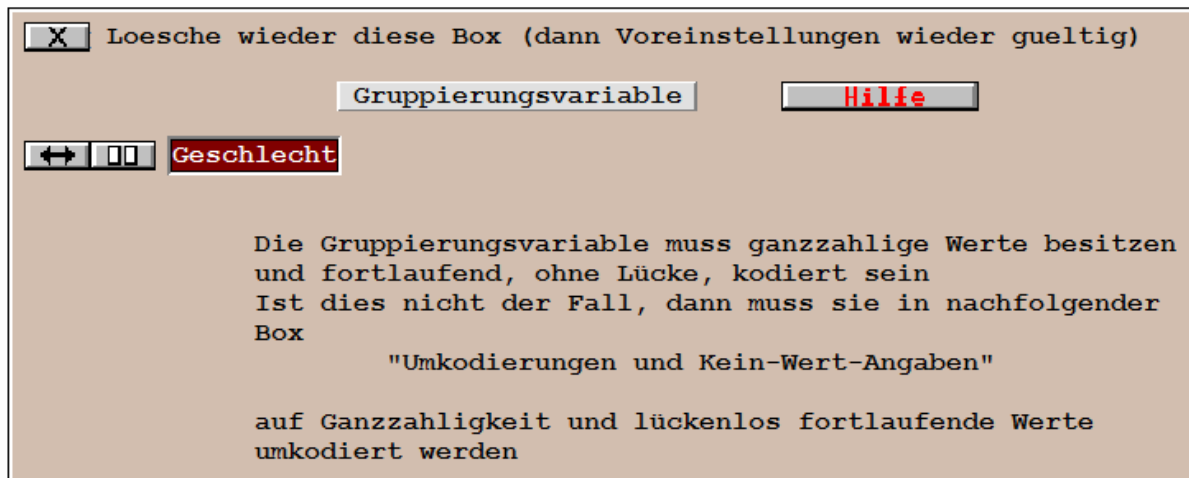
Die Datenpunkte können nach einer Gruppierungsvariablen verschieden markiert werden



Option: Gruppierungsvariable

Wird diese Optionsbox geöffnet, dann sieht man folgendes:

Die Datenpunkte können nach einer Gruppierungsvariablen verschieden markiert werden



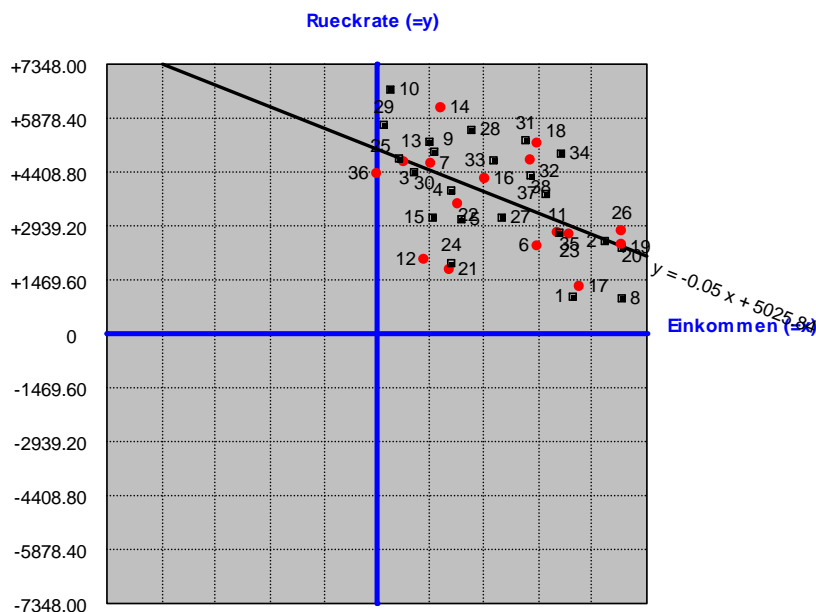
Wir haben das Geschlecht als Gruppierungsvariable eingesetzt. Die Datenpunkte werden in diesem Fall verschieden markiert, die Männer als roter Punkt, die Frauen als kleines schwarzes Viereck.

Die Gruppierungsvariable muss ganzzahlige Werte besitzen und fortlaufend, ohne Lücke, kodiert sein. Ist dies nicht der Fall, dann muss sie nachfolgend im 2. Eingabefeld auf Ganzzahligkeit und lückenlos fortlaufende Werte umkodiert werden.

## P45.14.2 Ausgabe für 2 Variable

Almo liefert ein Ergebnis, das nur aus Grafikdaten und einem Grafikknopf besteht. Nach Klick auf diesen sieht man folgendes

Streudiagramm  
 ● =m  
 ■ =w



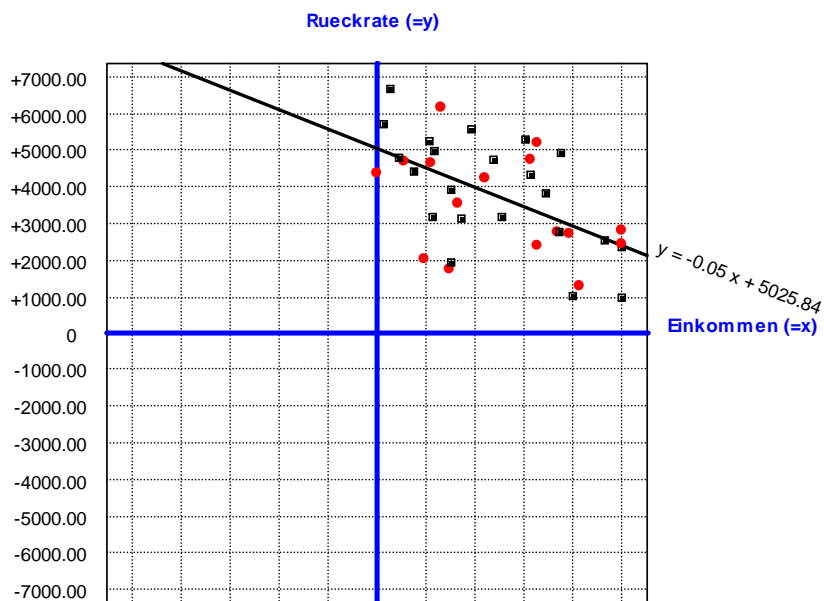
Diese Grafik kann "verschönert" werden. Wir benutzen dazu einige Knöpfe auf der rechten Grafikleiste im Grafik-Editor



Wir haben auf den Knopf "Wände: grau/weiß" geklickt. Dadurch wird der Hintergrund weiß. Dann haben wir die Checkbox "Name an Punkt" deaktiviert. Dadurch werden die Datensatznummern an den Punkten (deren "Namen") gelöscht. Bei "Schrittweite für Maßzahlen haben wir für die x-Achse (das Einkommen) 10000 und für die y-Achse (die Rückzahlungsrate) 1000 eingesetzt. Durch Klick auf die Regressions-Gleichung und Verschieben mit gedrückter linker Maustaste haben wir dann die Formel an eine besser sichtbare Stelle verschoben.

So erhalten wir folgende Grafik

Streudiagramm  
● =m  
■ =w



### P45.14.3 Streudiagramm für 3 Variable

In der Eingabe-Box "Analyse-Variable" geben wir 2 unabhängige Variable an

**Analyse-Variable**

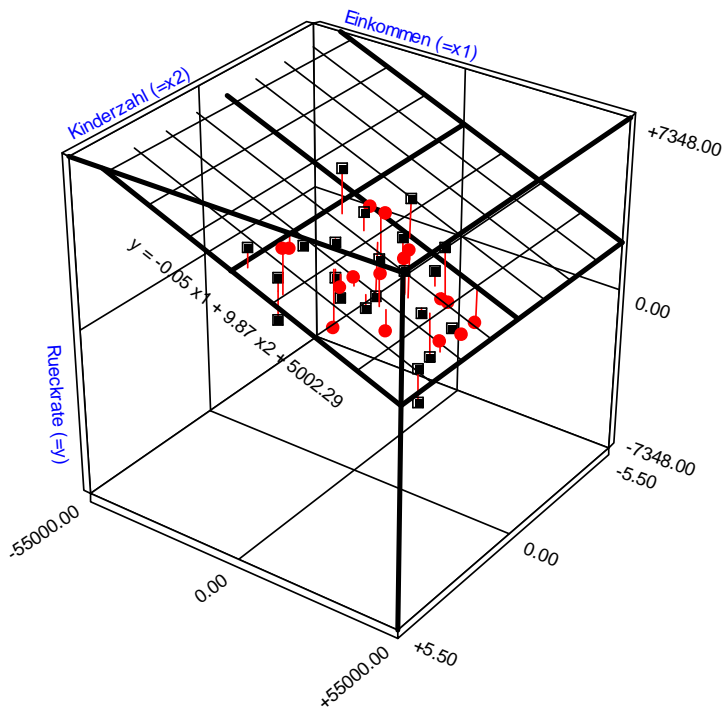
unabhängige quantitative Variable  
maximal 2 möglich

---

abhängige quantitative Variable  
nur 1 möglich

Die anderen Einstellungen lassen wir wie beim 2-Variablen-Streudiagramm. Also liefert wieder nur eine Ausgabe, die aus Grafikdaten und einem Grafikknopf besteht. Wir erhalten folgendes 3-dimensionale Streudiagramm (das wir ebenfalls durch einige Änderungen im Grafik-Editor "verschönert" haben).

Streudiagramm  
 ● =m  
 ■ =w



Die Gleichung der Regressionsebene ist

$$Y = -0.05 X_1 + 9.87 X_2 + 5002$$

Da der Regressionskoeffizient für X1 (das Einkommen) mit -0.05 sehr klein ist, empfiehlt es sich in der Eingabe-Box "Kein\_Wert-Angabe und Umkodierungen" das Einkommen mit 10000 zu dividieren.

Wir schreiben in diese Eingabe-Box

$$\text{Einkommen} = \text{Einkommen} / 10000 ;$$

Semikolon zum Schluß nicht vergessen !!

Der Regressionskoeffizient für X1 bleibt ziffernmäßig gleich, jedoch um den Faktor 10000 größer: -536.73

Die Grafik bleibt unverändert nur die Maßzahlen für die X1-Achse ändern sich selbstverständlich.

## Kapitel 8: Mehrfach-Zusammenhänge untersuchen

### **P45.15 Schritt 11a: Ursachen für die Zielvariable: Allgemeines Lineares Modell (ALM)**

Wir legen eine „Zielvariable“ fest, beispielsweise das Rückzahlungsverhalten. Unsere Frage lautet: Was sind die Ursachen dafür, daß ein Kredit zurückgezahlt wird bzw. nicht zurückgezahlt wird – und wie ist die jeweilige Einflußstärke der gefundenen ursächlichen Variablen?

Es ist nun sinnvoll 3 Typen von Zielvariablen zu unterscheiden:

- (1) Zielvariable, die nur 2 Ausprägungen haben – wie z.B. Rückzahlung: ja, nein. Wir bezeichnen sie als nominal-dichotome oder kurz als dichotome Zielvariable.
- (2) Zielvariable, die nominal sind und mehr als 2 Ausprägungen besitzen, z.B. gekaufte Produkt: Möbel, Technik, Kleidung. Wir bezeichnen sie als nominal-polytome oder kurz als polytome Zielvariable.
- (3) Zielvariable, die quantitativ (oder auch ordinal sind), z.B. Einkommen oder Laufzeit.

Anmerkung zur Terminologie: Wir bezeichnen die „Zielvariable“ gelegentlich auch als „abhängige Variable“ und die „ursächlichen Variablen“ auch als „unabhängige Variable“.

#### **Die Einbeziehung ordinaler Variable in das ALM**

Almo ermöglicht es, ordinale Variable als Zielvariable und/oder ursächliche Variable in das ALM einzusetzen. Zur Berechnung der Streuungsmatrix verwendet es dabei den Groß-Gamma-Kalkül (siehe dazu P45.12.3.1). Auf die so gewonnene Streuungsmatrix wird dann der übliche ALM-Kalkül angewendet. Diese Vorgehensweise hat ihre Probleme, insbesondere wenn über den F- bzw. t-Test Signifikanzen ermittelt werden sollen. Siehe dazu Handbuch „P20 Allgemeines Lineares Modell“, Abschnitt P20.6.9. Wir werden deswegen hier im Rahmen des Almo-Data-Mining-Systems auf die Einbeziehung ordinaler Variabler nicht weiters eingehen. Es sei aber noch darauf hingewiesen, daß Almo mit Prog20m8 eine zweite Möglichkeit anbietet, ordinale Variable in der Form der Rangvariablen (als Zielvariable) im Rahmen des ALM zu verarbeiten.

Auch bei der Logit- und Probitanalyse ist es möglich eine ordinale abhängige Variable zu verarbeiten. Siehe hierzu die Almo-Maskenprogramme Prog22m und Prog22mb sowie die Darstellung im Handbuch, Teil 4, Fortgeschrittene Verfahren, Abschnitt P22.2.6.

### **P45.15.1 Ursachen für die Zielvariable: Zielvariable ist dichotom**

#### **P45.15.1.0 Eine theoretische Vorbemerkung**

Im folgenden werden wir das Allgemeine Lineare Modell (ALM) auf den Fall anwenden, daß die Zielvariable dichotom-nominal oder polytom-nominal oder quantitativ ist. Die Anwendung des ALM auf quantitative Variable ist unproblematisch. Seine Anwendung auf dichotome und polytome Zielvariable schafft jedoch 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 die Rückzahlung eines Kredits  $p=1.08$  (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 werden wir in P45.15.1.6 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 Reproduzierbarkeit 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 Prog45mf die gewichtete Kleinste-Quadrate-Schätzung für nominal-dichotome Zielvariable als Option an. Siehe dazu die Erläuterungen zur Optionsbox „gewichtete Analyse“ in P45.15.1.2 und in Abschnitt P45.15.1.7. In Prog45gw in Abschnitt P45.15.2.2 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 werden es in P45.7.6 darstellen. 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.

Ein wichtiges Kriterium für die Brauchbarkeit eines Modells ist seine Reproduzierbarkeit: Kann die Zugehörigkeit von Untersuchungsobjekten aus der **Stichprobe** zu den Ausprägungen der Zielvariablen zufriedenstellend reproduziert werden. Beispiel: Kann das Modell zufriedenstellend reproduzieren, ob Personen ihren Kredit zurückzahlen oder nicht. Hier sind in der Praxis bei dichotomen Zielvariablen beide Modelle, das (ungewichtete) ALM und das Logit-Modell, gleich gut, während die gewichtete Kleinste-Quadrate-Schätzung hier etwas schlechter abschneidet.

Eine andere sehr wichtige Frage ist es, ob es ein Modell erlaubt, von den Stichprobenergebnissen auf die Grundgesamtheit zu schließen – ob es z.B. erlaubt ist, mit Hilfe der Regressionskoeffizienten und Effekte, die aus den Stichprobendaten errechnet wurden, Prognosen hinsichtlich einer oder mehrerer Personen zu leisten, die nicht zur Stichprobe gehören, aber aus derselben Grundgesamt stammen. Hier wird der Statistiker Zweifel gegenüber dem auf nominal-dichotome oder nominal-polytome Variable angewandten ALM anmelden. Unsere Empfehlung ist folgende:

## **Empfehlung**

1. Ist die Zielvariable nominal-dichotom oder nominal-polytom, dann sollte man das in Abschnitt P45.16 dargestellte Logit-Modell verwenden. Da dieses Modell dem weniger geübten Datenauswerter bei der inhaltlichen Interpretation der Ergebnisse Schwierigkeiten macht, sollte er zuvor "explorativ" das (ungewichtete) ALM rechnen. Es ermöglicht eine sehr einsichtige Interpretation seiner Ergebnisse und hilft damit, die komplexeren Ergebnisse des Logitmodells eher zu verstehen. In aller Regel stimmen die Ergebnisse des "nicht ganz korrekten" ALM und des korrekten Logit-Modells ohnehin überein.
2. Einen Kompromiß stellt das (in Abschnitt P45.15.1.8 und P45.15.2.2 dargestellte) gewichtete ALM dar. Es ist statistisch korrekt und seine Ergebnisse sind wie beim ungewichtete ALM inhaltlich gut zu interpretieren. Seine Reproduzierbarkeit ist allerdings, wie wir bereits ausführten, etwas schwächer.

Eine weitere Möglichkeit, nominale Zielvariable zu analysieren, besteht in der (kanonischen) Diskriminanzanalyse. Einige Autoren (etwa Urban, 1993) vertreten die Auffassung, daß bei der Diskriminanzanalyse auf Seiten der ursächlichen Variablen nur quantitative aber nicht nominale Variable zulässig sind - was natürlich die praktische Anwendung dieses Verfahrens sehr einschränkt. Außerdem entstehen, bei der auf nominal-polytome Zielvariable angewandten Diskriminanzanalyse, Koeffizienten, die inhaltlich nicht interpretierbar sind. Wir haben deswegen die Diskriminanzanalyse in das Almo-Data-Mining-System nicht aufgenommen. Es ist jedoch als Prog29 und Prog27 in Almo enthalten und im Handbuch, Teil 4, "Fortgeschrittene Verfahren" ausführlich dargestellt.

Wir wollen nun ein Beispiel rechnen bei dem wir die Variable „Rückzahl: nein, ja“ als dichotome Zielvariable verwenden. Wir rechnen zuerst das (ungewichtete) ALM.

#### ***P45.15.1.1 Eingabe in Prog45mf***

Das nachfolgende Programm ist das generelle Programm für das ALM. Es wird in dieser Form auch eingesetzt, wenn die Zielvariable quantitativ ist, aber natürlich auch wenn sie nominal-dichotom oder nominal-polytom ist, wobei dann auch die gewichtete Variante des ALM gerechnet werden kann. Als Beispiel für die Eingabe haben wir den Fall der nominal-dichotomen Zielvariablen verwendet.

Prog45mf.Msk

Wirkungsstärke der ursächlichen Variablen  
hinsichtlich der Zielvariablen

Die Zielvariable kann nominal, ordinal oder quantitativ sein  
Es wird ein Allgemeines Lineares Modell gerechnet

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

1

Vereinbare Variable=  ;

2

Option: Weitere Vereinbarungen - nur wenn Almo dazu auffordert

3

"C:\Almo7\TESTDAT\DatMin.nam"  
  **zeige**                    zeige = Namensdatei in Output zeigen  
leer = nicht

4

**erzeuge zusätzliche Namensfelder**

5

"C:\Almo7\TESTDAT\DatMin.dir"

6

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 Zielvariable

---

ordinale Zielvariable

---

nominale Zielvariable

**Rueckzahl**

**Ursächliche Variable**

ursächliche nominale Variable

**Wohnort, Hausbesitz, Produkt**

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

---

**Wohnort, Hausbesitz, Produkt**  
paarweise Vergleiche (Kontraste) für die  
ursächlichen nominalen Variablen rechnen

---

ursächliche quantitative Variable

**Einkommen, Rueckrate, Laufzeit**

---

ursächliche ordinale Variable

7

Option: Ein- und Ausschliessen von Untersuchungseinheiten

8

Loesche wieder diese Box

**Umkodierungen und Kein-Wert-Angaben**

Umkodierungen   
Kein\_Wert-Angabe

**Einkommen = Einkommen / 10000;**

erzeuge zusätzliche Felder für Umkodierungen / Kein\_Wert-Angaben

---

Kontrollieren, ob Umkodierung so erfolgt wie gewünscht  
diese Variablen ...

**Einkommen**

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

9

Option: Ausreisser vom Typ 1 identifizieren

- 10  Option: Spezielle Kein-Wert-Behandlung
- 11  Option: Untersuchungseinheiten gewichten
- 12  Option: Streuungsmatrix
- 13  Option: Verfahren
- 14  Option: Gewichtete Kleinste-Quadrate-Schätzung
- 15  Option: Prognosewerte und Residuen
- 16  Option: Wertemuster
- 17  Option: "Aussehen" der auszugehenden Tabelle bzw. Matrix
- 18  Grafik-Optionen
- 19  Option: Die errechneten Koeffizienten in eine Datei speichern
- 20   **Ausgabe der Ergebnisse**  
 0= Ergebnisse in voller Länge ausgeben  
 1= Ergebnisse etwas verkürzt ausgeben  
 2= Ergebnisse stark verkürzt ausgeben
- Ausgabe der Ergebnisse**  
 1= Basisstatistiken ausgeben  
 2= Basisstatistiken und "diverse Werte" ausgeben  
 0= nicht

## P45.15.1.2 Erläuterungen zu den Eingabe-Boxen

**Eingabe-Box 1:** Speicher für x Variable.

Siehe "Arbeiten mit Almo-Datenanalyse-System", Abschnitt P0.1.

**Eingabe-Box 2:** Weitere Vereinbarungen

Siehe "Arbeiten mit Almo-Datenanalyse-System", Abschnitt P0.2.

**Eingabe-Box 3:** Datei der Variablennamen

Siehe "Arbeiten mit Almo-Datenanalyse-System", Abschnitt P0.3.

**Eingabe-Box 4:** Freie Namensfelder

Siehe "Arbeiten mit Almo-Datenanalyse-System", Abschnitt P0.3.

**Eingabe-Box 5:** Datei aus der gelesen wird

Siehe "Arbeiten mit Almo-Datenanalyse-System", Abschnitt P0.4.

**Eingabe-Box 6:** Zielvariable

**Zielvariable** 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 Zielvariable

ordinale Zielvariable Hilfe

nominale Zielvariable Hilfe

Rueckzahl

Wir geben „Rueckzahl“ als nominale Zielvariable an. Daß sie dichotom ist brauchen wir nicht eigens zu vermerken.

## Eingabe-Box 7: Ursächliche Variable

**Ursächliche Variable**

ursächliche nominale Variable

**Wohnort, Hausbesitz, Produkt**

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

---

**Wohnort, Hausbesitz, Produkt**  
paarweise Vergleiche (Kontraste) für die  
ursächlichen nominalen Variablen rechnen

---

ursächliche quantitative Variable

**Einkommen, Rueckrate, Laufzeit**

---

ursächliche ordinale Variable

Die Frage ist, welche Variable wollen wir als ursächliche in unsere Analyse einbeziehen. Beim Korrelieren haben wir festgestellt, daß die Variable Rueckzahl mit folgenden Variablen mit Werten über 0.09 korreliert:

Mit	Wohnort	0.13
	Hausbesitz	0.11
	Produkt	0.15
	Einkommen	0.43
	Rückzahlungsrate	0.43

Es ist also sinnvoll, diese Variable als ursächliche zu verwenden.

Es spricht aber nichts dagegen, alle anderen Variablen als ursächliche einzusetzen, sofern sie aus inhaltlichen Überlegungen auch als Ursachen betrachtet werden dürfen.

*Eingabefeld 1: Ursächliche nominale Variable*

Wir haben uns entschlossen, folgende Variable einzusetzen:

Wohnart	(Stadt, Land)
Hausbesitz	(ja, nein)
Produkt	(Kleidung, Möbel, Technik)

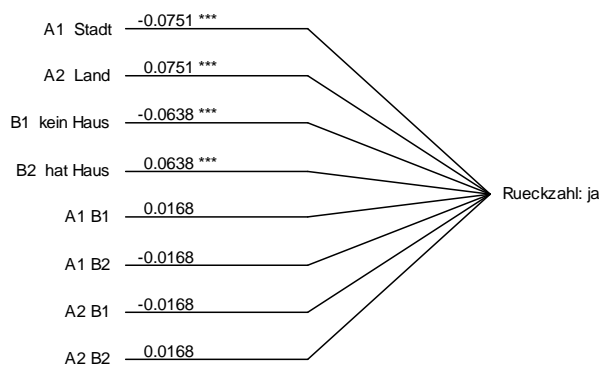
## Eingabefeld 2: Interaktionen

Wir würden empfehlen, in der 1. Phase des Data-Mining-Prozesses keine Interaktionen einzuschließen. Erst nachdem sich der Benutzer mit Prog45mf und seinem Output ausreichend beschäftigt hat, sollte er mit den Interaktionen experimentieren.

Wir wollen den Begriff der **Interaktion** kurz erläutern:

Betrachten wir ein medizinisches Beispiel. Wenn „hoher Cholesterinspiegel“ und „Raucher“ zusammenkommen, dann entsteht ein zusätzlicher Effekt auf die Wahrscheinlichkeit einen Herzinfarkt zu erleiden, der zu den Einzeleffekten von „Cholesterin“ und „Rauchen“ dazukommt. Das ist der Interaktionseffet „Cholesterin \* Rauchen“.

Betrachten wir nun unser „Rückzahlungs-Beispiel“. Wenn wir auf Seiten der ursächlichen Variablen nur die beiden nominalen Variablen Wohnort und Hausbesitz haben, dann läßt sich eines der Ergebnisse des „Allgemeinen Linearen Modells“ (das mit Prog45mf gerechnet wird) als folgendes Flußdiagramm darstellen



Die abhängige Variable (auf der rechten Seite des Flußdiagramms) ist die Wahrscheinlichkeit der Rückzahlung. Auf der linken Seite stehen die ursächlichen Variablen.

Auf den Strichen des Flußdiagramms stehen die Effekte der ursächlichen Variablen hinsichtlich der Rückzahlungswahrscheinlichkeit.

Betrachten wir eine Person mit A1 (Wohnort:Land) und B2 (hat Haus).

	<u>Effekt</u>
A2 Wohnort:Land	0.0751
B2 hat Haus	0.0638
Interaktion A2 B2	0.0168
Zwischensumme	0.1557
+ Konstante	0.7339
Summe	0.8896

Diese Person hat eine Rückzahlungswahrscheinlichkeit von 0.8896 also 88,96%.

Die „Konstante“ wird im Flußdiagramm nicht gezeigt, da sie inhaltlich kaum interpretierbar ist. Sie wird jedoch im Output ausgegeben und ist für unseren Kalkül notwendig. Siehe dazu auch P45.15.1.6 Prognosefähigkeit.

Die Interaktion A2 B2 kann verstanden werden als der zusätzliche Effekt der entsteht, wenn die beiden Ausprägungen „Land“ und „hat Haus“ zusammenkommen. In unserem Beispiel ist der Interaktionseffekt nicht signifikant.

Wenn 3 ursächliche nominale Variable vorhanden sind, dann können Interaktionen 3. Ordnung entstehen; bei 4 Variablen dann Interaktionen 4. Ordnung etc. Dabei entstehen dann Interaktionen, die inhaltlich nicht mehr interpretierbar sind, die rechnerische Artefakte sind.

Unsere Empfehlung ist, mit Interaktionen in Prog45mf zu experimentieren – sie aber nur im entgeltigen Modell zu belassen, wenn sie noch signifikant sind und inhaltlich eindeutig interpretierbar sind.

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

Die Effekte der nominalen Variablen werden untereinander verglichen. Die Vergleiche werden auf ihre Signifikanz getestet. In Abschnitt P45.15.1.4 werden wir die Vergleiche ausführlich erläutern.

#### *Eingabefeld 4: Ursächliche quantitative Variable*

Einkommen, Rückzahlungsrate und Laufzeit werden von uns als quantitative ursächliche Variable eingesetzt.

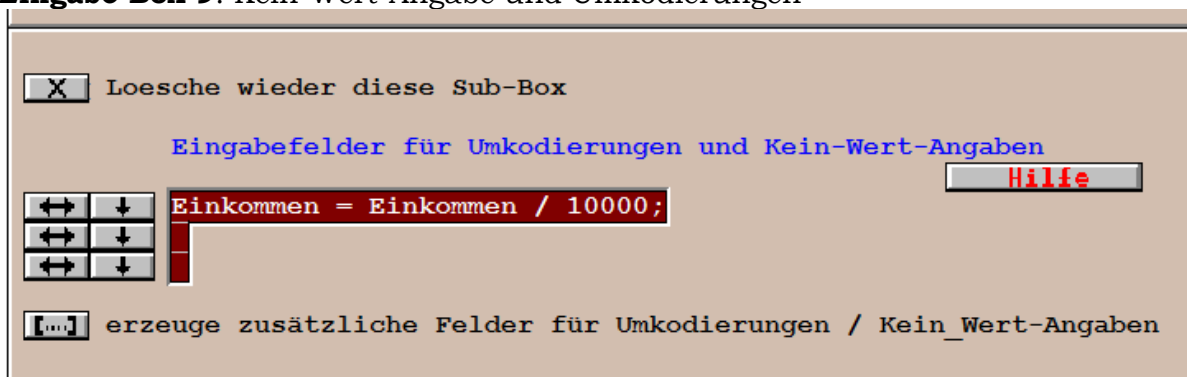
#### *Eingabefeld 5: Ursächliche ordinale Variable*

Keine der Variablen in unserem Beispiel wird von uns als ordinal eingeschätzt. Zur Problematik der ordinalen Variablen im ALM siehe Handbuch zum ALM, P20, Abschnitt P20.6.9.

#### **Eingabe-Box 8: Option: Ein- und Ausschliessen von Untersuchungseinheiten**

Siehe "Arbeiten mit Almo-Datenanalyse-System", Abschnitt P0.7.

#### **Eingabe-Box 9: Kein-Wert-Angabe und Umkodierungen**



Wie diese Eingabe-Box zu bedienen ist, haben wir ausführlich im "Arbeiten mit Almo-Datenanalyse-System", Abschnitt P0.5 dargestellt.

In unserem Beispiel haben wir Einkommen durch 10 000 dividiert. Der Grund dafür ist folgender. Die ursächlichen quantitativen Variablen sind

Einkommen mit einem Wertebereich von	0	bis	50 000
Rueckrate	ca. 600	bis ca.	7 000
Laufzeit	ca. 1	bis ca.	30

Wenn wir diese Variable, ohne sie umzukodieren, in die Analyse einfügen, dann erhalten wir folgende Regressionskoeffizienten.

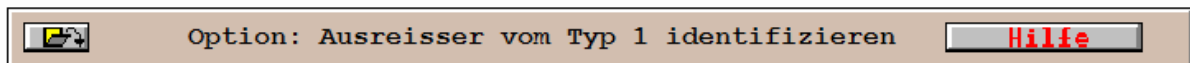
Einkommen	-0.0000098
Rueckrate	0.0001030
Laufzeit	-0.0073030

Almo gibt (teilweise) nur 4 Kommastellen aus, so daß der Regressionskoeffizient für Einkommen mit  $-0.0000$  ausgegeben wird. Wir haben deshalb Einkommen mit 10 000 dividiert. Dabei erhalten wir dann folgenden Regressionskoeffizienten

$-0.098402$

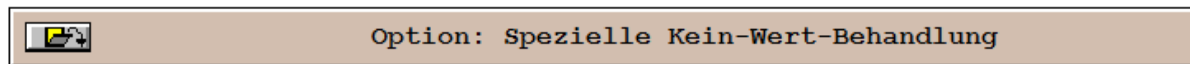
Das Komma ist also um 4 Stellen nach rechts verschoben. Eigentlich sinnvoll wäre es auch gewesen, Rueckrate mit 1 000 zu dividieren.

**Eingabe-Box:** Ausreisser identifizieren



Der Begriff "Ausreisser" und die Art und Weise, wie in Almo Ausreisser-Daten erkannt und behandelt werden, sind im Almo-Dokument 23 "Ausreisser entdecken" ausführlich dargestellt.

**Eingabe-Box 10:** Option: Spezielle Kein-Wert-Behandlung



Wurden mit Prog45mo oder Prog45mm fehlende Werte durch Schätzwerte ersetzt, dann wird diese Option nicht benötigt.

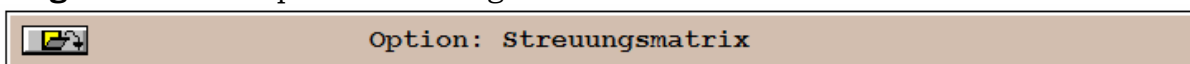
Diese Option wird auch nicht benötigt, wenn der Benutzer das "paarweise Ausscheiden" als Kein-Wert-Behandlung akzeptiert. Dies ist die Voreinstellung in Almo.

Wird die Optionsbox geöffnet, dann präsentiert Almo eine sehr große Eingabe-Box mit 7 Methoden der Kein-Wert-Behandlung. Wir haben die Methoden 1 bis 3 bereits ausführlich in P45.7.1.3, Eingabe-Box 9 bei den Erläuterungen zu Prog45mm und die Methoden 4 bis 7 in P45.6.1, Eingabe-Box 7 bei den Erläuterungen zu Prog45mo beschrieben.

Die Methode 1, das „paarweise Ausscheiden“ haben wir besonders ausführlich in P45.12.4 dargestellt. Wählt der Benutzer diese Methode, dann kann er auch eine Entscheidung darüber treffen, welche Fallzahl Almo für Signifikanzberechnungen einsetzen soll. Das Eingabefeld dafür steht ganz unten in der großen Eingabe-Box. Siehe dazu die ausführliche Darstellung in P45.12.4. Die Voreinstellung in Almo ist das harmonische Mittel.

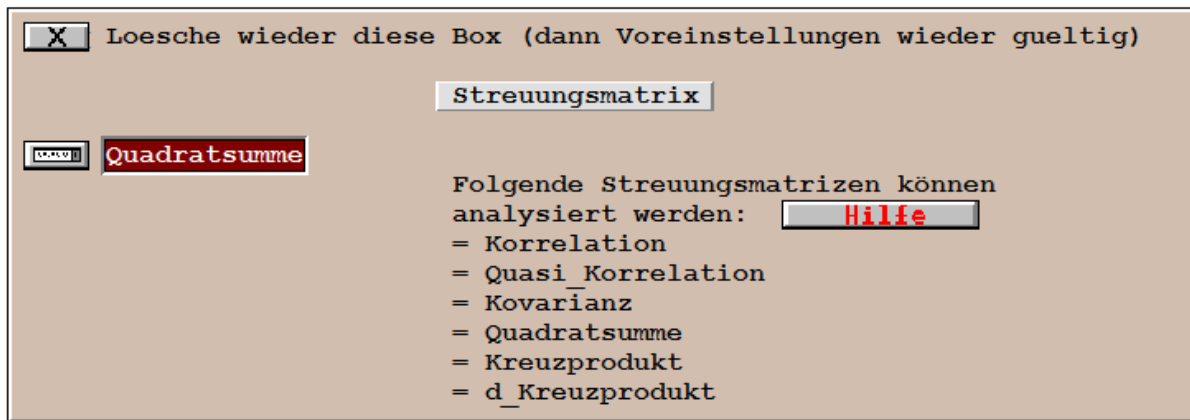
**Eingabe-Box 11:** Option: Untersuchungseinheiten gewichten  
Siehe P0.8.

**Eingabe-Box 12:** Option: Streuungsmatrix



Die Voreinstellung ist "Quadratsumme". 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 Abschnitt P45.12.4.2 und P20.8.5.3.1  
 \* standardisierter Variabler  
 \*\* und teilweise Produkte/Kreuzprodukte (siehe unten)  
 \*\*\* durch n dividiert

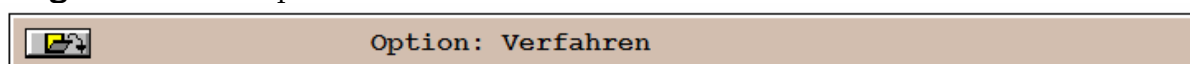
Vorzugsweise sollten Sie "Quadratsumme" verwenden. Nur ausnahmsweise verwenden sollten Sie "Kreuzprodukt" und "d\_Kreuzprodukt".

"d\_Kreuzprodukt" ist die durchschnittliche Kreuzprodukte-Matrix. Sie entsteht dadurch, daß die Kreuzprodukte-Matrix mit n (=der Zahl der Fälle) dividiert wird.

BEACHTET: Bei der Verwendung von "Kreuzprodukt" und "d\_Kreuzprodukt" gibt es noch ein weiteres Problem: Die von Almo ermittelte Gesamtstreuung, die durch alle unabhängigen Variablen erklärte Streuung, die durch die quantitativen/ordinalen Variablen insgesamt erklärte Streuung sind nicht Abweichungsquadratsummen sondern Summen quadrierter Rohwerte bzw. deren Kreuzprodukte. Die F-Werte und Signifikanzen dieser 3 Streuungen sind falsch. Almo teilt dies dem Benutzer auch mit.

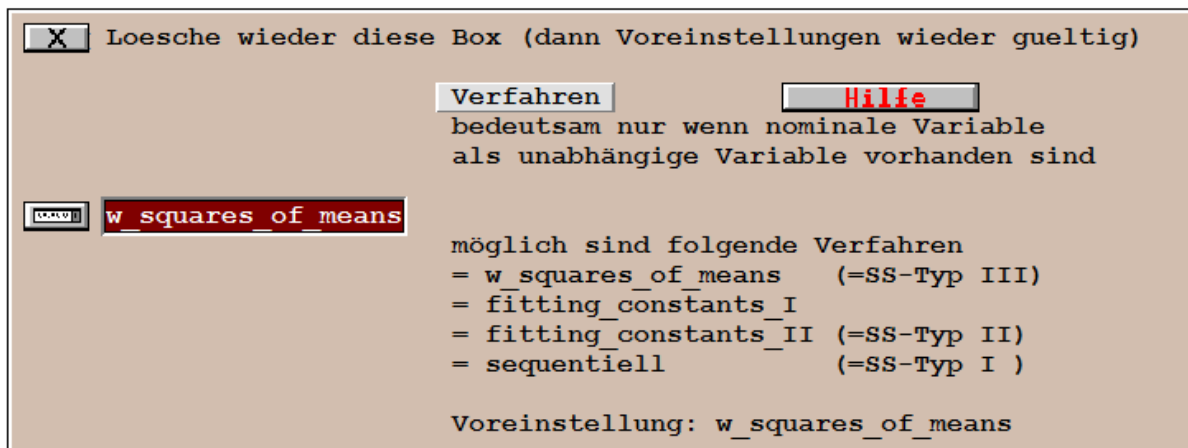
Also: Verwenden Sie "Kreuzprodukt" und "d\_Kreuzprodukt" nur wenn Sie dafür einen guten Grund haben.

### Eingabe-Box 13: 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:



3 Verfahren stehen zur Auswahl:

*w\_squares\_of\_means*. Bezeichnung in SAS und SPSS: SS Typ III

Alle Effekte (Haupteffekte und Interaktionseffekte) werden gegenseitig auspartiiert. Ist das empfehlenswerte Verfahren !!

*sequentiell*. Bezeichnung in SAS und SPSS: SS Typ I

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

*fitting\_constants\_I*

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.

*fitting\_constants\_II*

In SAS und SPSS wird diese spezielle Variante der "fitting constants" als SS Typ II bezeichnet. Diese spezielle Variante kann nur mit den Maskenprogrammen

Prog20my.Msk, Prog20mz.Msk, Prog20mm.Msk

gerechnet werden. Siehe dazu Handbuch zu ALM, P20, Abschnitt P20.8.9. oder das Almo-Dokument 13 "ALM Allgemeines Lineares Modell.PDF"

SS Typ II ist in folgenden zwei Fällen identisch mit "fitting\_constants\_I", wenn

- keine Interaktionen in das Modell aufgenommen werden

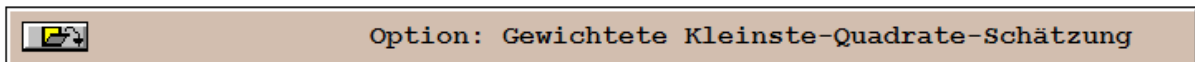
- wenn nur die beiden Faktoren A und B und ihre Interaktion AB, aber keine Kovarianten sich im Modell befinden

BEACHTTE: Bei gleichen Zellenhäufigkeiten erbringen alle Verfahren das gleiche Ergebnis. Geben Sie in diesem Falle am besten "w\_squares\_of\_means" als Verfahren an.

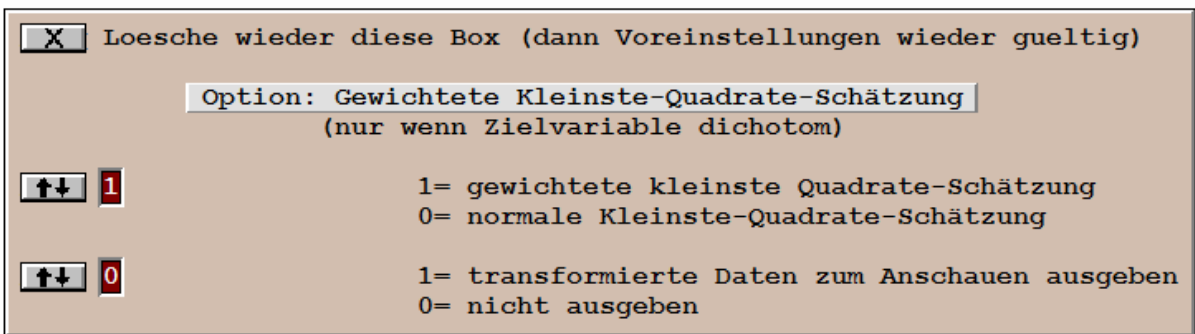
Siehe die ausführliche Darstellung im Handbuch zu P20 „Allgemeines Lineares Modell“, Abschnitt P20.7. bzw. im Almo-Dokument 13.

Der SS Typ IV aus SAS und SPSS ist in Almo nicht enthalten. Siehe dazu unsere "Anmerkungen zu SS Typ IV" im Handbuch zu P20 „Allgemeines Lineares Modell“, Abschnitt P20.7.4.1. bzw. im Almo-Dokument 13

**Eingabe-Box 14:** Option: Gewichtete Kleinste-Quadrate-Schätzung

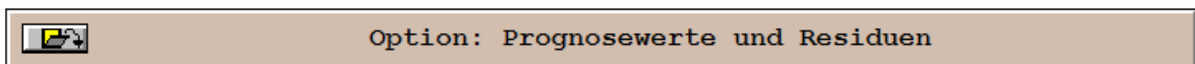


Wird die Optionsbox geöffnet, dann sieht man folgendes:



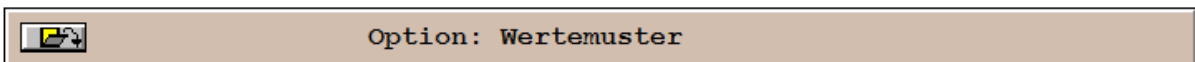
Ist die abhängige Variable nominal-dichotom, dann besteht modellbedingte Varianzheterogenität. Diese kann durch die Methode der "gewichteten Kleinste-Quadrate" beseitigt werden. Wir werden in Abschnitt P45.15.1.8 dieses Verfahren darstellen. Wir empfehlen dieses Verfahren in der ersten Phase des Data-Mining-Prozesses nicht zu verwenden.

**Eingabe-Box 15:** Option: Prognosewerte und Residuen

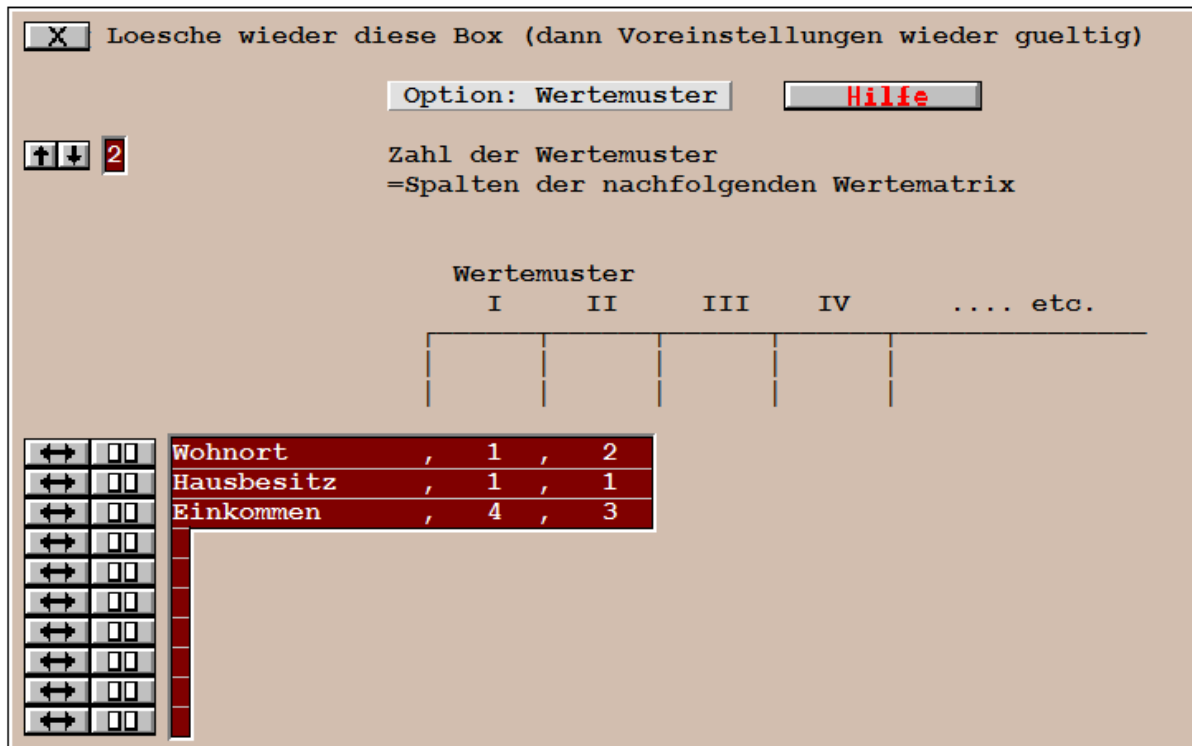


Wir werden diese Optionsbox später in Abschnitt P45.15.1.6 darstellen und erläutern.

**Eingabe-Box 16:** Option: Wertemuster



Wird die Optionsbox durch Klick auf den Knopf mit nach unten weisenden Pfeil geöffnet, dann sieht man folgendes:



Der Benutzer kann sich von Almo berechnen lassen, welchen Wert in der Zielvariablen eine Person mit bestimmten Werten in einer oder mehreren oder allen unabhängigen Variablen hat.

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 besitzen
2. Landbewohner, die kein Haus besitzen und ein Einkommen von 3 besitzen

(Einkommen wurde in der Umkodierungsbox mit 10 000 dividiert)

Wir geben als Zahl der Wertemuster = 2 an und schreiben in die Eingabefelder der Wertemustermatrix

	Wertemuster		
	I	II	III IV .... etc.
[ Wohnort	, 1 ,	2 ]	
[ Hausbesitz	, 1 ,	1 ]	
[ Einkommen	, 4 ,	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.

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. Almo 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:

	Wertemuster		
	I	II	III IV .... etc.
[Geschlecht	, 1 ,	2 ]	
[Alter	, 48 ,	58 ]	
[Einkommen	, 4 ,	kw ]	

kw eingesetzt

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

Almo 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.

	Wertemuster				
	I	II	III	IV	.... etc.
[Geschlecht	, 1	, 2	, Alter	, 48,58	]
[Einkommen	, 7200	, 3500	, Bildung,	5, 3	]

Sie schreiben in ein Eingabefeld 2 oder sogar mehrere Variable mit ihren Werten.

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 17:** Optionen, die das "Aussehen" der auszugebenden Matrix steuern. Siehe dazu "Arbeiten mit Almo-Datenanalyse-System", Abschnitt P0.9.

### Eingabe-Box 18: Grafik-Optionen



Nach Klick auf diesen Knopf erscheint folgende Eingabe-Box:

↓ Loesche wieder diese Box

**Grafik-Optionen**

Almo zeichnet  
 Liniendiagramme  
 Balkendiagramme  
 Flussdiagramme   
 lineare Funktionen

Almo Almo = Almo-Grafik ausgeben  
 = keine Grafik

---

**Wohnort**

**1** 0 = für jede Ausprägung der Grupp.variablen  
eine eigene Grafik zeichnen  
1 = alle Ausprägung der Grupp.variablen  
in einer gemeinsamen Grafik zeichnen

---

**1** 1 = Almo-Grafiken in Ergebnisliste einsetzen  
0 = nicht

Almo zeichnet - standardmäßig, auch ohne daß diese Optionsbox aktiviert wurde - ein Flußdiagramm, in dem die Koeffizienten der ursächlichen Variablen hinsichtlich der Zielvariablen eingetragen sind. Außerdem zeichnet Almo je eine lineare Funktion für die ursächlichen quantitativen Variablen hinsichtlich der Zielvariablen. Dabei wird die ursächliche quantitative Variable an die x-Achse geschrieben und die Zielvariable an die y-Achse.

In unserem Beispiel wird die lineare Funktion für "Einkommen" (x-Achse) und "Wahrscheinlichkeit der Kredit-Rückzahlung" (y-Achse) gezeichnet.

Danach wird die lineare Funktion für die Rückzahlungsrate und die Kredit-Laufzeit gezeichnet.

Die jeweils anderen ursächlichen quantitativen Variablen werden dabei auf ihren Mittelwert gesetzt. Auch die Dummies der ursächlichen nominalen Variablen (Wohnort, Hausbesitz, Produkt) werden auf ihre Mittelwert gesetzt. Dieser entspricht dem Anteilswert der Ausprägungen.

Bei der Interpretation der Almo-Ergebnisse in Abschnitt P45.15.1.4 werden wir ausführlich die von Almo erzeugten Kurvendiagramme besprechen.

*Eingabefeld 1:* Der Eintrag "Almo" bedeutet: Es werden, wie oben ausgeführt, Flußdiagramme und Kurvendiagramme erzeugt. Dies geschieht auch standardmäßig, ohne daß diese Optionsbox aktiviert wurde. Der Eintrag "0" (Null) bedeutet: Es wird keine Almo-Grafik erzeugt.

*Eingabefeld 2:* Es kann eine oder mehrere Gruppierungsvariable angegeben werden. In unserem Beispiel wird "Wohnort" als Gruppierungsvariable angegeben. Almo zeichnet dann die linearen Funktionen (so wie oben beschrieben) für die beiden Ausprägungen "Stadt" und "Land".

Wir werden später in Abschnitt P45.15.1.4.2 zeigen, daß man auch Gruppierungsvariable kombinieren kann

Beachte: Als Gruppierungsvariable können nur Variable verwendet werden, die in der Eingabe-Box "Ursächliche Variable" als nominale Variable angegeben wurden.

*Eingabefeld 3:* Betrachten wir ein Beispiel: Als Gruppierungsvariable wurde beispielsweise "Wohnort" mit den beiden Ausprägungen "Stadt" und "Land" eingesetzt.

Wenn Sie in das Eingabefeld 1 einsetzen, dann zeichnet Almo eine einzige Grafik, in der sich zwei Kurven befinden, eine für die Stadtbewohner und eine für die Landbewohner. Sie erkennen dann sehr gut den Unterschied zwischen den beiden Bewohnern.

Wenn Sie in das Eingabefeld 0 einsetzen, dann zeichnet Almo zwei Grafiken mit je einer Kurve, eine für die Stadtbewohner und eine für die Landbewohner. Da Sie dann 2 getrennte Grafiken besitzen, ist der Unterschied zwischen den Bewohnern nicht so leicht zu erkennen.

Wenn Sie eine Gruppierungsvariable angegeben haben, die viel Ausprägungen besitzt, beispielsweise 10, dann werden bei Eingabe von 1 alle 10 Kurven in einer Grafik dargestellt. Die Kurven können dann so dicht beieinander liegen, dass sie nicht mehr unterscheidbar sind. In einem solchen Fall ist es besser eine 0 in das Eingabefeld einzusetzen. Dieser Fall tritt vor allem dann auf, wenn sie mehrere Gruppierungsvariable durch das Wort MIT miteinander kombinieren. Es können dann sehr viele Ausprägungskombinationen entstehen, für die Almo je eine Kurve zeichnet. Siehe dazu P45.15.1.4.1 und 2.

*Eingabefeld 4:* Wenn Sie '1' eingeben, dann werden die Almo-Grafiken direkt in die Ergebnisliste eingesetzt.

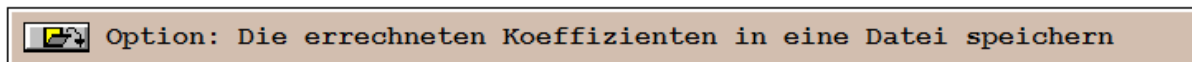
Wenn Sie durch die Ergebnisliste blättern oder scrollen, dann werden Ihnen (anschliessend an Tabellen und Matrizen) auch die Almo-Grafiken gezeigt.

Wenn Sie '0' eingeben, dann können die Almo-Grafiken nur im Grafik-Editor angeschaut werden.

In der Ergebnisliste ist dann (anschliessend an Tabellen und Matrizen) nur ein Grafikknopf enthalten. Durch Klick auf diesen Knopf gelangen Sie in den Grafik-Editor, wo Ihnen die Grafik gezeigt wird.

Der Grafikknopf ist auch vorhanden, wenn Sie '1' eingeben, wenn also die Grafiken in die Ergebnisliste eingesetzt werden. Durch Klick auf den Grafikknopf in der Ergebnisliste können die Grafiken dann im Grafik-Editor bearbeitet werden und von dort durch Klick auf den Knopf "Einsetzen" in der veränderten Form wieder in die Ergebnisliste übergeben werden. Eine Bearbeitung der Grafik wird häufig notwendig sein. Man möchte beispielsweise die Balken in einem Balkendiagramm schlanker abgebildet haben, als dies Almo standardmäßig tut. Oder man möchte mehr Perspektive in die Grafik bringen etc. Oder man möchte noch zusätzliche Beschriftungen einfügen.

**Eingabe-Box 19:** Option: Die errechneten Koeffizienten in eine Datei speichern

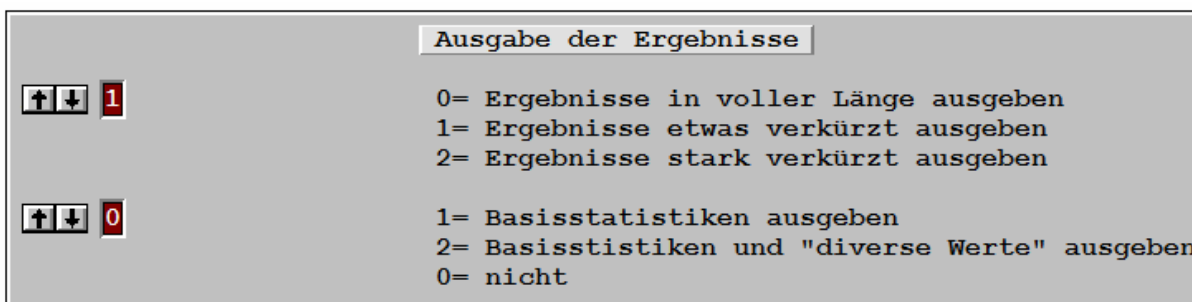


Wird die Optionsbox durch Klick auf den Knopf mit nach unten weisenden Pfeil geöffnet, dann sieht man folgendes



Die errechneten Koeffizienten werden mit einigen Zusatzinformationen in eine Datei gespeichert. Es besteht dann die Möglichkeit mit Prog45mp eine Prognose für die Personen einer anderen Datei zu leisten. Wir werden darauf ausführlich in Abschnitt P45.17 kommen.

**Eingabe-Box 20:** Ausgabe der Ergebnisse



In das Eingabefeld kann geschrieben werden:

- 0 dann wird das Ergebnis aus Prog45mf in voller Länge ausgegeben
- 1 Ergebnisse werden etwas verkürzt ausgegeben
- 2 Ergebnisse werden stark verkürzt ausgegeben

Unsere Empfehlung ist, drei Analysen zu rechnen; zuerst mit stark verkürzter, dann mit etwas verkürzter und dann mit voller Ausgabe.

### P45.15.1.3 Ausgabe bei stark verkürzten Ergebnissen

Wurde in der letzten Eingabe-Box 20 im 1. Eingabefeld 2 eingesetzt, dann liefert Almo folgende stark reduzierte Ausgabe.

```
Zahl der insgesamt eingelesenen Einheiten      1000

Zahl der in die Analyse einbezogenen Einheiten 1000
=====

Zusammenfassung

Streuungsquelle                                Korrel   Signifikanz
                                                Koeff.   p         (1-p)100
-----
alle unabh. Var. zusammen                    0.5561   0.0000   99.9995
quant./ordin. Var. zusammen                  0.5188   0.0000   99.9995
nominale Variable zusammen                   0.2757   0.0000   99.9995

V4 Einkommen                                0.3020   0.0000   99.9995
V7 Rueckrate                                0.2965   0.0000   99.9995
V8 Laufzeit                                  0.0988   0.0020   99.8023
V1 Wohnort                                    0.1523   0.0000   99.9995
V6 Hausbesitz                                0.1691   0.0000   99.9995
V9 Produkt                                    0.1729   0.0000   99.9984
```

Alle ursächlichen Variablen zusammen korrelieren mit der Zielvariablen mit

$$R = 0.5561$$

R ist der "multiple" Korrelationskoeffizient. Diese Korrelation ist mit 99.9995 % hoch signifikant.

Die ursächlichen quantitativen Variablen besitzen zusammen genommen hinsichtlich der Zielvariablen einen "partiellen multiplen Korrelationskoeffizient" von 0.5188, der ebenfalls mit 99.9995 % hoch signifikant ist.

Die ursächlichen nominalen Variablen besitzen zusammen genommen hinsichtlich der Zielvariablen einen "partiellen multiplen Korrelationskoeffizient" von 0.2757, der ebenfalls mit 99.9995 % hoch signifikant ist.

Almo liefert dann noch die "partielle Korrelation" der einzelnen ursächlichen Variablen hinsichtlich der Zieldimension - und bringt dann noch die Mitteilung:

```
***** MITTEILUNG
die Korrelationskoeffizienten der unabhaengigen quantitativen Variablen
sind vorzeichenlos, da die abhaengige Variable nominal-dichotom ist
```

#### P45.15.1.3.1 Zum Begriff der „partiellen Korrelation“:

Die partielle Korrelation der ursächlichen Variablen  $x_1$  mit der Zielvariablen  $y$  ist in folgender Weise zu verstehen:

Die anderen ursächlichen Variablen  $x_2, x_3, x_4, \dots$  korrelieren nicht nur mit der Zielvariablen, sondern auch mit  $x_1$ . Wird der Einfluß der anderen ursächlichen Variablen (rechnerisch) eliminiert, dann bleibt die „direkte“ Korrelation zwischen  $x_1$  und der Zielvariablen zurück. Man spricht dann von der „partiellen“ Korrelation und man spricht davon, daß die anderen ursächlichen Variablen „auspartielliert“ sind.

Als Symbol für diese partielle Korrelation schreibt man

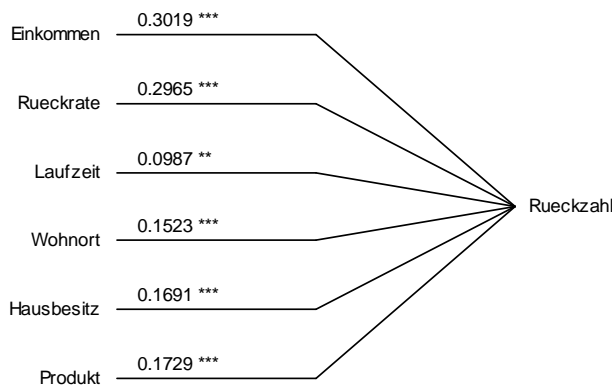
$$r_{x_1 y \cdot x_2 x_3 x_4}$$

wobei  $x_1$  die betrachtete ursächliche Variable ist,  $y$  die Zielvariable und die hinter dem Punkt stehenden Variablen  $x_2$ ,  $x_3$ ,  $x_4$  etc. die „auspartiellierten“ anderen ursächlichen Variablen sind.

Das Einkommen beispielsweise korreliert mit der „Rückzahlung“ mit 0.3020. Dies ist die „direkte“ Korrelation zwischen den beiden Variablen. Die anderen ursächlichen Variablen sind „auspartielliert“. Also teilt dann noch mit, daß dieser Korrelationskoeffizient vorzeichenlos ist. D.h. wir wissen zwar, daß das Einkommen die Rückzahlung mit mittlerer Stärke determiniert, aber nicht in welche Richtung. Das erfahren wir erst, wenn in Eingabe-Box 20 auf „1“ oder „0“ eingestellt wird. Immerhin teilt uns dieser Output mit, welche Variable nicht signifikant wirken, so daß wir ohne diese Variable eine 2. Analyse mit „1“ oder „0“ in der Eingabe-Box 20 rechnen können.

Dieses Ergebnis wird von Almo noch in Form eines Flußdiagramms ausgegeben:

Partielle  
Korrelationskoeffizienten



Auf die Linien dieses Flußdiagramms sind die partiellen Korrelationskoeffizienten geschrieben. 1 Stern hinter dem Zahlenwert bedeutet „mit 95% signifikant“, 2 Sterne „mit 99%“ und 3 Sterne „mit 99,9% signifikant“.

**P45.15.1.4 Ausgabe bei etwas verkürzten Ergebnissen**

Wird in Eingabe-Box 20 eine „1“ eingesetzt, dann entsteht folgende etwas verkürzte Ausgabe.

Prozentwerte der unabhaengigen nominalen Variablen (zeilenweise auf 100 % addiert)  
je Auspraegung der abhaengigen nominalen Variablen

		Rueckzahl nein V10-1	Rueckzahl ja V10-2
Wohnort	Stadt V1-1	36.95	63.05
Wohnort	Land V1-2	24.28	75.72
Hausbesitz	kein Hau V6-1	31.08	68.92
Hausbesitz	hat Haus V6-2	17.74	82.26
Produkt	Kleidung V9-1	39.71	60.29
Produkt	Möbel V9-2	30.23	69.77
Produkt	Technik V9-3	21.52	78.48

**\*\*\*\*\*Erläuterung:**

Von den 100 % Personen, die in der Stadt wohnen, haben 36.95 % ihren Kredit nicht zurückgezahlt; 63.05 % haben zurückgezahlt. Im Unterschied dazu haben von den 100 % Personen, die auf dem Land wohnen nur 24.28 % nicht zurückgezahlt; 75.72 % haben zurückgezahlt.

Entsprechend sind auch die nachfolgenden Zeilen dieser Tabelle zu interpretieren.

Mittelwerte der unabhaengigen quantitativen Variablen  
je Auspraegung der abhaengigen nominalen Variablen

		Rueckzahl nein V10-1	Rueckzahl ja V10-2
Einkommen	V4	1.54	2.85
Rueckrate	V7	4376.62	3092.54
Laufzeit	V8	15.51	14.47

**\*\*\*\*\*Erläuterung:**

Personen, die ihren Kredit nicht zurückgezahlt haben, verfügen im Durchschnitt über ein Einkommen von 1.54 (mal 10 000). Hingegen verfügen Personen, die ihren Kredit zurückgezahlt haben, über ein Durchschnitts-Einkommen von 2.85 (mal 10 000), also über beinahe doppelt soviel.

BEACHTTE: Wir haben die Variable "Einkommen" in der Umkodierungsbox mit 10 000 dividiert.

Entsprechend sind auch die nachfolgenden Zeilen dieser Tabelle zu interpretieren.

Haeufigkeiten je Auspraegung der nominalen Variablen

```
-----
```

V1 Wohnort	
V1-1 Stadt	341
V1-2 Land	659
V6 Hausbesitz	
V6-1 kein Haus	814
V6-2 hat Haus	186
V9 Produkt	
V9-1 Kleidung	204
V9-2 Möbel	387
V9-3 Technik	409
V10 Rueckzahl	
V10-1 nein	286
V10-2 ja	714

**\*\*\*\*\*Erläuterung:**

Von den 1000 Personen wohnen 341 in der Stadt und 659 auf dem Land  
Entsprechend sind auch die nachfolgenden Zeilen dieser Tabelle zu interpretieren.

Zahl der insgesamt eingelesenen Einheiten 1000  
Zahl der in die Analyse einbezogenen Einheiten 1000

**\*\*\*\*\*Erläuterung:**

Die beiden Zahlen können auseinandergehen, z.B. wenn eine eingelesene Person in allen Analysevariablen keinen Wert besitzt

Alle im folgenden angegebenen Streuungen und erklarte Streuungen sind  
Abweichungsquadratsummen  
=====

Koeffizienten fuer quantitat./ordinale Variable aus univariater Analyse

hinsichtlich der abhaeng. Var. V10-1 Rueckzahl: nein

Variable	Regr. koeff.	part. Korrel.	Signifikanz p	(1-p)100
V4 Einkommen	-0.0984	-0.302	0.000	100.00
V7 Rueckrate	0.0001	0.296	0.000	100.00
V8 Laufzeit	-0.0073	-0.099	0.002	99.80

hinsichtlich der abhaeng. Var. V10-2 Rueckzahl: ja

V4 Einkommen	0.0984	0.302	0.000	100.00
V7 Rueckrate	-0.0001	-0.296	0.000	100.00
V8 Laufzeit	0.0073	0.099	0.002	99.80

**\*\*\*\*\* Erläuterung:**

Den partiellen Korrelationskoeffizient haben wir bereits in P45.15.1.3.1 erläutert.  
Betrachten wir nun den Regressionskoeffizienten.

Die Regressionskoeffizienten können verwendet werden, um eine Prognose abzugeben. Für die Zielvariable „Rückzahlung: nein“ kann folgende Gleichung geschrieben werden:

$$p = -0.0984 * \text{Einkommen} + 0.0001 * \text{Rueckrate} - 0.0073 * \text{Laufzeit} + \text{Effekte} + \text{Konstante}$$

$p$  = Wahrscheinlichkeit, dass Rückzahlung nicht erfolgt

Effekte = dies sind die Effekte der ursächlichen nominalen Variablen.

Wir werden weiter unten darauf zurückkommen

Konstante= dies ist eine rein rechnerische Größe, die inhaltlich kaum interpretierbar ist. Sie beträgt in unserem Falle 0.2646. Dieser Wert wird von Almo nur bei der „vollständigen Ausgabe“ (mit „0“ in Eingabe-Box 20) ausgegeben.

Betrachten wir die 2. Person aus der Datei "DatMin.fre". Ihre Werte in den Analyse-Variablen sind folgende:

Wohnort	Einkommen	Hausbesitz	Rueckrate	Laufzeit	Produkt	Rückzahl
2	34770	1	4236	14	2	2

Wohnort : 1=Stadt, 2=Land;  
 Hausbesitz: 1=kein Haus, 2=hat Haus;  
 Produkt : 1=Kleidung, 2=Möbel, 3=Technik;  
 Rückzahl : 1=nein, 2=ja;

Beachte:

Das Einkommen haben wir in der Umkodierungsbox mit 10000 dividiert, so daß aus 34770 der Wert 3.477 wird.

Wenn wir diese Zahlenwerte in obige Gleichung einsetzen, dann erhalten wir

$$p = -0.0984 * 3.477 + 0.0001 * 4236 - 0.0073 * 14 + \text{Effekte} + 0.2646$$

Für die Effekte, wir werden das weiter unten ausrechnen, erhalten wir 0.0219. Damit ergibt sich

$$p = 0.28$$

Die Wahrscheinlichkeit, dass die 2. Person aus unserer Datei ihren Kredit nicht zurückzahlt ist also 28 % und umgekehrt 72%, daß sie ihn zurückbezahlt. Und tatsächlich hat diese Person ihren Kredit zurückbezahlt.

Der Regressionskoeffizient einer ursächlichen Variablen bestimmt also den Beitrag, den diese zur Wahrscheinlichkeit leistet, dass die betrachtete Ausprägung der nominalen Zielvariablen realisiert wird.

Betrachten wir die Laufzeit. Ihr Regressionskoeffizient ist -0.0073. Das bedeutet: Nimmt die Laufzeit um 1 Einheit (1 Monat) zu, dann verringert sich die Wahrscheinlichkeit, dass Kleidung gekauft wird um 0.73%.

Die Regressionskoeffizienten in der 1. Spalte sind kaum miteinander zu vergleichen. In ihrer Größe sind sie auch vom "Wertebereich" abhängig, den die ursächliche

Variable einnimmt. So bewegt sich die „Rückrate“ zwischen 6 und 6830, die Laufzeit aber nur zwischen 1 und 29.

Vergleichbar sind nun jedoch die "partiellen Korrelationskoeffizienten". Den höchsten mit 0.302 besitzt das Einkommen. Es ist damit die stärkste ursächliche Variable. Wir bezeichnen den Korrelationskoeffizienten als einen "partiellen". Was damit gemeint ist haben wir in Abschnitt P45.15.1.3 erläutert.

Koeffizienten der Dummies  
hinsichtlich der abh. Var. V10-1 Rueckzahl: nein

**\*\*\*\*\*Erläuterung:**

Almo gibt zuerst die Effekte der nominalen Variablen hinsichtlich der Zielvariablen „Rückzahlung: nein“ aus und erst dann die für „Rückzahlung: ja“. Wir wollen (um die Ausgabe übersichtlich zu halten) nur letztere betrachten. Erstere ergeben sich durch einfache Vorzeichenumkehr. Die partiellen Korrelationen, Signifikanzen und Kontraste sind gleich. Dies gilt nicht immer.

Koeffizienten der Dummies  
hinsichtlich der abh. Var. V10-2 Rueckzahl: ja

Effekte von A Wohnort

	Effekte	partielle	Signifikanz	
		Korrelat.	p	(1-p)100
A1 Stadt	-0.0611	-0.1523	0.0000	100.00%
A2 Land	0.0611	0.1523	0.0000	100.00%

Paarweise Vergleiche (Kontraste) von A Wohnort

	Differenz	Signifikanz	
		p	(1-p)100
A1 - A2	-0.1223	0.0000	100.00%

=====

Effekte von B Hausbesitz

	Effekte	partielle	Signifikanz	
		Korrelat.	p	(1-p)100
B1 kein Ha	-0.0831	-0.1691	0.0000	100.00%
B2 hat Hau	0.0831	0.1691	0.0000	100.00%

Paarweise Vergleiche (Kontraste) von B Hausbesitz

	Differenz	Signifikanz	
		p	(1-p)100
B1 - B2	-0.1662	0.0000	100.00%

=====

Effekte von C Produkt

	Effekte	partielle	Signifikanz	
		Korrelat.	p	(1-p)100
C1 Kleidun	-0.0874	-0.1391	0.0000	100.00%
C2 Möbel	0.0001	0.0002	0.9985	0.15%

C3 Technik 0.0872 0.1653 0.0000 100.00%

Paarweise Vergleiche (Kontraste) von C Produkt

	Differenz	Signifikanz p	(1-p)100
C1 - C2	-0.0875	0.0075	99.25%
C1 - C3	-0.1746	0.0000	100.00%
C2 - C3	-0.0871	0.0013	99.87%

**\*\*\*\*\* Erläuterung:**

**Effekte** sind die Regressionskoeffizienten der Dummy-Variablen der nominalen Variablen. Wir haben oben ausgeführt, dass die Regressionskoeffizienten der quantitativen ursächlichen Variablen verwendet werden können, um eine Prognose abzugeben. Wir haben folgende Gleichung geschrieben:

$$p = -0.0984 * \text{Einkommen} + 0.0001 * \text{Rueckrate} - 0.0073 * \text{Laufzeit} + \text{Effekte} + \text{Konstante}$$

Wir können jetzt die mit "Effekte" bezeichnete Größe auf der rechten Gleichungsseite bestimmen. Wir tun dies zuerst für „Rückzahlung: ja“. Es gilt

$$\begin{aligned} \text{Effekte} = & 0.0611 * \text{Stadt} + 0.0611 * \text{Land} \\ & 0.0831 * \text{keinH} + 0.0831 * \text{hatH} \\ & 0.0874 * \text{Kleidung} + 0.0001 * \text{Möbel} + 0.0872 * \text{Technik} \end{aligned}$$

In dieser Gleichung sind die Bezeichnungen Stadt, Land, keinH, hatH, Kleidung, Möbel, Technik die 0-1 kodierte Dymmy-Variable der 3 ursächlichen nominalen Variablen Wohnort, Hausbesitz, Produkt.

Für die 2. Person in unserer Datei mit folgenden Merkmalen: Land, keinHaus, Möbel lautet die Gleichung:

$$\begin{aligned} \text{Effekte} = & - 0.0611 * 0 + 0.0611 * 1 \\ & - 0.0831 * 1 + 0.0831 * 0 \\ & - 0.0874 * 0 + 0.0001 * 1 + 0.0872 * 0 \\ = & - 0.0219 \end{aligned}$$

Für „Rückzahlung: nein“ ergibt sich dann einfach der umgedrehte Wert

$$\text{Effekte} = 0.0219$$

**Paarweise Vergleiche** oder **Kontraste** sind die Differenzen zwischen den Effekten einer nominalen ursächlichen Variable. Für uns bedeutsam ist, ob diese paarweisen Vergleiche signifikant sind. Betrachten wir die Variable des Produkts. Die Käufer von Kleidung, Möbel und Technik unterscheiden sich hoch signifikant voneinander in ihrer Rückzahlungswahrscheinlichkeit.

Zusammenfassung

Streuungsquelle	Streuung	Korrel Koeff.	F-Wert	df	Signifikanz p	(1-p)100
Gesamtstreuung	204.2040					
Fehlerstreuung	141.0617			992		
alle unabh. Var. zusammen	63.1423	0.5561	63.4344	7	0.0000	99.9995
quant./ordin. Var. zusammen	51.9529	0.5188	121.7842	3	0.0000	99.9995
nominale Variable zusammen	11.6078	0.2757	20.4076	4	0.0000	99.9995
V4 Einkommen	14.1521	0.3020	99.5229	1	0.0000	99.9995
V7 Rueckrate	13.5952	0.2965	95.6066	1	0.0000	99.9995
V8 Laufzeit	1.3897	0.0988	9.7729	1	0.0020	99.8023

V1 Wohnort	3.3506	0.1523	23.5624	1	0.0000	99.9995
V6 Hausbesitz	4.1504	0.1691	29.1870	1	0.0000	99.9995
V9 Produkt	4.3452	0.1729	15.2786	2	0.0000	99.9984

**\*\*\*\*\*Erläuterung:**

In dieser Zusammenfassung der Ergebnisse werden, wie in P45.15.1.3 bereits beschrieben, die Korrelationskoeffizienten ausgegeben. Daneben werden nun auch die Streuungen mitgeteilt.

**Zum Begriff der Streuung**

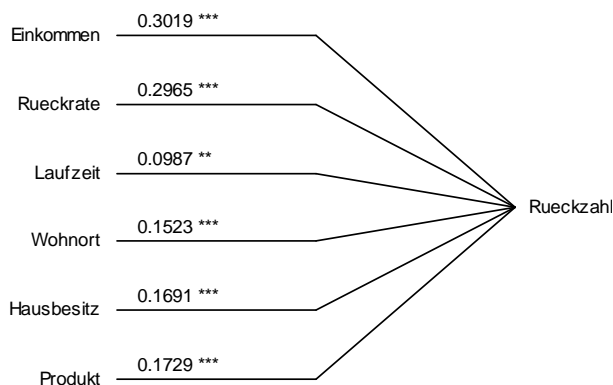
Die Gesamtstreuung in der Zielvariablen „Rückzahlung“ beträgt 204.2040. Davon werden durch alle ursächlichen Variablen 63.1423 erklärt. Das sind 31.2%. Die verbleibende, durch die ursächlichen Variablen nicht erklärte „Fehlerstreuung“ beträgt dann 141.0617. Die durch die quantitativen ursächlichen Variablen zusammen erklärte Streuung ist 51.9529, die durch die nominalen ursächlichen Variablen 11.6078. Die beiden Streuungen ergeben nicht genau 63.1423 (die durch alle ursächlichen Variablen erklärte Streuung), da die quantitativen einerseits und die nominalen Variablen andererseits etwas miteinander korrelieren.

Almo gibt noch die durch die einzelnen ursächlichen Variablen erklärten Streuungen aus. Auch diese addieren sich nicht zu 63.1423, da sie untereinander korreliert sind. Den höchsten Streuungsanteil mit 14.1521 erklärt das Einkommen. Die erklärten Streuungen, wie auch die Korrelationskoeffizienten, können miteinander verglichen werden. Dadurch wird es möglich, die Wirkungsstärke der ursächlichen Variablen vergleichend zu beurteilen. Der (partielle) Korrelationskoeffizient  $r_i$  der Variablen  $i$  kann aus der erklärten Streuung gemäß folgender Formel errechnet werden.

$$r_i = \sqrt{\frac{\text{erkl. Streuung von } i}{\text{erkl. Streuung von } i + \text{Fehlerstreuung}}}$$

Almo liefert folgendes Flußdiagramm der partiellen Korrelationskoeffizienten

Partielle  
Korrelationskoeffizienten



Zusammenfassung: Effekte und Regressionskoeffizienten  
 und ihre Signifikanzen  
 hinsichtlich der abhaengigen Variablen  
 Rueckzahl: nein

	Effekte Regress.koeff	Signifikanz (1-p)*100
	-----	
A1 Stadt	0.061133	99.995000
A2 Land	-0.061133	99.995000
B1 kein Haus	0.083115	99.995000
B2 hat Haus	-0.083115	99.995000
C1 Kleidung	0.087351	99.995000
C2 Möbel	-0.000120	0.148511
C3 Technik	-0.087232	99.995000
Einkommen	-0.098402	99.999500
Rueckrate	0.000103	99.999500
Laufzeit	-0.007303	99.802273

**\*\*\*\*\*Erläuterung:**

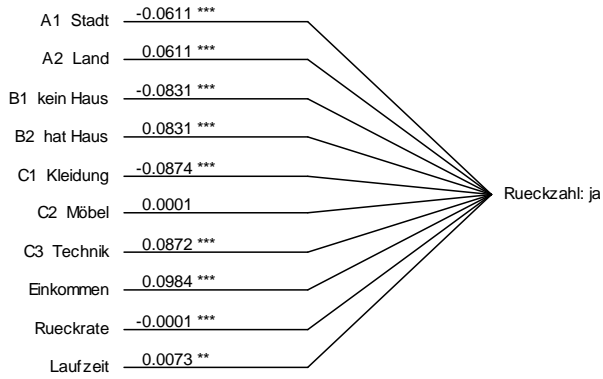
Die Effekte und Regressionskoeffizienten können wir verwenden, um für jede Person die Wahrscheinlichkeit zu berechnen, daß sie den Kredit nicht zurückzahlen. Wir werden das in Abschnitt P45.17 ausführlich darstellen.

Zusammenfassung: Effekte und Regressionskoeffizienten  
 und ihre Signifikanzen  
 hinsichtlich der abhaengigen Variablen  
 Rueckzahl: ja

	Effekte Regress.koeff	Signifikanz (1-p)*100
	-----	
A1 Stadt	-0.061133	99.995000
A2 Land	0.061133	99.995000
B1 kein Haus	-0.083115	99.995000
B2 hat Haus	0.083115	99.995000
C1 Kleidung	-0.087351	99.995000
C2 Möbel	0.000120	0.148511
C3 Technik	0.087232	99.995000
Einkommen	0.098402	99.999500
Rueckrate	-0.000103	99.999500
Laufzeit	0.007303	99.802273

Almo liefert folgendes Flussdiagramm der Effekte und Regressionskoeffizienten

Effekte und Regressionskoeffizienten  
 A Wohnort: A1=Stadt A2=Land  
 B Hausbesitz: B1=kein Haus B2=hat Haus  
 C Produkt: C1=Kleidung C2=Möbel C3=Technik



Almo zeichnet nun noch eine Reihe von linearen Funktionen. Wir zeigen hier beispielhaft nur eine. In der Almo-Ergebnisliste ist folgendes zu erkennen:

**Lineare Funktion fuer**  
 abhaengige Variable: U10 Rueckzahl: 1.Auspraegung: nein  
 unabhagige Variable: U4 Einkommen

Hilfe

Lineare Funktion  
 Grafik 04

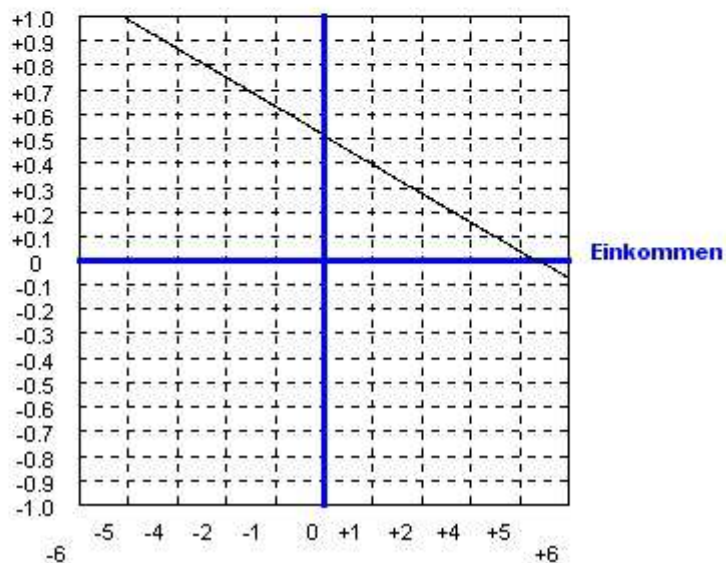
Hilfe  
 Grafik  
 1083918360

Grafik erzeugen  
 und bearbeiten

Grafik loeschen

Lineare Funktion  
 $Y = -0.098402 * X + 0.51382$

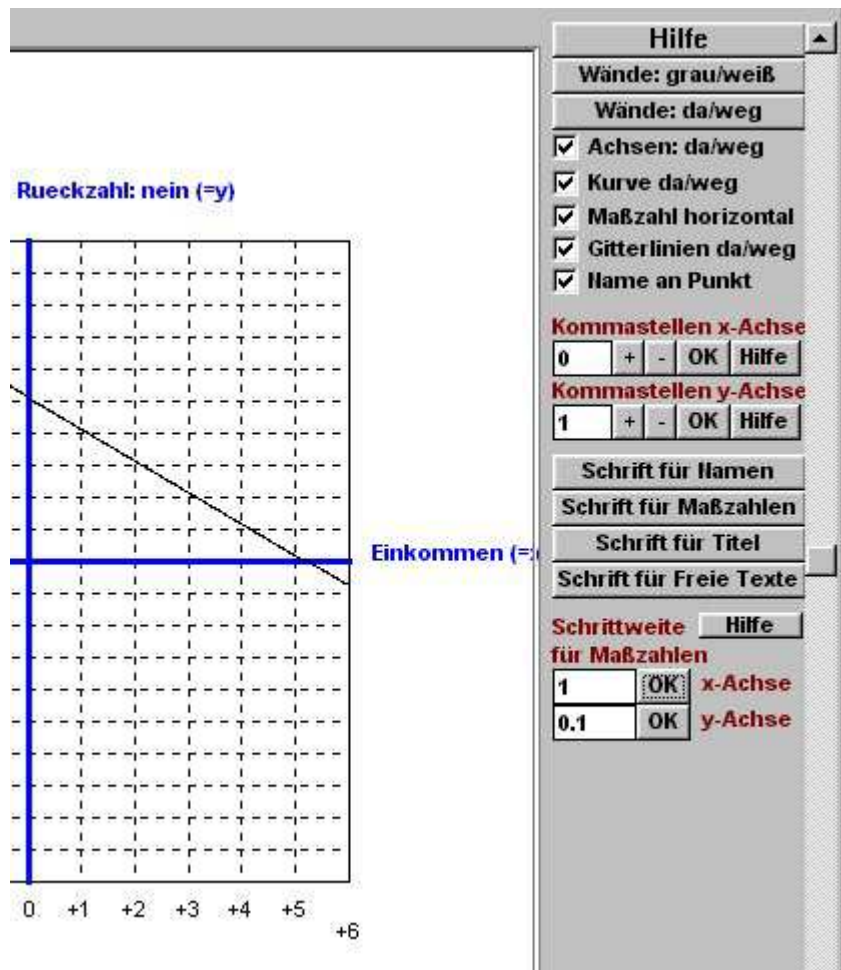
Wahrscheinlichkeit von Rueckzahl: nein



Die Zielvariable ist "Wahrscheinlichkeit von Rueckzahl: Nein". Sie bildet die y-Achse. Die ursächliche quantitative Variable ist "Einkommen". Sie wird an der x-Achse abgebildet.

Zu beachten ist, daß das Einkommen in der Umkodierungsbox mit 10000 dividiert wurde, so daß an der x-Achse jetzt die Werte 1 bis 6 stehen.

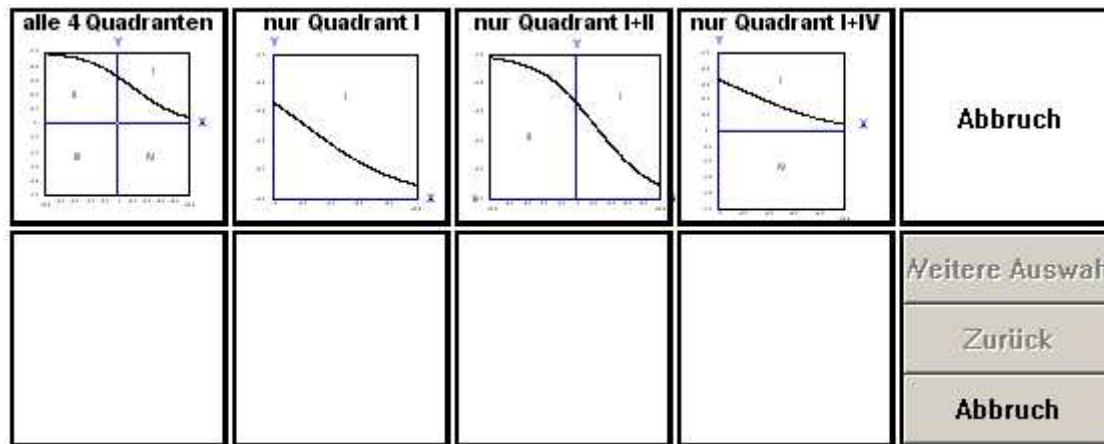
Wir haben diese Grafik im Almo-Grafik-Fenster etwas "verschönert". Wir zeigen einen Ausschnitt von der rechten Hälfte des Grafikfensters.



Wir haben folgende Aktionen vorgenommen:

1. Zuerst haben wir auf den Knopf "Wände: grau / weiß" geklickt. Dadurch wurde der Hintergrund weiß.
2. Die "Kommastellen an der x-Achse" wurden auf 0 gesetzt. Dadurch werden an die x-Achse Ganzzahlwerte geschrieben. Für die y-Achse wurde 1 Kommastelle eingesetzt.
3. Bei der "Schrittweite für Maßzahlen" haben wir für die x-Achse "1" eingesetzt. An der x-Achse stehen dann die Ziffern 1, 2, .... 6. Für die y-Achse wurde 0.1 eingesetzt.
4. Die Checkbox "Maßzahl horizontal" wurde selektiert. Dadurch werden alle Maßzahlen horizontal geschrieben.

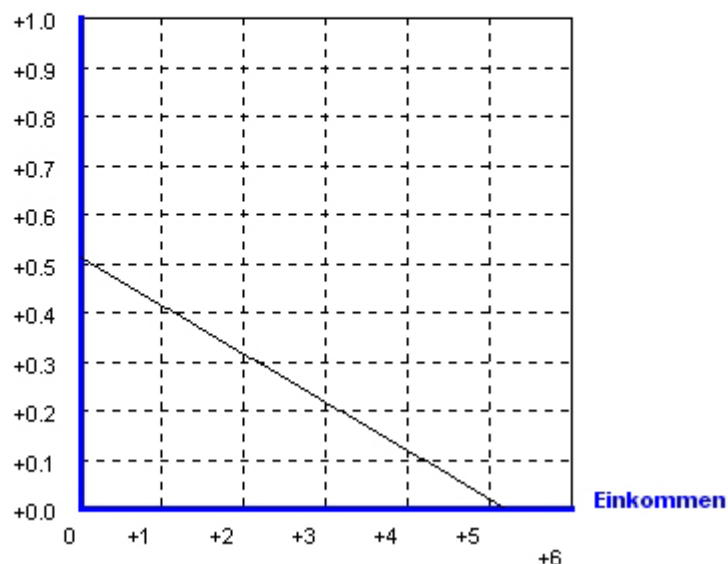
Der jeweils negative Ast der x- und y-Achse ist natürlich irrelevant. Wenn Sie im Grafik-Editor auf der linken Seite auf den Knopf „Diverse Positionen“ klicken, dann bietet Ihnen Almo folgende Auswahl an:



Sie können sich nur den positiven Ast der Kurve zeigen lassen. Klicken Sie auf das 2. kleine Fenster. Almo beschränkt dann die Grafik auf den 1. Quadranten des Koordinatensystems. Sie sehen dann folgende Grafik:

Lineare Funktion  
 $Y = -0.098402 * X + 0.51382$

Wahrscheinlichkeit von Rückzahl: nein



Wir erkennen, daß beispielsweise bei einem Einkommen von 30000 (3 Einheiten) die Wahrscheinlichkeit, daß der Kredit nicht zurück bezahlt wird, etwas über 0.2 liegt. Die Gerade fällt von links nach rechts. Die Wahrscheinlichkeit, daß der Kredit nicht zurückbezahlt wird, fällt also mit steigendem Einkommen. Wir erkennen weiters, dass die Gerade bei einem Einkommen von ca. 5.5 Einheiten die x-Achse schneidet, also negative Wahrscheinlichkeiten prognostiziert. Das ist eines der Gebrechen des Allgemeinen Linearen Modells im Falle, daß die Zielvariable dichotom ist. Wir können interpretieren:

- a. Der Anwendungsbereich unseres Modells ist auf Einkommen von 0 bis 5.5 beschränkt  
 b. oder: Wahrscheinlichkeiten unter 0 werden als 0 verstanden und über 1 als 1.

Almo zeichnet nun je eine lineare Funktion für alle ursächlichen Variablen, also noch für "Rückrate" und "Laufzeit" hinsichtlich der Zielvariablen "Wahrscheinlichkeit von Rueckzahl: Nein".

Die jeweils anderen ursächlichen quantitativen Variablen werden dabei auf ihren Mittelwert gesetzt. Auch die Dummies der ursächlichen nominalen Variablen werden auf ihren Mittelwert gesetzt. Dieser entspricht dem Anteilswert der Ausprägungen. Wurde mit dem Verfahren der "weighted squares of means" gerechnet, dann sind die Anteile der Ausprägungen gleich groß (siehe dazu Handbuch zu P20, Abschnitt P20.7.3).

Die Gleichung für unser Beispiel ist nachfolgend in (1) angegeben, die Gleichung, die Almo zeichnet, in (2)

Gleichung (1)	Gleichung (2)		
$p = \beta_1 * E$	$p = \beta_1 * E$	Einkommen	Variable an der x-Achse  alle anderen werden auf ihren Mittelwert gesetzt
$+ \beta_2 * R$	$+ \beta_2 * MR$	Rückrate	
$+ \beta_3 * L$	$+ \beta_3 * ML$	Laufzeit	
$+ \beta_4 * Ws$	$+ \beta_4 * Mws$	Wohnort: Stadt	
$+ \beta_5 * Wl$	$+ \beta_5 * MWl$	Wohnort: Land	
$+ \beta_6 * Hk$	$+ \beta_6 * MHk$	Hausbesitz: kein Haus	
$+ \beta_7 * Hh$	$+ \beta_7 * MHh$	Hausbesitz: hat Haus	
$+ \beta_8 * Pk$	$+ \beta_8 * MPk$	Produkt: Kleidung	
$+ \beta_9 * Pm$	$+ \beta_9 * MPm$	Produkt: Möbel	
$+ \beta_{10} * Pt$	$+ \beta_{10} * MPt$	Produkt: Technik	
$+ const$	$+ const$	Konstante	Konstante

- p = "Wahrscheinlichkeit für Rückzahlung:Nein"
- $\beta_1$  =Regressionskoeffizient für Einkommen  
 $\beta_2$  =Regressionskoeffizient für Rückrate  
 $\beta_3$  =Regressionskoeffizient für Laufzeit
- $\beta_4$  =Effekt für Wohnort: Stadt  
 $\beta_5$  =Effekt für Wohnort: Land
- $\beta_6$  =Effekt für Hausbesitz: kein Haus  
 $\beta_7$  =Effekt für Hausbesitz: hat Haus
- $\beta_8$  =Effekt für Produkt: Kleidung  
 $\beta_9$  =Effekt für Produkt: Möbel  
 $\beta_{10}$  =Effekt für Produkt: Technik
- const =Konstante
- MR, ML =Mittelwert aus Rückrate, Laufzeit  
 Mws, MWl =Mittelwert für Wohnort: Stadt bzw. Land  
 MHk, MHh =Mittelwert für Hausbesitz: kein Haus bzw. hat Haus  
 MPk, MPm, MPt =Mittelwert für Produkt: Kleidung bzw. Möbel bzw. Technik

Für die ursächlichen quantitativen Variablen Rückrate und Laufzeit ist in (2) deren Mittelwert eingesetzt worden. Ebenso für die Dummies der unabhängigen nominalen Variablen. Das entspricht der Einsetzung einer "Durchschnittsperson".

Wir können also etwas verkürzt formulieren:

In der Almo-Grafik wird für die "Durchschnittsperson" der lineare Zusammenhang zwischen Einkommen und "Wahrscheinlichkeit für Rückzahlung:Nein" gezeichnet.

Im Titel der Almo-Graphik wird Gleichung (2) angegeben. Dabei wird der Gleichungsteil

$\beta_2 * MR$		
$+ \beta_3 * ML$		
$+ \beta_4 * MWS$		
$+ \beta_5 * MWL$		
$+ \beta_6 * MHk$		die anderen Variablen die auf ihren Mittelwert gesetzt wurden
$+ \beta_7 * MHh$		
$+ \beta_8 * MPk$		
$+ \beta_9 * MPm$		
$+ \beta_{10} * MPt$		

aus obiger Gleichung der Konstanten "const" hinzugefügt

Die Summe von "Effekt mal Anteilswert" der Dummies einer unabhängigen nominalen Variablen ist immer =0.

Es ist also  $\beta_4 * MWS + \beta_5 * MWL = 0$

$$\beta_6 * MHk + \beta_7 * MHh = 0$$

$$\beta_8 * MPk + \beta_9 * MPm + \beta_{10} * MPt = 0$$

Tatsächlich lautet also die Gleichung, die Almo zeichnet:

$$p = \beta_1 * E + \beta_2 * MR + \beta_3 * ML + const$$

So entsteht für "Einkommen" (x-Achse) versus "Wahrscheinlichkeit für Rückzahlung:Nein" (y-Achse) folgende Gleichung

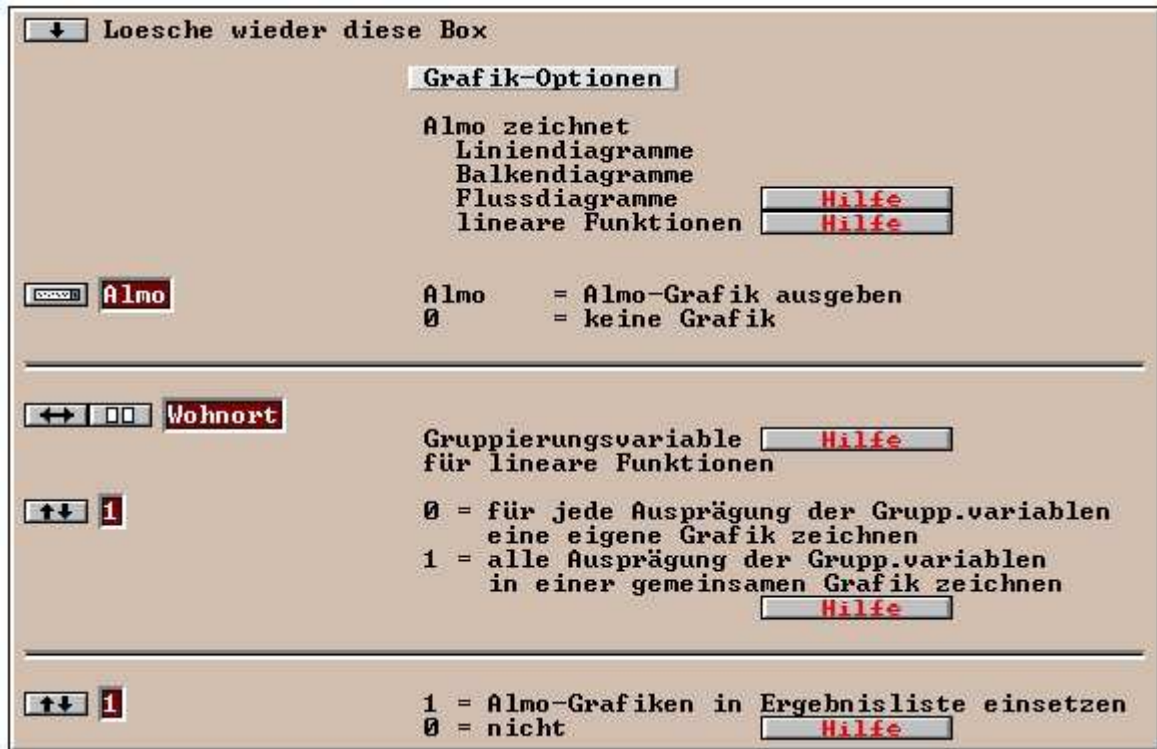
$$p = -0.098402 * E + 0.52936$$

#### P45.15.1.4.1 Gruppierungsvariable

Nun besteht die Möglichkeit eine oder mehrere Gruppierungsvariable anzugeben.

BEACHTTE: Als Gruppierungsvariable können nur Variable verwendet werden, die als ursächliche nominale Variable angegeben wurden.

Es wird beispielsweise der "Wohnort" als Gruppierungsvariable angegeben. In die Eingabe-Box "Grafik-Optionen" wird dann eingetragen



Almo zeichnet dann die linearen Funktionen (so wie oben beschrieben) für die beiden Ausprägungen des Wohnorts. Es werden also folgende Kurven gezeichnet:

1. Einkommen (x-Achse) mit "Wahrscheinlichkeit für Rückzahlung:Nein" (y-Achse) für die Städter und die Landbewohner
2. Rückrate (x-Achse) mit "Wahrscheinlichkeit für Rückzahlung:Nein" (y-Achse) für die Städter und die Landbewohner
3. Laufzeit (x-Achse) mit "Wahrscheinlichkeit für Rückzahlung:Nein" (y-Achse) für die Städter und die Landbewohner

Die jeweils anderen ursächlichen Variablen sind dabei auf ihren Mittelwert gesetzt. Bei (1) wird also

Hausbesitz:keinHaus  
 Hausbesitz:hatHaus

Produkt:Kleidung  
 Produkt:Möbel  
 Produkt:Technik

Rückrate  
 Laufzeit

auf den Mittelwert bzw. Anteilswert gesetzt.

Wie wir oben bereits ausgeführt haben, ist die Summe der Effekte einer ursächlichen nominalen Variablen, wenn der Anteilswert eingesetzt ist, gleich 0.

Wir wollen nur die erste Funktion betrachten

Lineare Funktion fuer  
 abhaengige Variable: U10 Rueckzahl: 1.Auspraegung: nein  
 unabhengige Variable: U4 Einkommen

Gruppierungsvariable:  
 1: gruene Linie: U1 Wohnort 1.Auspraegung: Stadt  
 2: blaue Linie: U1 Wohnort 2.Auspraegung: Land

Lineare Funktion  
 Grafik 04

Hilfe  
 Grafik  
 1083919103

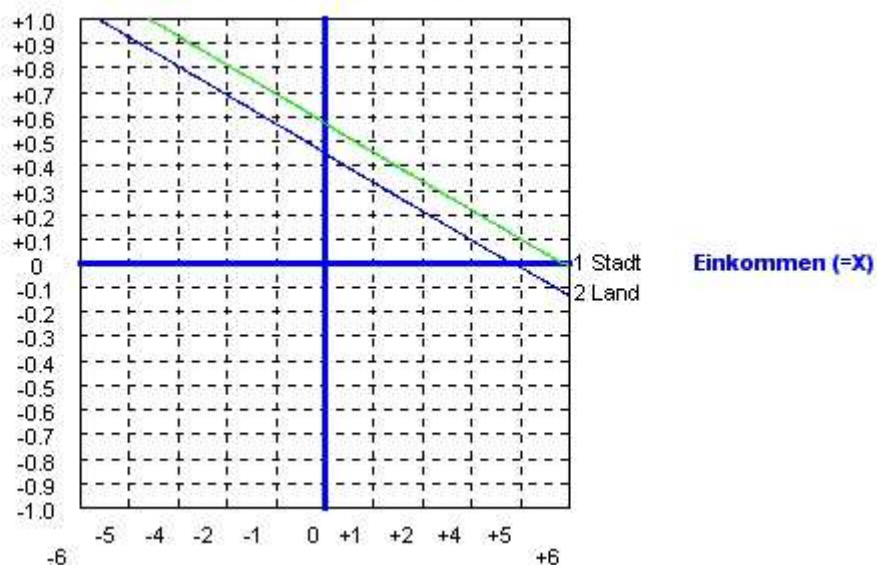
Grafik erzeugen  
 und bearbeiten

Grafik loeschen

Lineare Funktion

1: gruene Linie: Stadt  $Y = -0.098402 * X + 0.57496$   
 2: blaue Linie: Land  $Y = -0.098402 * X + 0.45269$

Wahrscheinlichkeit von Rueckzahl: nein (=Y)



Auf dem Bildschirm ist die obere Gerade grün und die untere Gerade blau dargestellt.

Folgendes ist deutlich zu erkennen:

1. Die Steigung der beiden Geraden ist dieselbe.
2. Die Gerade für die Städter liegt höher. D.h. bei gleichem Einkommen haben die Städter eine durchwegs höhere Wahrscheinlichkeit, ihren Kredit nicht zurückzuzahlen.

Almo zeichnet bei der oberen Geraden folgende Gleichung:

$$(3) \quad p = \beta_1 * E + \beta_2 * MR + \beta_3 * ML + \beta_4 * 1 + \beta_5 * 0 + \text{const}$$

p = "Wahrscheinlichkeit für Rückzahlung:Nein"

$\beta_1$  =Regressionskoeffizient für Einkommen

$\beta_2$  =Regressionskoeffizient für Rückrate

$\beta_3$  =Regressionskoeffizient für Laufzeit

MR, ML =Mittelwert aus Rückrate, Laufzeit

$\beta_4$  =Effekt für Wohnort: Stadt

$\beta_5$  =Effekt für Wohnort: Land

const =Konstante

Bei der oberen Geraden wird die Gruppierungsvariable "Wohnort" mit ihrer Ausprägung "Stadt" verwendet. Der Effekt  $\beta_4$  ("Stadt") wird demzufolge mit 1 und  $\beta_5$  ("Land") mit 0 multipliziert. Wie wir oben bereits ausgeführt haben ist die Summe der Effekte der anderen ursächlichen nominalen Variablen (Hausbesitz,Produkt), wenn der Anteilswert eingesetzt ist, gleich 0, so daß wir sie nicht in die Gleichung einschreiben.

Bei der unteren Geraden zeichnet Almo folgende Gleichung

$$(4) \quad p = \beta_1 * E + \beta_2 * MR + \beta_3 * ML + \beta_4 * 0 + \beta_5 * 1 + \text{const}$$

Die Gruppierungsvariable ist jetzt "Wohnort" mit ihrer Ausprägung "Land". Der Effekt "Stadt" ist deswegen mit 0 und "Land" mit 1 multipliziert.

Würden sich nicht die ursächlichen Variable "Rückrate" und "Laufzeit" im Modell befinden, dann würden die beiden Funktionen sehr einfach lauten:

$$(3a) \quad A = \beta_1 * E + \beta_4 * 1 + \beta_5 * 0 + \text{const}$$

$$(4a) \quad A = \beta_1 * E + \beta_4 * 0 + \beta_5 * 1 + \text{const}$$

Die Steigung der Geraden ist gleich  $\beta_1$ , dem Regressionskoeffizienten des Einkommens. Zur Konstanten "const" kommt dann noch der Effekt der jeweiligen Ausprägung der Gruppierungsvariablen hinzu.

#### P45.15.1.4.2 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. Betrachten wir ein Beispiel:

In die Eingabe-Box "Grafik-Optionen" schreiben Sie in das Eingabefeld für die Gruppierungsvariable beispielsweise

Wohnort MIT Hausbesitz

BEACHTTE: Es können maximal 4 Variable durch MIT kombiniert werden.



Wir bilden nur den oberen Teil der Grafik-Optionsbox ab.

Almo erzeugt dann folgende Kombinationen in folgender Reihenfolge

Stadt mit hat kein Haus  
 Stadt mit hat Haus  
 Land mit hat kein Haus  
 Land mit hat Haus

Die jeweils hintere Variable "läuft" über ihre Ausprägungen. Für jede Kombination wird eine Funktionsgrafik gezeichnet.

#### P45.15.1.4.3 Mehrere Gruppierungsvariable

Es können mehrere Gruppierungsvariable (durch Beistrich getrennt) angegeben werden. Beispiel:

Wohnort, Hausbesitz

Almo zeichnet dann für die ursächlichen quantitativen Variablen (Einkommen, Rückrate, Laufzeit) je 4 Kurven, eine für die Städter, eine für die Landbewohner, eine für die Hausbesitzer und eine für die Nicht-Hausbesitzer.

Mehrere einzelne Gruppierungsvariable und mehrere durch MIT kombinierte Gruppierungsvariable können angegeben werden.

Beispiel:

Produkt, Wohnort mit Hausbesitz

#### P45.15.1.5 Wertemuster

Wenn die Options-Box "Wertemuster" aktiviert wurde, dann gibt Almo noch die Prognosewerte für die Wertemuster aus.

Prognosewerte fuer Wertemuster fuer abhaengige Variable V10 Rueckzahl

-----  
Wertemuster 1

Werte der unabhaengigen Variablen

V1 Wohnort	1 Stadt	1	
V1 Wohnort	2 Land	0	
V6 Hausbesitz	1 kein Haus	1	
V6 Hausbesitz	2 hat Haus	0	
V9 Produkt	1 Kleidung	0.204	(=Anteilswert)
V9 Produkt	2 Möbel	0.387	(=Anteilswert)
V9 Produkt	3 Technik	0.409	(=Anteilswert)
V4 Einkommen		4	
V7 Rueckrate		3459.78	(=Mittelwert)
V8 Laufzeit		14.771	(=Mittelwert)

Prognosewerte fuer abhaengige Variable

p-Wert fuer V10 Rueckzahl	1 nein	0.24724
p-Wert fuer V10 Rueckzahl	2 ja	0.75276

-----  
Wertemuster 2

Werte der unabhaengigen Variablen

V1 Wohnort	1 Stadt	0	
V1 Wohnort	2 Land	1	
V6 Hausbesitz	1 kein Haus	1	
V6 Hausbesitz	2 hat Haus	0	
V9 Produkt	1 Kleidung	0.204	(=Anteilswert)
V9 Produkt	2 Möbel	0.387	(=Anteilswert)
V9 Produkt	3 Technik	0.409	(=Anteilswert)
V4 Einkommen		3	
V7 Rueckrate		3459.78	(=Mittelwert)
V8 Laufzeit		14.771	(=Mittelwert)

Prognosewerte fuer abhaengige Variable

p-Wert fuer V10 Rueckzahl	1 nein	0.223376
p-Wert fuer V10 Rueckzahl	2 ja	0.776624

-----

**\*\*\*\*\*Erläuterung:**

Almo gibt zuerst die Werte der ursächlichen Variablen aus, die der Benutzer in der Eingabe-Box „Wertemuster“ eingesetzt hat. Für die Variablen, für die der Benutzer keine Werte eingesetzt hat, verwendet Almo die Mittelwerte bzw. die Anteilswerte. Dann gibt Almo die errechneten Prognosewerte für die Zielvariable aus.

**P45.15.1.6 Volle Ausgabe**

Die volle Ausgabe entsteht, wenn in der Eingabe-Box 20 eine „0“ eingesetzt wird. Die volle Ausgabe ist außerordentlich umfangreich. Wir geben hier nur jene Teile an, die gegenüber der „etwas verkürzten Ausgabe“ neu sind und die für den Data-Mining-Prozeß noch interessant sind.

Zellenmittelwerte der  
unabhaengigen quantitativen/ordinalen Variablen

Wohnort	Hausbesi	Produkt	Einkommen	Rueckrate	Laufzeit
Stadt	kein Hau	Kleidung	2.14	3636.16	16.39
		Möbel	2.86	3276.52	14.37
		Technik	2.51	3534.04	14.23
	hat Haus	Kleidung	2.11	3045.44	13.39
		Möbel	1.74	4355.03	15.94
		Technik	2.62	2604.72	12.44
Land	kein Hau	Kleidung	2.67	3387.92	14.48
		Möbel	2.46	3488.44	14.85
		Technik	2.44	3447.06	15.27
	hat Haus	Kleidung	2.25	3734.16	14.84
		Möbel	2.40	3479.40	14.78
		Technik	2.49	3278.89	13.85
Gesamtmittel			2.47	3459.78	14.77

**\*\*\*\*\*Erläuterung:**

Betrachten wir die Zahl 2.14 im linken oberen Eck der Tabelle. Personen, die in der Stadt wohnen, kein Haus besitzen und Kleidung gekauft haben, besitzen ein durchschnittliches Einkommen von 2.14 Einheiten.

Beachte: Wir haben das Einkommen in der Eingabe-Box "Kein-Wert-Angabe und Umkodierungen" mit 10 000 dividiert.

Haeufigkeitstabelle:

Wohnort	Hausbesi	Produkt	Rueckzahl		Summe
			nein	ja	
Stadt	kein Hau	Kleidung	27	29	56
		Möbel	35	57	92
		Technik	44	81	125
	hat Haus	Kleidung	6	12	18
		Möbel	13	19	32
		Technik	1	17	18
Land	kein Hau	Kleidung	42	69	111
		Möbel	64	154	218
		Technik	41	171	212
	hat Haus	Kleidung	6	13	19
		Möbel	5	40	45
		Technik	2	52	54
Summe			286	714	1000

**\*\*\*\*\*Erläuterung:**

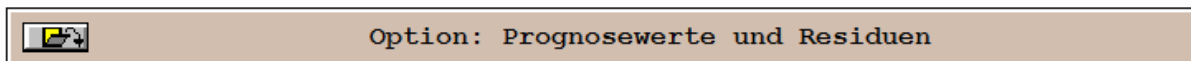
Betrachten wir die erste Zeile der Tabelle. Sie bezieht sich auf die Personen, die in der Stadt wohnen, kein Haus besitzen und Kleidung gekauft haben. Insgesamt sind dies in unserem Datenmaterial 56 Personen. 27 von ihnen haben ihren Kredit nicht zurückgezahlt. 29 haben ihn zurückgezahlt.

***P45.15.1.7 Prognosefähigkeit bzw. Reproduzierbarkeit des Modells***

Wir wollen überprüfen, wie gut wir mit den ursächlichen Variablen und den für sie errechneten Koeffizienten prognostizieren bzw. reproduzieren können, ob eine Person ihren Kredit zurückzahlt oder nicht zurückzahlt. Genauer: Für jede der 1 000 Personen unserer Beispieldaten wird (indem wir die Effekte und Regressionskoeffizienten verwenden) die Wahrscheinlichkeit errechnet, daß sie zurückzahlt bzw. nicht zurückzahlt. Das ist der „Prognosewert“ oder „reproduzierte“ Wert.

Nun wissen wir aber von jeder Person, ob sie zurückgezahlt hat oder nicht. Das ist der „tatsächliche Wert“. So können wir nun auszählen, bei wie vielen Personen unser Modell richtig bzw. falsch prognostiziert hat.

Wir suchen im Programm Prog45mf nach der Eingabe-Box „Option: Prognosewerte und Residuen“.



Nach Klick auf den Knopf wird folgende Eingabe-Box inkludiert.

X Loesche wieder diese Box (dann Voreinstellungen wieder gueltig)

**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 %

---

Schreibe die Prognosewerte und Residuen  
 in eine Datei im Format FREI  
 was wird gespeichert ? --->

Eingabefeld leer = nicht in Datei speichern

*Eingabefeld 1:* Hier geben wir 1 ein. Damit werden für alle 1 000 Personen die „Prognosewerte“ ermittelt und 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 unsere 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:* Die Prognosewerte und Residuen können in eine Datei gespeichert werden. Was wird dabei gespeichert ?

Für die Datei der Prognosewerte und Residuen gelten folgende Bedingungen:

1. Wurden in der Box

Option: Umkodierungen und Kein-Wert-Angaben

Variable umkodiert oder wurden KeinWert-Angaben vorgenommen, dann werden diese für die Berechnung der Prognosewerte und Residuen auch durchgeführt

2. Wurden hingegen in der Box

Option: Ein- und Ausschliessen von Untersuchungseinheiten

Datensätze ausgeschlossen, so werden für diese auch keine Prognosewerte und keine Residuen ermittelt. Wenn der Benutzer z.B. Männer ausschliesst, dann werden nur für die Frauen Prognosewerte und Residuen errechnet

3. Wurde die Box

Option: Ausreisser identifizieren

geöffnet und als Reaktion auf gefundene Ausreisser angegeben:

5 = Ausreisser melden und ganzen Datensatz ausschliessen

so fehlen auch die Datensätze in der Datei der Prognosewerte und Residuen, in denen mindestens ein Ausreisser in den Analysevariablen gefunden wurde.

Prog45mf liefert folgende zusätzliche **Ausgabe** (gekürzt):

Berechnung der Prognosewerte und Residuen

\*\*\*\*\* WARNUNG  
 Es koennen prognostizierte Wahrscheinlichkeiten ausserhalb 0-1 auftreten  
 Als Alternative ist die Logit-Analyse (Prog22m oder Prog45m9) moeglich  
 Bei ihr treten keine Wahrscheinlichkeiten ausserhalb 0-1 auf

Prognosewerte und Residuen fuer Variable V10 Rueckzahl: nein, ja

Die Gruppe mit maximaler Wahrscheinlichkeit ist mit \* markiert

Die tatsaechliche Gruppenzugehoerigkeit wird hinter  
 der Datensatznummer in Klammern angegeben

Datensatz	prognostizierte Wahrscheinlichkeit der Zugehoerigkeit zu Gruppe		Residuen (Differenz) zu Gruppe	
	1	2	1	2
	nein	ja	nein	ja
1 (2)	0.570*	0.430	-0.570	0.570
2 (2)	0.280	0.720*	-0.280	0.280
3 (1)	0.593*	0.407	0.407	-0.407
4 (2)	0.135	0.865*	-0.135	0.135
5 (2)	0.105	0.895*	-0.105	0.105
6 (1)	0.328	0.672*	0.672	-0.672
7 (2)	0.387	0.613*	-0.387	0.387
8 (2)	0.507*	0.493	-0.507	0.507
9 (2)	0.025	0.975*	-0.025	0.025
10 (2)	-0.036	1.036*	0.036	-0.036
.	.	.	.	.
.	.	.	.	.
995 (1)	0.308	0.692*	0.692	-0.692
996 (2)	0.156	0.844*	-0.156	0.156
997 (2)	0.110	0.890*	-0.110	0.110
998 (1)	0.369	0.631*	0.631	-0.631
999 (1)	0.422	0.578*	0.578	-0.578
1000 (1)	0.317	0.683*	0.683	-0.683

\*\*\*\*\* **Erläuterung:**

In der 1. Spalte steht die (fortlaufende) Nummer der jeweiligen Person. Dahinter steht die tatsächliche Zugehörigkeit zur Gruppe der Nicht-Rückzahler (Gruppe 1) bzw. der Rückzahler (Gruppe 2).

Dann kommt die vom Modell prognostizierte Wahrscheinlichkeit, dass die Person der Gruppe 1 bzw. der Gruppe 2 angehört.

Betrachten wir die Person 1. Für sie wird prognostiziert, dass sie mit einer Wahrscheinlichkeit von 0.570 zur Gruppe 1 gehört und mit einer Wahrscheinlichkeit von 0.430 zur Gruppe 2 gehört. Die größere Wahrscheinlichkeit ist die für Gruppe 1. Tatsächlich gehört aber diese Person zur Gruppe 2. Unser Modell hat also bezüglich der Person 1 eine falsche Prognose abgegeben.

Bei Person 10 prognostiziert unser Modell mit -0.036 eine negative Wahrscheinlichkeit und mit 1.036 (für Gruppe 2) eine Wahrscheinlichkeit größer als 1. Eine Wahrscheinlichkeit muß aber zwischen 0 und 1 liegen. Eine mögliche Lösung des Problems bestünde darin, Werte unter 0 auf 0 und Werte über 1 auf 1 zu setzen. Wie auch immer: Almo behält sein Entscheidungskriterium bei, die Gruppenzugehörigkeit zu prognostizieren, die den höchsten Wahrscheinlichkeitswert besitzt. Eine solche Situation ereignet sich, wie uns weiter unten mitgeteilt wird, in 125 Fällen, also in 12.5 % aller Fälle. Das kann bei dem von uns gewählten Ansatz, dem "Allgemeinen Linearen Modell", durchaus geschehen, vor allem dann, wenn die Zielvariable, wie in unserem Beispiel, dichotom ist. Bei polytomen Zielvariablen tritt das sehr viel seltener auf. Da es für die Prognose, welcher Gruppe die Person angehört, nur wichtig ist, welche Wahrscheinlichkeit die größere ist, ist dieser Defekt des "Allgemeinen Linearen Modell" nicht so tragisch. Wir werden anschließend als alternatives Modell die "Logit-Analyse" kennen lernen, bei der dieser Defekt nicht auftritt - dafür jedoch andere Probleme sich bemerkbar machen.

Prognoseerfolg fuer Variable V10 Rueckzahl: nein, ja

		Haeufigkeit in Gruppe		
		tatsaechlich	davon richtig prognostiziert	zufaellig richtig
		-----	-----	-----
Gruppe 1	nein	286	148 (51.7 %)	82 (28.6 %)
Gruppe 2	ja	714	650 (91.0 %)	510 (71.4 %)

In 0 Faellen ist die tatsaechliche Gruppenzugehoerigkeit nicht bekannt (=Kein\_Wert)

Bei 125 Datensaeetzen (=12.5 %) liegt die prognostizierte Wahrscheinlichkeit ausserhalb des zulaessigen Wertebereichs von 0 bis 1

**\*\*\*\*\* Erläuterung:**

286 Personen haben ihren Kredit nicht zurückbezahlt. Für 148 Personen (=51.7 %) hat unser Modell eine richtige Prognose abgegeben. Das ist nun doch etwas enttäuschend. Andererseits haben wir von den 714, die ihren Kredit zurückgezahlt haben 91 % richtig prognostiziert.

Um die Leistung unseres Modells gerecht zu beurteilen, sollte man folgenden Vergleich anstellen: Würden wir zufällig aus den 1000 Personen 286 auswählen, dann wären davon  $286 \cdot 286 / 1000 = 82$  Personen (28.6 %) Nicht-Rückzahler. Das ist in der 3. Spalte obiger Tabelle angegeben. Unser Modell hat aber 148 Personen richtig als Nicht-Rückzahler erkannt. Die Trefferquote unseres Modells ist mit 51.7 % um den Faktor 1.81 besser als der Zufall.

Allgemein gilt: Je disproportionaler eine Verteilung auf 2 (oder mehr) Gruppen ist, umso ungünstiger wird der Prognoseerfolg für die kleinere Gruppe und umso besser für die größere Gruppe.

### ***P45.15.1.8 Gewichtete Kleinste-Quadrate-Schätzung für dichotome Zielvariable***

Ist die abhängige Variable nominal-dichotom, dann besteht modellbedingte Varianzheterogenität. Diese kann durch die Methode der "gewichteten Kleinste-Quadrate" beseitigt werden. Wir haben dies in Abschnitt P45.15.1.0 dargestellt.

Die Theorie dieses Verfahrens (mit einem Rechenbeispiel) ist sehr übersichtlich dargestellt bei Aldrich/Nelson (1984, S. 14ff). Das Beispiel in Aldrich/Nelson ist unter dem Namen „Hetero.Alm“ als Beispielprogramm auch in Almo enthalten.

Wir verwenden im folgenden unser „Rückzahlungs-Beispiel“, das wir allerdings vereinfachen.

Abhängige dichotome Variable ist die Rückzahlung. Ursächliche quantitative Variable ist das Einkommen (dividiert durch 10 000). Ursächliche nominale Variable sind der Wohnort und der Hausbesitz.

Almo rechnet zuerst eine normale Analyse. Dabei werden folgende Effekte für die unabhängigen nominalen Variablen bzw. Regressionskoeffizienten für die unabhängigen quantitativen Variablen ermittelt:

Effekte und Regressionskoeffizienten  
hinsichtlich der abhängigen Variablen Rueckzahl: nein

	Effekte
	Regress.koeff
	-----
A1 Stadt	0.0835
A2 Land	-0.0432
B1 kein Haus	0.0315
B2 hat Haus	-0.1378
Einkommen	-0.1426
Konstante	0.6387

Für jede Untersuchungseinheit wird nun die Wahrscheinlichkeit  $p_1$  der „Rückzahlung: nein“ durch folgende lineare Wahrscheinlichkeitsfunktion prognostiziert:

$$P_1 = 0.0835 * A1 - 0.0432 * A2 + 0.0315 * B1 - 0.1378 * B2 - 0.1462 * \text{Einkommen} + 0.6387$$

Für die 1. Person in unserer Datei, die in der Stadt lebt, kein Haus besitzt und ein Einkommen von 3.2384 Einheiten (= 32 384 Geldeinheiten) hat, ergibt sich:

$$P_1 = 0.0835 * 1 - 0.0432 * 0 + 0.0315 * 1 - 0.1378 * 0 - 0.1462 * 3.2384 + 0.6387 = 0.2919$$

also eine Wahrscheinlichkeit  $p_1$  der Nicht-Rückzahlung von 0.2919.

Wäre dieses  $p_1$  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.

Der Standardfehler für diese Person ist dann

$$s = \sqrt{(p * (1 - p))} = \sqrt{(0.2919 * (1 - 0.2919))} = 0.4546$$

Sämtliche Variablenwerte (inklusive der auf 1 gesetzten Konstantenvariablen) dieser Person werden nun mit s dividiert. Der neue Datensatz wird zwischengespeichert. So wird mit jeder Person verfahren. Die zwischengespeicherten Daten werden dann einer 2. Analyse unterworfen. Diese liefert die Ergebnisse der "gewichteten" Analyse:

Zusammenfassung: Effekte und Regressionskoeffizienten  
und ihre Signifikanzen  
hinsichtlich der abhaengigen Variablen  
Rueckzahl: nein

	Effekte Regress.koeff	Signifikanz (1-p)*100
	-----	
A1 Stadt	0.049709	99.995000
A2 Land	-0.025722	99.995000
B1 kein Haus	0.018361	99.995000
B2 hat Haus	-0.080355	99.995000
Einkommen	-0.077482	99.999500
Konstante	0.412667	99.999500

Beim Vergleich mit den Ergebnissen aus der ungewichteten Analyse erkennt man, daß die Koeffizienten ungefähr halb so groß sind. Das muß aber nicht immer notwendigerweise so sein.

Anmerkungen:

1. Sind die Zusammenhänge zwischen den ursächlichen Variablen und der Zielvariablen schwach, dann können die Ergebnisse aus der gewichteten Analyse teilweise sehr verschieden sein von denen der normalen ungewichteten Analyse.
2. Die Regressionskoeffizienten und Effekte aus der gewichteten Analyse sind "unverzerrt" und "effizient". Ihre Standardfehler sind die kleinst möglichen, also kleiner als die aus der normalen ungewichteten Analyse. Ihre "Prognosefähigkeit" ist jedoch geringer, d.h. sie vermögen die Werte der Untersuchungspersonen in der Stichprobe in der Zielvariablen nicht so gut zu reproduzieren (prognostizieren) wie die aus der normalen ungewichteten Analyse.

### Multiple Korrelation

Die gesamte erklärte Streuung und damit auch die multiple Korrelation, die nach der 2. Analyse ausgegeben werden, beziehen sich auf die veränderte (mit dem individuellen Standardfehler dividierte) Zielvariable. Wir benötigen jedoch diese Koeffizienten für die Original-Zielvariable. Um sie zu erhalten müssen die Prognosewerte und die Residuen errechnet werden. Aus ihnen können wir dann die gesuchten Koeffizienten errechnen.

Der Benutzer muß die Optionsbox "Prognosewerte und Residuen ermitteln" öffnen. Er sieht dann folgendes:

Siehe unsere ausführliche Beschreibung dieser Eingabe-Box in Abschnitt P45.15.1.6.

Das 2. *Eingabefeld* muß leer bleiben (oder man trägt 100 ein). Es müssen alle Untersuchungseinheiten analysiert werden.

Im 1. *Eingabefeld* kann man eine "2" eintragen. Dann werden die Prognosewerte zwar berechnet, aber nicht ausgegeben. Selbstverständlich kann man sich die Prognosewerte auch mit "1" ausgeben lassen.

Almo liefert dann folgende zusätzliche Ausgabe

Mittelwert und Standardabweichung der Residuen  
 fuer Variable V10 Rueckzahl: nein, ja

	Mittelwert	Standardabweichung
Gruppe 1 nein	0.0514963	0.391733
Gruppe 2 ja	-0.0514963	0.391733

**\*\*\*\*\* Erläuterung:**

"Residuen" sind die Differenz zwischen Prognosewert und tatsächlichem Wert. Ihr Mittelwert hat keine Bedeutung. Wird die Standardabweichung quadriert, dann erhält

man die Fehlervarianz. Wird diese mit n (der Zahl der Untersuchungseinheiten multipliziert) dann erhält man die Fehlerstreuung als Quadratsumme:

$$0.391733 * 0.391733 * 1000 = 153.4547$$

Gesamt-    erklarte    multiple    Freiheitsgrade    F-Wert    Signifikanz

	streuung	Streuung	Korrelation	Zaehler	Nenner		(1-p)*100
Gruppe 1 nein	204.2040	50.7492	0.4985	992	7	46.8665	100.0000
Gruppe 2 ja	204.2040	50.7492	0.4985	992	7	46.8665	100.0000

**\*\*\*\*\* Erläuterung:**

Hier wird uns nun die multiple Korrelation und ihre Signifikanz mitgeteilt. Alle ursächlichen Variablen (gemessen in ihren Originalwerten) zusammen korrelieren mit der Original-Zielvariablen mit 0.4985. Dieser Koeffizient ist etwas kleiner als jener aus dem ungewichteten ALM (aus der 1. Analyse). Dieser war: 0.5561. Damit wird auch sichtbar, dass die Reproduzierbarkeit der gewichteten Analyse schlechter ist als die der ungewichteten. Das ist immer so.

Prognoseerfolg fuer Variable V10 Rueckzahl: nein, ja

		Haeufigkeit in Gruppe		
		tatsaechlich	davon richtig prognostiziert	zufaellig richtig
		-----	-----	-----
Gruppe 1	nein	286	31 (10.8 %)	82 (28.6 %)
Gruppe 2	ja	714	711 (99.6 %)	510 (71.4 %)

In 0 Faellen ist die tatsaechliche Gruppenzugehoerigkeit nicht bekannt (=Kein\_Wert)

Bei 44 Datensaeetzen (=4.4 %) liegt die prognostizierte Wahrscheinlichkeit ausserhalb des zulaessigen Wertebereichs von 0 bis 1

**\*\*\*\*\* Erlaeuterung:**

Die Tabelle des Prognoseerfolgs haben wir schon in Abschnitt P45.15.1.6 erkluert.

## **P45.15.2 Ursachen fuer die Zielvariable: Zielvariable ist nominal-polytom**

Unsere Zielvariable war seither eine dichotome, nominale Variable. Wir wollen nun den Fall betrachten, daB die Zielvariable polytom ist, d.h. mehr als 2 Auspraegungen besitzt.

Wir wollen hier nochmals auf die 2 Einschrueankungen hinweisen, die das Allgemeine Lineare Modell besitzt, wenn es auf nominale Zielvariable angewendet wird (siehe dazu die ausfuehrliche Darstellung in P45.15.1.0).

1. Es koennen Wahrscheinlichkeit auerhalb 0 – 1 auftreten.
2. Es besteht modellbedingte Varianz-Heterogenitaet.

Die 2. Einschrueankung kann durch die „gewichtete Kleinste-Quadrate-Schaetzung“ beseitigt werden. Wir werden diese Methode nachfolgend in P45.15.2.2 darstellen).

Als Alternative, die diese Probleme nicht besitzt (dafuer aber schwer zu interpretierende Ergebnisse liefert) werden wir in P45.16.2 die Logitanalyse darstellen.

Aus unseren Beispieldaten wollen wir nun die Variable „Produkt“ als Zielvariable verwenden. Die Frage lautet: Was sind die Ursachen, daB jemand Kleidung, oder Moebel oder Technik auf Kredit kauft?

Da unsere seitherigen Beispieldaten nicht gut geeignet sind, verwenden wir nun das etwas modifizierte Beispiel aus der Datei ".\Testdat\DatMin2.dir".

Wir rechnen eine 1. Analyse mit Pro45mf, wobei wir „Produkt“ als Zielvariable und alle anderen Variable als ursaechliche einsetzen.

Prog45mf haben wir in Abschnitt P45.15.1.1 abgebildet und die einzelnen Eingabe-Boxen in aller Ausfuehrlichkeit in P45.15.1.2 erlaeuert. Wir betrachten deswegen hier nur die Eingabe-Boxen, in denen die Eintraege geaendert wurden.

Die Eingabe-Box für die Datei der Variablennamen ist nun folgende:

Variablennamen

Datei der Variablennamen Hilfe

↔ 📁 "C:\Almo15\Testdat\DatMin2.nam"

↔ ↓ zeige      zeige = Namensdatei in Output zeigen  
leer = nicht zeigen

Durch Doppelklick auf den Dateinamen kann die Namensdatei geladen werden. Die Variablennamen sind folgende:

```
Name1=Wohnort:Stadt,Land;  
Name2=Geschlecht:m,w;  
Name3=Einkommen:bis 10,10-20,20 bis 30,30 bis 40,40-50;  
Name4=Hausbesitz:kein Haus,hat Haus;  
Name5=Rueckrate;  
Name6=Produkt:Kleidung,Möbel,Technik;  
Name7=Alter;  
Name8=Bildung;
```

Die Eingabe-Box für die zu lesende Datei ist folgende:

Datei aus der gelesen wird Hilfe

📁 "C:\Almo15\TESTDAT\DatMin2.dir"

Die beiden Eingabe-Boxen für die Zielvariable und die ursächlichen Variablen sind folgendermaßen auszufüllen.

Zielvariable

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 Zielvariable



ordinale Zielvariable

Hilfe



nominale Zielvariable

Hilfe



Ursächliche Variable

ursächliche nominale Variable

Wohnort, Geschlecht, Hausbesitz

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

paarweise Vergleiche (Kontraste) für die  
ursächlichen nominalen Variablen rechnen

---

ursächliche quantitative Variable

Einkommen, Alter, Bildung, Rueckrate

---

ursächliche ordinale Variable

Da die Zahlenwerte von Einkommen und Rückrate im Vergleich zu den Zahlenwerten der anderen ursächlichen Variablen sehr hoch sind, dividieren wir sie in der Umkodierungsbox mit 10 000 bzw. 100. Dadurch erreichen wir, daß die Regressionskoeffizienten, die sonst winzig klein ausfallen würden, um vier bzw. zwei Kommastellen nach vorne verschoben werden. Ansonsten ändert sich nichts.

Loesche wieder diese Sub-Box (Voreinstellungen wieder gueltig)

Eingabefelder für Umkodierungen und Kein-Wert-Angaben

Einkommen = Einkommen / 10000;  
  Rueckrate = Rueckrate / 100;

erzeuge zusätzliche Felder für Umkodierungen / Kein\_Wert-Angaben

In der Eingabe-Box „Ausgabe der Ergebnisse“ geben wir eine 2 ein, da wir nur eine stark verkürzte Ausgabe erhalten wollen.

### P45.15.2.1 Ausgabe

Almo liefert folgendes Ergebnis:

Streuungsquelle	Korrel Koeff.	Signifikanz p	(1-p)100
alle unabh. Var. zusammen	0.5529	0.0000	99.9995
quant./ordin. Var. zusammen	0.5196	0.0000	99.9995
nominale Variable zusammen	0.3932	0.0000	99.9995
V3 Einkommen	0.1467	0.0001	99.9851
V7 Alter	0.6970	0.0001	99.9950
V8 Bildung	0.5505	0.0001	99.9950
V5 Rueckrate	0.3569	0.0001	99.9950
V1 Wohnort	0.1129	0.0021	99.7852
V2 Geschlecht	0.4535	0.0000	99.9995
V4 Hausbesitz	0.4053	0.0000	99.9995

#### \*\*\*\*\* Erläuterung:

Alle ursächlichen Variablen liegen mit ihrer Signifikanz über der konventionellen 95 % - Grenze.

Die wirksamsten ursächlichen Variablen sind:

1. das Alter mit einer Korrelation von  $r = 0.697$
2. die Bildung 0.551
3. das Geschlecht 0.454
4. der Hausbesitz 0.405

#### \*\*\*\*\* Eine technische Anmerkung zum Korrelationskoeffizienten:

Wenn die Zielvariable eine nominale ist, die mehr als 2 Ausprägungen besitzt, dann rechnet Almo eine "multivariate" Analyse mit den Dummies der Zielvariablen als abhängige Variablen. Im Verlauf des Kalküls entsteht dabei "Pillais Spur". Der Korrelationskoeffizient in obiger Tabelle wird berechnet nach der Formel:

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

Siehe dazu auch Handbuch zu P20, Abschnitt P20.9.4.1

Wir rechnen eine 2. Analyse. In der Optionsbox „Prognosewerte und Residuen“ geben wir eine 1 in das Eingabefeld 1 ein, d.h. wir wollen den Prognoseerfolg kennen lernen.

Loesche wieder diese Box (dann Voreinstellungen wieder gueltig)

**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 %



In der Eingabe-Box „Ausgabe der Ergebnisse“ stellen wir auf 1 (= etwas verkürzte Ausgabe) ein.

**Ausgabe der Ergebnisse**

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

1= Basisstatistiken ausgeben  
 2= Basisstatistiken und "diverse Werte" ausgeben  
 0= nicht

Almo liefert folgende Ausgabe:

Prozentwerte der unabhängigen nominalen Variablen (zeilenweise auf 100 % addiert)  
je Ausprägung der abhängigen nominalen Variablen

(zeilenweise auf 100 % addiert)

	Produkt Kleidung	Produkt Möbel	Produkt Technik	
Wohnort Stadt	44.92	30.27	24.80	100 %
Wohnort Land	47.13	26.64	26.23	100 %
Geschlec m	34.03	32.77	33.19	100 %
Geschlec w	56.87	24.62	18.51	100 %
Hausbesitz kein Haus	59.48	26.01	14.52	100 %
Hausbesitz hat Haus	32.74	30.95	36.31	100 %

**\*\*\*\*\* Erläuterung:**

Stadt- und Landbewohner unterscheiden sich kaum. Frauen kaufen im Vergleich zu den Männern eher Kleidung (56.87% zu 34.03%). Ansonsten sind die Unterschiede zwischen den Geschlechtern gering. Hausbesitzer kaufen eher Technik (36.31% zu 14.52%). Nicht-Hausbesitzer kaufen eher Kleidung (59.48% zu 32.74%).

Mittelwerte der unabhängigen quantitativen Variablen  
je Ausprägung der abhängigen nominalen Variablen

	Produkt Kleidung	Produkt Möbel	Produkt Technik
Einkommen	2.90	2.91	2.91
Alter	59.31	52.90	46.45
Bildung	3.23	3.46	3.58
Rueckrate	612.93	595.53	568.54

**\*\*\*\*\* Erläuterung:**

Die Käufer der 3 Produktgruppen unterscheiden sich nicht in ihrem durchschnittlichen Einkommen.

Technische Produkte werden eher von "Jüngeren" gekauft (Durchschnittsalter 46.45), Kleidung eher von Älteren (Durchschnittsalter 59.31).

Haeufigkeiten je Auspraegung der nominalen Variablen

```

-----
V1 Wohnort
  V1-1 Stadt          512
  V1-2 Land           488

V2 Geschlecht
  V2-1 m              476
  V2-2 w              524

V4 Hausbesitz
  V4-1 kein Haus     496
  V4-2 hat Haus      504

V6 Produkt
  V6-1 Kleidung      460
  V6-2 Möbel         285
  V6-3 Technik       255
  
```

\*\*\*\*\* **Erläuterung:**  
 512 Personen wohnen auf dem Land etc.

Haeufigkeitstabelle:

Wohnort	Geschl.	Hausbes.	Produkt			Summe
			Kleidung	Möbel	Technik	
Stadt	männl	kein Hau	57	41	24	122
		hat Haus	25	38	56	119
	weibl	kein Hau	84	29	12	125
		hat Haus	64	47	35	146
Land	männl	kein Hau	61	38	26	125
		hat Haus	19	39	52	110
	weibl	kein Hau	93	21	10	124
		hat Haus	57	32	40	129
Summe			460	285	255	1000

\*\*\*\*\* **Erläuterung:**  
 Bei welcher Merkmalskombination wird vorzugsweise Kleidung gekauft? Von den 460 Personen, die Kleidung gekauft haben, sind die meisten (93 Personen) Frauen vom Land, die kein Haus haben.

```

Zahl der insgesamt eingelesenen Einheiten      1000
Zahl der in die Analyse einbezogenen Einheiten  1000
=====
  
```

Koeffizienten fuer quantitat./ordinale Variable aus univariater Analyse

hinsichtlich der abhaeng. Var. V6-1 Produkt: Kleidung

Variable	Regr. koeff.	part. Korrel.	Signifikanz p	(1-p)100
V3 Einkommen	-0.0776	-0.122	0.000	100.00
V7 Alter	0.0280	0.618	0.000	100.00
V8 Bildung	-0.2528	-0.480	0.000	100.00
V5 Rueckrate	0.0961	0.300	0.000	100.00

\*\*\*\*\* **Erläuterung:**

Hier werden einige Koeffizienten angegeben, die ausdrücken, wie stark die jeweilige ursächliche Variable die Zielvariable determinieren. Die Zielvariable ist eine nominale Variable mit 3 Ausprägungen. Zuerst wird nun die 1. Ausprägung "Kleidung" determiniert.

Die Regressionskoeffizienten können verwendet werden, um eine Prognose abzugeben. Es kann folgende Gleichung geschrieben werden:

$$p = -0.0776 * \text{Einkommen} + 0.0280 * \text{Alter} - 0.2528 * \text{Bildung} + 0.0961 * \text{Rueckrate} + \text{Effekte} + \text{Konstante}$$

$p$  = Wahrscheinlichkeit, dass Kleidung gekauft wird (zwischen 0 und 1)

Effekte = dies sind die Effekte der ursächlichen nominalen Variablen.

Wir werden weiter unten darauf zurückkommen

Konstante= dies ist eine rein rechnerische Größe, die inhaltlich kaum interpretierbar ist. Sie beträgt in unserem Falle -0.5548

Betrachten wir die 2. Person aus der Datei "DatMin2.fre". Ihre Werte in den Variablen sind folgende:

Wohnort	Geschlecht	Einkommen	Hausbesitz	Rueckrate	Produkt	Alter	Bildung
1	2	38790	1	627	1	57	3

Wohnort : 1=Stadt, 2=Land;  
 Geschlecht: 1=männl, 2=weibl;  
 Hausbesitz: 1=kein Haus, 2=hat Haus;  
 Produkt : 1=Kleidung, 2=Möbel, 3=Technik;

Beachte:

Das Einkommen haben wir in der Umkodierungsbox mit 10000 dividiert, so daß aus 38790 der Wert 3.8790 wird.

Die Rückrate haben wir in der Umkodierungsbox mit 100 dividiert, so dass aus 627 der Wert 6.27 wird.

Wenn wir diese Zahlenwerte in obige Gleichung einsetzen, dann erhalten wir

$$p = -0.0776 * 3.8790 + 0.0280 * 57 - 0.2528 * 3 + 0.0961 * 6.27 + \text{Effekte} - 0.5548$$

Für die Effekte, wir werden das weiter unten ausrechnen, erhalten wir 0.2487. Damit ergibt sich

$$p = 0.833$$

Die Wahrscheinlichkeit, dass die 2. Person aus unserer Datei Kleidung kauft ist also 83.3 %. Und tatsächlich hat diese Person auch Kleidung gekauft.

Der Regressionskoeffizient einer ursächlichen Variablen bestimmt also den Beitrag, den diese zur Wahrscheinlichkeit leistet, dass die betrachtete Ausprägung der nominalen Zielvariablen (hier: Kleidung) realisiert wird.

Betrachten wir die Bildung. Ihr Regressionskoeffizient ist -0.2528. Das bedeutet: Nimmt die Bildungsstufe um 1 Einheit zu, dann verringert sich die Wahrscheinlichkeit, dass Kleidung gekauft wird um 25.28 %.

Die Regressionskoeffizienten in der 1. Spalte sind kaum miteinander zu vergleichen. In ihrer Größe sind sie auch vom "Wertebereich" abhängig, den die ursächliche Variable einnimmt. So bewegt sich das Alter zwischen 19 und 89, die Bildung aber nur zwischen 1 und 6.

Vergleichbar sind nun jedoch die "partiellen Korrelationskoeffizienten". Den höchsten mit 0.618 besitzt das Alter. Es ist damit die stärkste ursächliche Variable. Wir bezeichnen den Korrelationskoeffizienten als einen "partiellen". Was damit gemeint ist, haben wir in Abschnitt P45.15.1.3.1 erläutert.

Koeffizienten hinsichtlich der abhaeng. Var. V6-2 Produkt: Möbel

Variable	Regr. koeff.	part. Korrel.	Signifikanz p	(1-p)100
V3 Einkommen	0.0177	0.022	0.489	51.11
V7 Alter	-0.0050	-0.109	0.001	99.93
V8 Bildung	0.0577	0.098	0.002	99.78
V5 Rueckrate	-0.0201	-0.051	0.105	89.52

**\*\*\*\*\* Erläuterung:**

Es fällt auf, dass die partiellen Korrelationen der ursächlichen quantitativen Variablen hinsichtlich "Produkt: Möbel" außerordentlich niedrig sind - während sie gegenüber "Kleidung" und "Technik" sehr hoch sind (ausgenommen das Einkommen). Dies gilt auch, wie weiter unten zu sehen ist, für die ursächlichen nominalen Variablen (Wohnort, Geschlecht und Hausbesitz). Es ist also zu erwarten, dass unser Modell hinsichtlich der abhängigen Variablen "Produkt: Möbel" keine zufriedenstellende Prognosefähigkeit besitzt.

Koeffizienten hinsichtlich der abhaeng. Var. V6-3 Produkt: Technik

Variable	Regr. koeff.	part. Korrel.	Signifikanz p	(1-p)100
V3 Einkommen	0.0599	0.099	0.002	99.82
V7 Alter	-0.0230	-0.563	0.000	100.00
V8 Bildung	0.1950	0.406	0.000	100.00
V5 Rueckrate	-0.0760	-0.253	0.000	100.00

"multivariate" partielle Korrelation zwischen der abhaengigen nominalen Variablen und den einzelnen unabhaengigen quantitat./ordinalen Variablen

Variable	part. Korrel	Signifikanz p	(1-p)100
V3 Einkommen	0.1467	0.000	99.99
V7 Alter	0.6970	0.000	99.99
V8 Bildung	0.5505	0.000	99.99
V5 Rueckrate	0.3569	0.000	99.99

**\*\*\*\*\* Erläuterung:**

Hier werden nun die 3 Dummy-Variablen der nominalen Variablen "Produkt" in einer multivariaten Analyse zusammengefasst. Dabei entsteht "Pillais Spur" (siehe dazu Almo-Handbuch zu P20, Abschnitt P20.9.4.1, Punkt 12), aus der ein Korrelationskoeffizient berechnet werden kann. Dadurch wird ein pauschaler Korrelationskoeffizient zwischen den ursächlichen Variablen und der polytomen Variablen "Produkt" gewonnen.

Wir erkennen, dass das Alter die am stärksten wirkende Variable ist und das Einkommen die schwächste.

Koeffizienten der Dummies  
hinsichtlich der abh. Var. V6-1 Produkt: Kleidung

Effekte von A Wohnort

	Effekte	partielle	Signifikanz	
		Korrelat.	p	(1-p)100
Stadt	-0.0403	-0.1093	0.0007	99.93%
Land	0.0403	0.1093	0.0007	99.93%

Effekte von B Geschlecht

	Effekte	partielle	Signifikanz	
		Korrelat.	p	(1-p)100
männl	-0.1626	-0.4028	0.0000	100.00%
weibl	0.1626	0.4028	0.0000	100.00%

Effekte von C Hausbesitz

	Effekte	partielle	Signifikanz	
		Korrelat.	p	(1-p)100
kein Haus	0.1264	0.3398	0.0000	100.00%
hat Haus	-0.1264	-0.3398	0.0000	100.00%

**\*\*\*\*\* Erläuterung:**

Wir haben oben ausgeführt, dass die Regressionskoeffizienten der quantitativen ursächlichen Variablen verwendet werden können, um eine Prognose abzugeben. Wir haben folgende Gleichung geschrieben:

$$p = -0.0776 \cdot \text{Einkommen} + 0.0280 \cdot \text{Alter} - 0.2528 \cdot \text{Bildung} + 0.0961 \cdot \text{Rueckrate} + \text{Effekte} + \text{Konstante}$$

Wir können jetzt die mit "Effekte" bezeichnete Größe auf der rechten Gleichungsseite bestimmen. Es gilt

$$\begin{aligned} \text{Effekte} = & -0.0403 \cdot \text{Stadt} + 0.0403 \cdot \text{Land} \\ & -0.1626 \cdot \text{männl} + 0.1626 \cdot \text{weibl} \\ & + 0.1264 \cdot \text{keinH} - 0.1264 \cdot \text{hatH} \end{aligned}$$

In dieser Gleichung sind die Bezeichnungen Stadt, Land, männl, weibl, keinH, hatH die 0-1 kodierte Dymmy-Variable der 3 ursächlichen nominalen Variablen Wohnort, Geschlecht, Hausbesitz.

Für eine Person mit folgenden Merkmalen: Stadt, weiblich, keinHaus lautet die Gleichung:

$$\begin{aligned} \text{Effekte} = & -0.0403 \cdot 1 + 0.0403 \cdot 0 \\ & -0.1626 \cdot 0 + 0.1626 \cdot 1 \\ & + 0.1264 \cdot 1 - 0.1264 \cdot 0 \\ = & 0.2487 \end{aligned}$$

Diesen Wert setzen wir oben in die Gleichung für p ein.

"Koeffizienten fuer quantitat./ordinale Variable aus univariater Analyse"  
 =====

Koeffizienten der Dummies  
 hinsichtlich der abh. Var. V6-2 Produkt: Möbel

Effekte von A Wohnort

	Effekte partielle Korrelat.	Signifikanz p	(1-p)100
Stadt	0.0247	0.0528	90.38%
Land	-0.0247	-0.0528	90.38%

=====

Effekte von B Geschlecht

	Effekte partielle Korrelat.	Signifikanz p	(1-p)100
männl	0.0510	0.1077	99.92%
weibl	-0.0510	-0.1077	99.92%

=====

Effekte von C Hausbesitz

	Effekte partielle Korrelat.	Signifikanz p	(1-p)100
kein Haus	-0.0241	-0.0540	91.13%
hat Haus	0.0241	0.0540	91.13%

=====

**\*\*\*\*\* Erläuterung:**

Die partiellen Korrelationen der ursächlichen nominalen Variablen hinsichtlich "Produkt: Möbel" sind niedrig - während sie gegenüber "Kleidung" und "Technik" eher hoch sind. Dies gilt auch, wie oben zu sehen war, für die ursächlichen quantitativen Variablen. Es ist also zu erwarten, dass unser Modell hinsichtlich der abhängigen Variablen "Produkt: Möbel" keine zufriedenstellende Prognosefähigkeit besitzt.

=====

Koeffizienten der Dummies  
 hinsichtlich der abh. Var. V6-3 Produkt: Technik

Effekte von A Wohnort

	Effekte partielle Korrelat.	Signifikanz p	(1-p)100
Stadt	0.0156	0.0447	84.08%
Land	-0.0156	-0.0447	84.08%

=====

Effekte von B Geschlecht

	Effekte partielle Korrelat.	Signifikanz p	(1-p)100
männl	0.1116	0.3025	100.00%

```

weibl  -0.1116  -0.3025  0.0000 100.00%
=====

```

Effekte von C Hausbesitz

```

          Effekte partielle  Signifikanz
          Korrelat.  p      (1-p)100
-----
kein Haus -0.1023  -0.2938  0.0000 100.00%
hat Haus  0.1023   0.2938  0.0000 100.00%
=====

```

Zusammenfassung

Streuungsquelle	generalisierte Streuung	Wilks Lambda	Korrel Koeff.	F-Wert	df	Signifikanz p	(1-p)100
Gesamtstreuung	33430.5000						
Fehlerstreuung	13082.1985				1982		
alle unabh. Var. zusammen	20348.3015	0.3913	0.5529	84.7401	14	0.0000	99.9995
quant./ordin. Var. zusammen	15328.9400	0.4605	0.5196	117.3549	8	0.0000	99.9995
nominale Variable zusammen	5838.5822	0.6914	0.3932	66.9328	6	0.0000	99.9995
V3 Einkommen	287.8408	0.9785	0.1467	10.9022	2	0.0001	99.9851
V7 Alter	12362.1748	0.5141	0.6970	468.2285	2	0.0001	99.9950
V8 Bildung	5688.4916	0.6969	0.5505	215.4567	2	0.0001	99.9950
V5 Rueckrate	1909.8539	0.8726	0.3569	72.3374	2	0.0001	99.9950
V1 Wohnort	169.0194	0.9872	0.1129	6.4018	2	0.0021	99.7852
V2 Geschlecht	3386.7212	0.7944	0.4535	128.2751	2	0.0000	99.9995
V4 Hausbesitz	2571.6031	0.8357	0.4053	97.4018	2	0.0000	99.9995

\*\*\*\*\* Erläuterung:

Die in der Zusammenfassung angegebenen Korrelationskoeffizienten sind "partielle". Sie entstanden aus "Pillais Spur" (siehe dazu P45.15.2.1). Wir erkennen, daß die am stärksten die Wahl des Produkts beeinflussende Variable das Alter ist.

Ist die Zielvariable nominal-polytom, dann wird sie von Almo in Dummy-Variable aufgelöst. Almo rechnet in diesem Falle eine multivariate Analyse. Dabei entstehen „generalisierte“ Streuungen, „Wilks Lambda“, „Pillais Spur“ usw. Zu diesen Begriffen siehe Handbuch zu P20, Abschnitt P20.9.4.

Multiple Korrelation aus univariater Analyse

hinsichtlich der abhaengigen Variablen V6-1 Produkt: Kleidung

```

-----
Fehlerstreuung                                121.545644
Durch alle unabhaeng. Variablen erklärte Streuung 126.854356
Multiples Bestimmtheitsmass                    0.510686
Multiple Korrelation                           0.714623
F-Wert f. erklarte Streuung                    147.903897
Freiheitsgrade Nenner = 7
          Zaehler= 992
Signifikanz: p                                0.000005
Signifikanz: (1-p)*100                        99.999500 %
Teststaerke von F                             1.000000

```

Multiple Korrelation aus univariater Analyse

hinsichtlich der abhaengigen Variablen V6-2 Produkt: Möbel

```

-----
Fehlerstreuung                                197.515626
Durch alle unabhaeng. Variablen erklärte Streuung 6.259374
Multiples Bestimmtheitsmass                    0.030717
Multiple Korrelation                           0.175263
F-Wert f. erklarte Streuung                    4.491000

```

```

Freiheitsgrade Nenner = 7
                  Zaehler= 992
Signifikanz: p                0.000169
Signifikanz: (1-p)*100       99.983126 %
Teststaerke von F            0.993680

```

Multiple Korrelation aus univariater Analyse  
hinsichtlich der abhaengigen Variablen V6-3 Produkt: Technik

```

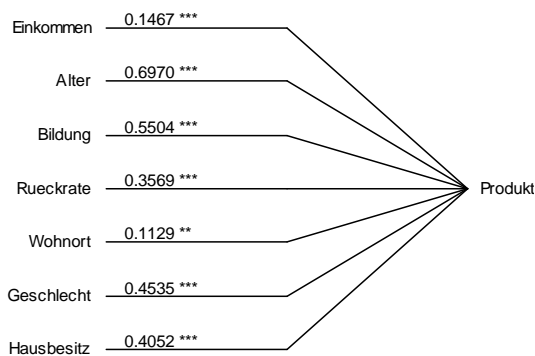
-----
Fehlerstreuung                110.016150
Durch alle unabhaeng. Variablen erklarte Streuung 79.958850
Multiples Bestimmtheitsmass    0.420891
Multiple Korrelation           0.648761
F-Wert f. erklarte Streuung   102.996799
Freiheitsgrade Nenner = 7
                  Zaehler= 992
Signifikanz: p                0.000005
Signifikanz: (1-p)*100       99.999500 %
Teststaerke von F            1.000000

```

**\*\*\*\*\* Erluterung:**

Almo liefert nun noch fur jede einzelne Auspragung der nominal-polytomem Zielvariablen "Produkt" die multiplen Korrelationen und ihre Signifikanzen. Wir erkennen, da alle ursachlichen Variablen zusammen mit der Auspragung "Kleidung" recht gut mit 0.714623 korrelieren. Hingegen ist die multiple Korrelation hinsichtlich "Moebel" mit 0.175263 unbefriedigend.

Almo zeichnet folgendes Fludiagramm der partiellen Korrelationen



3 Sterne hinter einem Koeffizienten bedeuten, dass die betreffende ursachliche Variable mit einer Signifikanz von  $(1-p)100 = 99.9\%$  wirkt. Bei 2 Sternen ist die Signifikanz 99 %, bei 1 Stern 95 %. Ist kein Stern vorhanden, dann liegt die Signifikanz irgendwo unterhalb 95 %. Die Wirkung der ursachlichen Variablen ist also nicht signifikant.

Der Benutzer kann die Signifikanzwerte fur die Sterne in der rechten Grafikleiste (siehe Abschnitt P45.15.1.3) anders definieren.

Zusammenfassung: Effekte und Regressionskoeffizienten  
und ihre Signifikanzen  
hinsichtlich der abhaengigen Variablen  
Produkt: Kleidung

	Effekte Regress.koeff	Signifikanz (1-p)*100
	-----	-----
A1 Stadt	-0.040346	99.930899
A2 Land	0.040346	99.930899
B1 m	-0.162560	99.995000
B2 w	0.162560	99.995000
C1 kein Haus	0.126379	99.995000
C2 hat Haus	-0.126379	99.995000
Einkommen	-0.077567	99.996875
Alter	0.027998	99.999500
Bildung	-0.252761	99.999500
Rueckrate	0.096067	99.999500

Zusammenfassung: Effekte und Regressionskoeffizienten  
und ihre Signifikanzen  
hinsichtlich der abhaengigen Variablen  
Produkt: Möbel

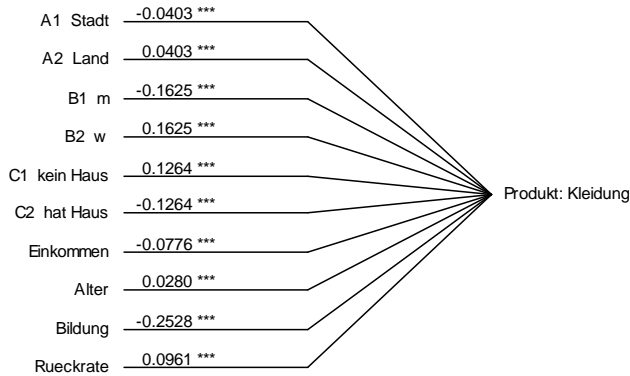
	Effekte Regress.koeff	Signifikanz (1-p)*100
	-----	-----
A1 Stadt	0.024732	90.378043
A2 Land	-0.024732	90.378043
B1 m	0.050997	99.917361
B2 w	-0.050997	99.917361
C1 kein Haus	-0.024116	91.130313
C2 hat Haus	0.024116	91.130313
Einkommen	0.017653	51.110490
Alter	-0.004959	99.930431
Bildung	0.057728	99.779093
Rueckrate	-0.020084	89.522155

Zusammenfassung: Effekte und Regressionskoeffizienten  
und ihre Signifikanzen  
hinsichtlich der abhaengigen Variablen  
Produkt: Technik

	Effekte Regress.koeff	Signifikanz (1-p)*100
	-----	-----
A1 Stadt	0.015614	84.082919
A2 Land	-0.015614	84.082919
B1 m	0.111563	99.995000
B2 w	-0.111563	99.995000
C1 kein Haus	-0.102263	99.995000
C2 hat Haus	0.102263	99.995000
Einkommen	0.059915	99.815200
Alter	-0.023039	99.999500
Bildung	0.195033	99.999500
Rueckrate	-0.075983	99.999500

Almo zeichnet folgendes Flußdiagramm der Effekte und Regressionskoeffizienten.

Effekte und Regressionskoeffizienten  
 A Wohnort: A1=Stadt A2=Land  
 B Geschlecht: B1=m B2=w  
 C Hausbesitz: C1=kein Haus C2=hat Haus



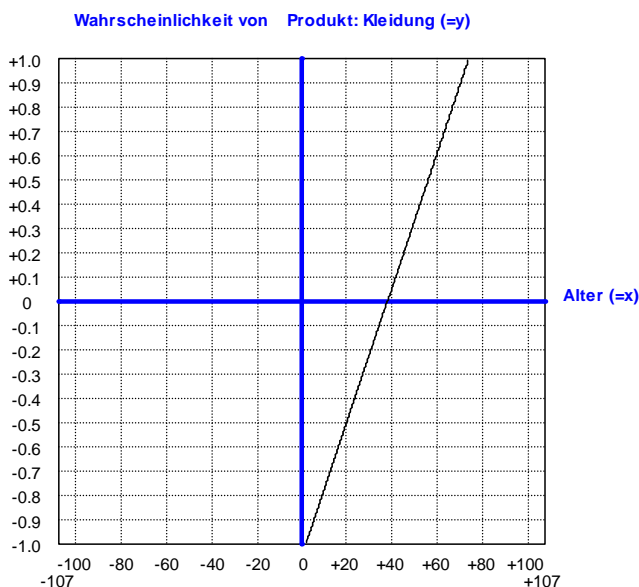
Wir geben hier nur das Flußdiagramm für "Produkt: Kleidung" wieder. Die anderen beiden Flußdiagramme für Möbel und Technik erhält der Benutzer, wenn er in der rechten Grafikleiste (siehe Abschnitt P45.15.1.4) bei

Welche Spalte der Grafikmatrix zeigen ?

auf "2" und "3" erhöht.

Almo zeichnet mehrere Funktionsgrafiken der ursächlichen quantitativen Variablen hinsichtlich der polytomen Zielvariablen. Wir wollen eine herausgreifen.

Lineare Funktion  
 $Y = 0.027998 * X - 1.0619$



Die Grafik zeigt den Zusammenhang zwischen Alter und der Wahrscheinlichkeit, dass Kleidung (im Versandhandel) gekauft wird. Die anderen ursächlichen quantitativen Variablen sind dabei auf ihren Mittelwert und die ursächlichen nominalen Variablen auf ihren Anteilswert gesetzt.

Die Grafik zeigt folgendes:

1. Die Gerade verläuft sehr steil, d.h. der Zusammenhang zwischen Alter und Kleidungskauf ist sehr stark.
2. Ab etwa 38 Jahren treten negative y-Werte, also Wahrscheinlichkeiten kleiner .0 auf
3. Ab etwa 75 Jahren treten y-Werte, also Wahrscheinlichkeiten über 1 auf

Wir interpretieren: Bis zu einem Alter von 38 Jahren ist die Wahrscheinlichkeit, daß Kleidung auf Kredit gekauft wird gleich 0, und ab einem Alter von 75 Jahren gleich 1.

Betrachten wir eine 60-jährige „Durchschnitts-Person“. Aus der Grafik lesen wir ab, daß diese Person eine Wahrscheinlichkeit von ca. 0.6 (=60%) besitzt Kleidung zu kaufen. Wir bezeichnen diese Person als „Durchschnitts-Person“. Damit meinen wir, daß sie in allen ursächlichen Variablen den Mittelwert (bzw. den Anteilswert bei nominalen ursächlichen Variablen) besitzt. Nur in der Variablen des Alters besitzt sie nicht den Mittelwert, sondern den Wert 60.

Berechnung der Prognosewerte und Residuen

-----

\*\*\*\*\* WARNUNG

Es koennen prognostizierte Wahrscheinlichkeiten ausserhalb 0-1 auftreten  
 Als Alternative ist die Logit-Analyse (Prog22m oder Prog45m9) moeglich  
 Bei ihr treten keine Wahrscheinlichkeiten ausserhalb 0-1 auf

Prognosewerte und Residuen fuer Variable V6 Produkt: Kleidung, Möbel, Technik

-----

Die Gruppe mit maximaler Wahrscheinlichkeit ist mit \* markiert

Die tatsaechliche Gruppenzugehoerigkeit wird hinter  
 der Datensatznummer in Klammern angegeben

Datensatz	prognostizierte Wahrscheinlichkeit der Zugehoerigkeit zu Gruppe			Residuen (Differenz) zu Gruppe		
	1 Kleidu	2 Möbel	3 Techni	1 Kleidu	2 Möbel	3 Techni
1 (1)	1.490*	0.099	-0.589	-0.490	-0.099	0.589
2 (1)	0.833*	0.211	-0.044	0.167	-0.211	0.044
3 (3)	0.068	0.357	0.575*	-0.068	-0.357	0.425
4 (1)	0.858*	0.222	-0.081	0.142	-0.222	0.081
5 (2)	0.416*	0.297	0.287	-0.416	0.703	-0.287
6 (1)	0.496*	0.256	0.248	0.504	-0.256	-0.248
7 (3)	0.171	0.343	0.485*	-0.171	-0.343	0.515
8 (1)	0.659*	0.247	0.094	0.341	-0.247	-0.094
9 (1)	0.608*	0.214	0.178	0.392	-0.214	-0.178
10 (1)	0.719*	0.246	0.035	0.281	-0.246	-0.035
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.

Prognoseerfolg fuer Variable V6 Produkt: Kleidung, Möbel, Technik

		Haeufigkeit in Gruppe		
		tatsaechlich	davon richtig prognostiziert	zufaellig richtig
		-----	-----	-----
Gruppe 1	Kleidung	460	435 (94.6 %)	212
Gruppe 2	Möbel	285	6 (2.1 %)	81
Gruppe 3	Technik	255	224 (87.8 %)	65

In 0 Faellen ist die tatsaechliche Gruppenzugehoerigkeit nicht bekannt (=Kein\_Wert)

Bei 289 Datensatzen (=28.9 %) liegt die prognostizierte Wahrscheinlichkeit ausserhalb des zulaessigen Wertebereichs von 0 bis 1

**\*\*\*\*\* Erläuterung:**

Almo gibt die Prognosewerte aus. Die Residuen interessieren hier nicht. Für Person 1 prognostiziert Almo für die 3 Gruppen Kleidung, Möbel, Technik folgende Wahrscheinlichkeiten

1.490    0.099    -0.589

Der höchste Wert wird für Gruppe 1 (Kleidung) ausgegeben. Tatsächlich gehört Person 1 auch zu dieser Gruppe.

Hier haben wir nun auch gleich ein Beispiel dafür, dass der zulässige Wahrscheinlichkeitsbereich von 0 bis 1 verlassen wird. Insgesamt tritt dieser Fall bei 28.9% unserer 1000 Datensätze auf. Das ist ein relativ hoher Wert. Er erklärt sich dadurch, dass die Zusammenhänge der ursächlichen Variablen mit den Zielvariablen "Produkt: Kleidung" und "Produkt: Technik" sehr stark sind (allerdings nicht mit "Produkt: Möbel"). Die Regressionsebene, die von den ursächlichen Variablen aufgespannt wird, ist damit sehr steil. Sie verlässt den zulässigen 0-1-Bereich sehr früh. Siehe die obige Grafik über den Zusammenhang von Alter und „Produkt:Kleidung“.

Die Prognosefähigkeit des Modells ist hinsichtlich Kleidung und Technik hervorragend. Die Gruppe der Möbelkäufer wird aber nur zu 2.1 % richtig prognostiziert. Würden wir aus den 1000 Personen zufällig 285 Möbelkäufer auswählen, dann würden wir nach den Gesetzen des Zufalls 81 richtig gefunden haben (das sind 28%). Wir müssen hier folgende Schlussfolgerung ziehen:

Wenn wir unser Modell auf Personen anwenden, von denen wir noch nicht wissen, welches Produkt sie kaufen werden, von denen wir aber die Werte aller ursächlichen Variablen kennen - wenn wir also echte Prognose leisten wollen - dann dürfen wir den Aussagen unseres Modells hinsichtlich Kleidung und Technik voll vertrauen, aber keinesfalls hinsichtlich Möbel.

Das Logit-Modell, das wir später in Abschnitt P45.16.2 rechnen erbringt hinsichtlich Kleidung und Technik eine geringfügig schlechtere Prognose, hinsichtlich Möbel aber eine sehr viel bessere.

### ***P45.15.2.2 Schritt 11b: Gewichtete Kleinste-Quadrate-Schätzung für nominal- polytome Zielvariable***

In Abschnitt P45.15.1.0 haben wir ausgeführt, dass bei nominaler Zielvariablen das Problem der Varianzheterogenität auftritt. In Abschnitt P45.15.1.7 haben wir gezeigt, wie dieses Problem bei dichotomer Zielvariablen durch eine "gewichtete Kleinste-Quadrate-Schätzung" gelöst werden kann. Dies gilt auch für polytome Zielvariable.

Die Vorgehensweise bei der gewichteten Analyse mit polytomer Zielvariablen ist folgende: Die polytome Zielvariable wird in so viele Dummy-Variable aufgelöst wie sie Ausprägungen besitzt. Für jede Dummy-Variable als Zielvariable wird eine separate Analyse mit Prog45gw gerechnet. Die jeweils anderen Dummy-Variablen der Zielvariablen werden zusammengefasst.

Im nachfolgend abgebildeten Programm Prog45gw haben wir unser Beispiel aus P45.15.2 verwendet.





Durch Umkodierung der Zielvariablen  
werden Dummies gebildet

Loesche wieder diese Box

**Umkodierungen und Kein-Wert-Angaben**

Umkodierungen   
Kein\_Wert-Angabe

Kleidung = Produkt(1=1; Sonst=2);  
  Moebel = Produkt(2=1; Sonst=2);  
  Technik = Produkt(3=1; Sonst=2);  
  Einkommen = Einkommen / 10000;

erzeuge zusätzliche Felder für Umkodierungen / Kein\_Wert-Angaben

Kontrollieren, ob Umkodierung so erfolgt wie gewünscht

diese Variablen ...

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

10

Option: Spezielle Kein-Wert-Behandlung

11

Option: Prognosewerte und Residuen

12

Option: Wertemuster

13

Option: "Aussehen" der auszugebenden Tabelle bzw. Matrix

14

Grafik-Optionen

15

Option: Die errechneten Koeffizienten in eine Datei speichern

16

**Ausgabe der Ergebnisse**

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

1= Basisstatistiken ausgeben  
2= Basisstatistiken und "diverse Werte" ausgeben  
0= nicht

17

#### P45.15.2.2.1 Erläuterung zu den Eingabe-Boxen

Prog45gw ist auch für dichotome Zielvariable verwendbar. Es ist für den Benutzer jedoch bequemer, so wie in Abschnitt P45.15.1.7 beschrieben, mit dem Programm Prog45mf zu rechnen.

Im abgebildeten Prog45gw ist die Zielvariable das gekaufte Produkt. Die 3 Ausprägungen sind

1. Kleidung
2. Moebel
3. Technik

Im Programm werden 3 Dummy-Variable gebildet, die diesen 3 Ausprägungen entsprechen. Für jede Dummy-Variable als Zielvariable wird eine Analyse gerechnet. D.h. Prog45gw muß 3 mal gerechnet werden.

Das abgebildete Programm Prog45gw entspricht weitgehend dem in Abschnitt P45.15.1.1 dargestellten Programm Prog45mf. Einige Options-Boxen, die in Prog45mf enthalten sind, sind in Prog45gw nicht vorhanden, weil die entsprechenden Optionen im Fall der gewichteten Analyse nicht verfügbar sind.

#### **Eingabe-Box 1 bis Eingabe-Box 3:**

Wie bei Prog45mf in Abschnitt P45.15.1.2.

#### **Eingabe-Box 4:** Ausprägungsnamen

Namen für die Ausprägungen der nominal-polytomen Zielvariablen Hilfe

verwenden Sie freie Variablennummern

```
Name 10=Kleidung:ja,nein;  
Name 11=Moebel:ja,nein;  
Name 12=Technik:ja,nein;
```

erzeuge zusätzliche Namensfelder

Hier werden die Namen für die Dummy-Variable, die als Zielvariable verwendet werden geschrieben. Die Namensgebung ist beliebig. Es ist aber sinnvoll, die Namen der Ausprägungen der nominal-polytomen Variablen als Variablennamen zu verwenden. Verwenden Sie Variablennummern die frei sind; am besten Nummern, die höher sind als die Nummer der letzten eingelesenen Variablen, aber niedriger als die Zahl der vereinbarten Variablen.

#### **Eingabe-Box 5:** Datei aus der gelesen wird

Wie Eingabe-Box 5 bei Prog45mf in Abschnitt P45.15.1.2.

### Eingabe-Box 6: Nominal-polytome Zielvariable

nominal-polytome Zielvariable Hilfe

nominale Zielvariable Hilfe

↔ □ □ Produkt

---

Dummy-Variable (=Ausprägung der Zielvariablen)

↔ □ □ Kleidung

↑ ↓ □ 1 das ist die x-te Ausprägung

*Eingabefeld 1:* Geben Sie hier die nominal-polytome Variable an. In unserem Beispiel ist das die Variable des Produkts.

*Eingabefeld 2:* Geben Sie hier die 1. Dummy-Variable an. In unserem Beispiel ist das die Variable "Kleidung".

*Eingabefeld 3:* Geben Sie hier an, für die wievielte Ausprägung die Dummy-Variable steht. In unserem Beispiel ist "Kleidung" die 1. Ausprägung von "Produkt". Also schreiben Sie: 1.

Wie wir bereits gesagt haben, muß Prog45gw 3 mal gerechnet werden - für jede der 3 Dummy-Variablen als Zielvariable.

Bei der 1. Analyse geben sie an  
im Eingabefeld 1: Produkt  
im Eingabefeld 2: Kleidung  
im Eingabefeld 3: 1

Bei der 2. Analyse geben sie an  
im Eingabefeld 1: Produkt  
im Eingabefeld 2: Moebel  
im Eingabefeld 3: 2

Bei der 3. Analyse geben sie an  
im Eingabefeld 1: Produkt  
im Eingabefeld 2: Technik  
im Eingabefeld 3: 3

Im Eingabefeld 1 steht also immer: Produkt.

Beachte: Dies ist die einzige Änderung, die sie bei den 3 Analysen vornehmen müssen.

### Eingabe-Box 7: Ursächliche Variable

Wie Eingabe-Box 7 bei Prog45mf in Abschnitt P45.15.1.2.

### Eingabe-Box 8: Option: Ein- und Ausschliessen von Untersuchungseinheiten

Wie Eingabe-Box 8 bei Prog45mf in Abschnitt P45.15.1.2.

## Eingabe-Box 9: Kein-Wert-Angabe und Umkodierungen

Durch Umkodierung der Zielvariablen  
werden Dummies gebildet

---

Loesche wieder diese Box

**Umkodierungen und Kein-Wert-Angaben**

Umkodierungen   
Kein\_Wert-Angabe

↔	↓	Kleidung = Produkt(1=1; Sonst=2);
↔	↓	Moebel = Produkt(2=1; Sonst=2);
↔	↓	Technik = Produkt(3=1; Sonst=2);
↔	↓	
↔	↓	Einkommen = Einkommen / 10000;

erzeuge zusätzliche Felder für Umkodierungen / Kein\_Wert-Angaben

---

Kontrollieren, ob Umkodierung so erfolgt wie gewünscht

diese Variablen ...

↔	□	
↔	□	

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

Wir haben hier gleich alle 3 Dummy-Variable erzeugt. Für die 1. Analyse wäre nur diejenige für die "Kleidung" notwendig gewesen. Die beiden anderen Umkodierungen zu Moebel und Technik stören jedoch nicht.

Die Dummy-Variable wird durch folgende zwei Umkodierung erzeugt.

```
Produkt (1=1; Sonst=2)  
Kleidung = Produkt;
```

Das Produkt wird dichotomisiert. Die Kleidung wird zu 1. Die beiden anderen Ausprägungen, Möbel und Technik, werden zu 2 zusammengefasst.

Die beiden obigen Anweisungen können in einer zusammengefasst werden:

```
Kleidung = Produkt (1=1; Sonst=2) ;      (Semikolon nicht vergessen !)
```

**Eingabe-Box 10:** Option: Spezielle Kein-Wert-Behandlung

**Eingabe-Box 11:** Option: Prognosewerte und Residuen

**Eingabe-Box 12:** Option: Wertemuster

**Eingabe-Box 13:** Option: "Aussehen" der auszugebenden Tabelle bzw. Matrix

**Eingabe-Box 14:** Grafik-Optionen

**Eingabe-Box 15:** Option: Die errechneten Koeffizienten in eine Datei speichern

**Eingabe-Box 16:** Ausgabe der Ergebnisse

**Eingabe-Box:** Ausreisser vom Typ 1 identifizieren

Wie die entsprechenden Eingabe-Boxen bei Prog45mf in Abschnitt P45.15.1.2.

Wir wollen es nochmals wiederholen: Für jede Dummy-Variable als Zielvariable muß eine Analyse gerechnet werden. In unserem Beispiel sind das 3 Dummy-Variablen und entsprechend 3 Analysen. Bei den wiederholten Analysen muß nur in Eingabe-Box 6 "Nominal-polytome Zielvariable" in der beschriebenen Weise geändert werden - sofern man in der Eingabe-Box 9 (Umkodierungen) gleich alle Dummy-Variable auf ein Mal bildet, wie wir es getan haben.

#### P45.15.2.2.2 Ausgabe der Ergebnisse

Wir unterstellen, dass die beiden Optionsboxen "Prognosewerte und Residuen" und "Errechneten Koeffizienten in eine Datei speichern" aktiviert wurden.

Für die 1. Analyse für die Dummy-Variable "Kleidung" als Zielvariable erhalten wir folgende Ausgabe - wenn in der Eingabe-Box "Ausgabe der Ergebnisse" auf 1, also auf "etwas verkürzte Ergebnisse" eingestellt wurde:

Almo gibt zuerst die Ergebnisse für die ungewichtete Kleinste-Quadrate-Lösung aus, die wir hier nicht zeigen. Dann folgt der Ausgabe-Abschnitt mit den Ergebnissen aus der gewichteten Analyse.

Die Ergebnisse sind genauso zu interpretieren, wie wir dies in Abschnitt P45.15.2.1 für das ungewichtete ALM vorgeführt haben. Wir ersparen uns deswegen die Erläuterung der nachfolgenden Ausgabe.

```
=====
=====
Ergebnisse aus gewichteter Analyse
(nach Korrektur der Varianzheterogenitaet)
=====
=====
```

Koeffizienten fuer quantitat./ordinale Variable aus univariater Analyse

hinsichtlich der abhaeng. Var. V10-1 Kleidung: ja

Variable	Regr. koeff.	part. Korrel.	Signifikanz p	(1-p)100
V3 Einkommen	-0.0409	-0.1783	0.0000	100.00
V7 Alter	0.0199	0.8987	0.0000	100.00
V8 Bildung	-0.2030	-0.7894	0.0000	100.00
V5 Rueckrate	0.0007	0.5552	0.0000	100.00
V91 Konstante	-0.2039	-0.2049	0.0000	100.00

hinsichtlich der abhaeng. Var. V10-2 Kleidung: nein

V3 Einkommen	0.0409	0.1783	0.0000	100.00
V7 Alter	-0.0199	-0.8987	0.0000	100.00
V8 Bildung	0.2030	0.7894	0.0000	100.00
V5 Rueckrate	-0.0007	-0.5552	0.0000	100.00
V91 Konstante	1.2039	0.7774	0.0000	100.00

Koeffizienten der Dummies

hinsichtlich der abh. Var. V10-1 Kleidung: ja

Effekte von A Wohnort

	Effekte	partielle Korrelat.	Signifikanz p	(1-p)100
A1 Stadt	-0.0192	-0.1466	0.0000	100.00%
A2 Land	0.0201	0.1466	0.0000	100.00%

Effekte von B Geschlecht

	Effekte partielle		Signifikanz	
	Korrelat.	p	(1-p)100	
B1 m	-0.1334	-0.6672	0.0000	100.00%
B2 w	0.1212	0.6672	0.0000	100.00%

Effekte von C Hausbesitz

	Effekte partielle		Signifikanz	
	Korrelat.	p	(1-p)100	
C1 kein Ha	0.0900	0.5444	0.0000	100.00%
C2 hat Hau	-0.0885	-0.5444	0.0000	100.00%

Die nachfolgenden Effekte beziehen sich auf „Kleidung: nein“, d.h. auf die zusammengefassten anderen Ausprägungen Möbel und Technik der polytomen Variablen „Produkt“. Diese Ergebnisse interessieren nicht.

Koeffizienten der Dummies  
hinsichtlich der abh. Var. V10-2 Kleidung: nein

Effekte von A Wohnort

	Effekte partielle		Signifikanz	
	Korrelat.	p	(1-p)100	
A1 Stadt	0.0192	0.1466	0.0000	100.00%
A2 Land	-0.0201	-0.1466	0.0000	100.00%

Effekte von B Geschlecht

	Effekte partielle		Signifikanz	
	Korrelat.	p	(1-p)100	
B1 m	0.1334	0.6672	0.0000	100.00%
B2 w	-0.1212	-0.6672	0.0000	100.00%

Effekte von C Hausbesitz

	Effekte partielle		Signifikanz	
	Korrelat.	p	(1-p)100	
C1 kein Ha	-0.0900	-0.5444	0.0000	100.00%
C2 hat Hau	0.0885	0.5444	0.0000	100.00%

Zusammenfassung: Effekte und Regressionskoeffizienten  
und ihre Signifikanzen  
hinsichtlich der abhaengigen Variablen  
Kleidung: ja

Effekte                      Signifikanz

	Regress.koeff	(1-p)*100
	-----	
A1 Stadt	-0.019190	99.995000
A2 Land	0.020134	99.995000
B1 m	-0.133374	99.995000
B2 w	0.121156	99.995000
C1 kein Haus	0.089973	99.995000
C2 hat Haus	-0.088545	99.995000
Einkommen	-0.040937	99.999500
Alter	0.019907	99.999500
Bildung	-0.202956	99.999500
Rueckrate	0.000662	99.999500
Konstante	-0.203873	99.999500

Zusammenfassung: Effekte und Regressionskoeffizienten  
 und ihre Signifikanzen  
 hinsichtlich der abhaengigen Variablen  
 Kleidung: nein

	Effekte Regress.koeff	Signifikanz (1-p)*100
	-----	
A1 Stadt	0.019190	99.995000
A2 Land	-0.020134	99.995000
B1 m	0.133374	99.995000
B2 w	-0.121156	99.995000
C1 kein Haus	-0.089973	99.995000
C2 hat Haus	0.088545	99.995000
Einkommen	0.040937	99.999500
Alter	-0.019907	99.999500
Bildung	0.202956	99.999500
Rueckrate	-0.000662	99.999500
Konstante	1.203873	99.999500

Almo liefert dann noch eine Reihe von Grafiken, die bereits in Abschnitt P45.15.2.1 gezeigt wurden.

\*\*\*\*\* MITTEILUNG  
 Koeffizientenmatrix wird erst in letzter Analyse gespeichert

\*\*\*\*\* MITTEILUNG  
 Prognosewerte werden erst in letzter Analyse ausgegeben

\*\*\*\*\* **Erläuterung:**

Almo weist darauf hin, dass erst nach der letzten Analyse, in unserem Beispiel nach der 3. Analyse, die Koeffizientenmatrix gespeichert und die Prognosewerte ausgegeben werden.

**Ausgabe nach der letzten Analyse**

Nachdem die 3. und letzte Analyse für die Dummy-Variable "Technik" als Zielvariable gerechnet wurde, gibt Almo zusätzlich noch aus

1. die Matrix der Regressionskoeffizienten und Effekte
2. die Prognosewerte
3. die multiple Korrelation und ihre Signifikanz
4. den Prognoseerfolg des Modells

2 wird nur ausgegeben, wenn der Benutzer die Eingabe-Box "Option: Prognosewerte und Residuen" aktiviert hatte und im 1. Eingabefeld 1 eingesetzt wurde. 1 wird ausgegeben, wenn die Eingabe-Box "Option: Die errechneten Koeffizienten in eine Datei speichern" und/oder die Eingabe-Box "Option: Prognosewerte und Residuen" aktiviert wurde.

Matrix der Regressionskoeffizienten und Effekte  
 der ursachlichen Variablen hinsichtlich der Dummies  
 der abhaengigen nominalen Variablen - aus gewichtetem ALM

			Produkt Kleidung V6-1	Produkt Möbel V6-2	Produkt Technik V6-3
Wohnort	Stadt	A1	-0.0192	0.0253	0.0133
Wohnort	Land	A2	0.0201	-0.0266	-0.0140
Geschlec	m	B1	-0.1334	0.0655	0.0586
Geschlec	w	B2	0.1212	-0.0595	-0.0533
Hausbesi	kein Hau	C1	0.0900	-0.0312	-0.0554
Hausbesi	hat Haus	C2	-0.0885	0.0307	0.0545
Einkomme		V3	-0.0409	0.0214	0.0385
Alter		V7	0.0199	-0.0069	-0.0126
Bildung		V8	-0.2030	0.0755	0.1082
Rueckrat		V5	0.0007	-0.0003	-0.0004
Konstant		V91	-0.2039	0.4957	0.6870

**\*\*\*\*\* Erläuterung:**

Almo hat diese Matrix aus den zuvor gerechneten 3 Analysen zusammengestellt. Die 1. Spalte obiger Matrix kann vom Benutzer aus den Ergebnissen der 1. Analyse (für "Kleidung") selbst rekonstruiert werden; entsprechend auch die 2. und 3. Spalte.

Die partiellen Korrelationen und die Signifikanzen gibt Almo nicht aus. Der Benutzer muß sie sich aus den 3 aufeinander folgenden Analysen selbst zusammenstellen.

Berechnung der Prognosewerte und Residuen

-----

\*\*\*\*\* MITTEILUNG  
 die Wahrscheinlichkeiten werden auf Summe=1.0 korrigiert

\*\*\*\*\* MITTEILUNG  
 Almo hat im Ordner "Zwisch" 3 Dateien mit dem Namen  
 "gew\_poly\_Koeffiz.mat\_x"  
 angelegt, die jetzt geloescht werden koennen

\*\*\*\*\* MITTEILUNG  
 Almo unterstellt, dass die eingelesene Koeffizientenmatrix  
 aus einem Allgemeinen Linearen Modell hervorgegangen ist

\*\*\*\*\* WARNUNG  
 Es koennen prognostizierte Wahrscheinlichkeiten ausserhalb 0-1 auftreten  
 Als Alternative ist die Logit-Analyse (Prog22m oder Prog45m9) moeglich  
 Bei ihr treten keine Wahrscheinlichkeiten ausserhalb 0-1 auf

\*\*\*\*\* MITTEILUNG  
 Sind unabhaengige Variable, die für die Berechnung  
 des Prognosewerts benoetigt werden, gleich "Kein\_Wert"  
 dann wird fuer sie "Kein-Wert-Behandlung = 7" durchgefuehrt  
 \*\*Hilfel74\*\*

Prognosewerte und Residuen fuer Variable V6 Produkt:  
 -----  
 Kleidung  
 Möbel  
 Technik

Die Gruppe mit maximaler Wahrscheinlichkeit ist mit \* markiert

Die tatsächliche Gruppenzugehörigkeit wird hinter der Datensatznummer in Klammern angegeben

Datensatz	prognostizierte Wahrscheinlichkeit der Zugehörigkeit zu Gruppe			auf 0-1 korrigierte Wahrscheinlichkeit der Zugehörigkeit zu Gruppe			Residuen (Differenz) zu Gruppe		
	1	2	3	1	2	3	1	2	3
	Kleidu	Möbel	Techni	Kleidu	Möbel	Techni	Kleidu	Möbel	Techni
1 (1)	1.201*	0.023	-0.235	0.848*	0.152	0.000	-0.201	-0.023	0.235
2 (1)	0.770*	0.179	0.071	0.755*	0.175	0.070	0.230	-0.179	-0.071
3 (3)	0.170	0.380	0.382*	0.182	0.408	0.410*	-0.170	-0.380	0.618
4 (1)	0.768*	0.189	0.054	0.760*	0.187	0.053	0.232	-0.189	-0.054
5 (2)	0.380*	0.297	0.217	0.425*	0.333	0.243	-0.380	0.703	-0.217
6 (1)	0.528*	0.248	0.203	0.539*	0.253	0.208	0.472	-0.248	-0.203
7 (3)	0.253	0.358*	0.338	0.267	0.377*	0.356	-0.253	-0.358	0.662
8 (1)	0.592*	0.232	0.121	0.627*	0.245	0.128	0.408	-0.232	-0.121
9 (1)	0.580*	0.198	0.174	0.608*	0.208	0.183	0.420	-0.198	-0.174
10 (1)	0.611*	0.226	0.088	0.660*	0.245	0.095	0.389	-0.226	-0.088
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
990 (1)	0.870*	0.126	-0.023	0.857*	0.143	0.000	0.130	-0.126	0.023
991 (1)	0.789*	0.135	0.050	0.810*	0.139	0.052	0.211	-0.135	-0.050
992 (1)	0.904*	0.126	0.001	0.877*	0.122	0.001	0.096	-0.126	-0.001
993 (2)	0.279	0.350*	0.294	0.302	0.379*	0.319	-0.279	0.650	-0.294
994 (2)	0.409*	0.307	0.275	0.413*	0.310	0.277	-0.409	0.693	-0.275
995 (3)	0.083	0.404	0.440*	0.089	0.436	0.475*	-0.083	-0.404	0.560
996 (3)	-0.017	0.412	0.548*	0.000	0.432	0.568*	0.017	-0.412	0.452
997 (2)	0.542*	0.255	0.130	0.585*	0.275	0.140	-0.542	0.745	-0.130
998 (3)	0.305	0.344*	0.337	0.309	0.349*	0.342	-0.305	-0.344	0.663
999 (2)	0.271	0.321	0.374*	0.280	0.332	0.388*	-0.271	0.679	-0.374
1000 (3)	0.272	0.389*	0.301	0.283	0.405*	0.313	-0.272	-0.389	0.699

**\*\*\*\*\* Erläuterung:**

Beim gewichteten ALM mit polytomer Zielvariablen tritt ein spezifisches Problem auf: Die Prognosewerte für die 3 Ausprägungen addieren sich nicht mehr zu 1.0. Wir führen deshalb eine einfache lineare Normierung der Prognosewerte auf Summe 1.0 durch, die gleichzeitig bewirkt, dass die einzelnen Prognosewerte nicht den Bereich 0 bis 1 verlassen. Betrachten wir Datensatz 996. Dort tritt ein negativer Wert auf. Alle 3 Prognosewerte werden um dessen Wert erhöht. Es entsteht also

$$0 \quad 0.429 \quad 0.565$$

Alle 3 Werte werden dann durch Ihre Summe 0.994 dividiert. Dadurch entstehen die auf den Bereich 0 bis 1 begrenzten und auf Summe 1.0 normierten "korrigierten" Prognosewerte.

Diese Korrektur ist nicht notwendig. Sie ist mehr eine ästhetische Maßnahme. Die Zugehörigkeit einer Person zu einer der 3 Ausprägungen wird nach dem maximalen Prognosewert prognostiziert. Ob wir die unkorrigierten oder die korrigierten Prognosewerte betrachten, die prognostizierte Zugehörigkeit ist dieselbe.

Mittelwert und Standardabweichung der Residuen  
 fuer Variable V6 Produkt: Kleidung, Möbel, Technik

	Mittelwert	Standardabweichung
Gruppe 1 Kleidung	-0.0034929	0.362313
Gruppe 2 Möbel	0.0047176	0.445081
Gruppe 3 Technik	0.0353111	0.355909

**\*\*\*\*\* Erläuterung:**

"Residuen" sind die Differenz zwischen unkorrigierten Prognosewert und tatsächlichem Wert. Ihr Mittelwert ist nahe 0. Wird die Standardabweichung quadriert, dann erhält man die Fehlervarianz. Wird diese mit n (der Zahl der Untersuchungseinheiten multipliziert) dann erhält man die Fehlerstreuung als Quadratsumme - beispielsweise für "Kleidung":

$$0.362313 * 0.362313 * 1000 = 131.2707$$

	Gesamt- streuung	erklärte Streuung	multiple Korrelation	Freiheitsgrade Zähler	Freiheitsgrade Nenner	F-Wert	Signifikanz (1-p)*100
Gruppe 1 Kleidung	248.4000	117.1294	0.6867	992	7	126.4480	100.0000
Gruppe 2 Möbel	203.7750	5.6779	0.1669	992	7	4.0619	99.9612
Gruppe 3 Technik	189.9750	63.3040	0.5773	992	7	70.8219	100.0000

**\*\*\*\*\* Erläuterung:**

Hier wird uns nun die multiple Korrelation und ihre Signifikanz mitgeteilt. Alle ursächlichen Variablen (gemessen in ihren Originalwerten) zusammen korrelieren mit der Original-Zielvariablen "Kleidung" mit 0.6867. Dieser Koeffizient ist etwas kleiner als jener aus dem ungewichteten ALM aus Prog45mf in Abschnitt P45.15.2.1. Dieser war: 0.714623. Damit wird auch sichtbar, dass die Reproduzierbarkeit der gewichteten Analyse schlechter ist als die der ungewichteten. Das ist immer so.

Prognoseerfolg fuer Variable V6 Produkt: Kleidung, Möbel, Technik

Haeufigkeit in Gruppe

	tatsaechlich	davon richtig prognostiziert	zufaellig richtig
Gruppe 1 Kleidung	460	444 (96.5 %)	212 (46.0 %)
Gruppe 2 Möbel	285	57 (20.0 %)	81 (28.5 %)
Gruppe 3 Technik	255	118 (46.3 %)	65 (25.5 %)

**\*\*\*\*\* Erläuterung:**

Der Prognoseerfolg hinsichtlich Kleidung und Technik ist deutlich besser als die zufällige Prognose, die für Möbel aber schlechter. Der Benutzer vergleiche diese Tabelle mit der, die das ungewichtete ALM erzeugt hat (Abschnitt P45.15.2.1).

In 0 Faellen ist die tatsaechliche Gruppenzugehoerigkeit nicht bekannt (=Kein\_Wert)

Bei 107 Datensaezten (=10.7 %) liegt die prognostizierte Wahrscheinlichkeit ausserhalb des zulaessigen Wertebereichs von 0 bis 1

```
***** MITTEILUNG
Lesen oder Schreiben korrekt beendet in Datei
"C:\Almo6\Progs\Koeffiz.fre"
```

```
***** MITTEILUNG
Lesen oder Schreiben korrekt beendet in Datei
"C:\Almo6\Zwisch\gew_poly_Koeffiz.mat_3"
```

**\*\*\*\*\* Erlaeuterung:**

Zur 2. Mitteilung: Almo hat bereits oben mitgeteilt, dass es 3 Dateien mit dem Namen "gew\_poly\_Koeffiz.mat\_x" angelegt hat, die jetzt geloescht werden koennen.

Zur 1. Mitteilung: Almo hat die Matrix der Koeffizienten in eine Datei gespeichert, fuer die der Benutzer einen Namen angegeben hatte. Diese Datei wird fuer die Prognose mit denselben Variablennamen aber anderen Daten verwendet. Siehe dazu unsere ausfuehrliche Darstellung in Kapitel 8 "Prognose leisten", Abschnitt P45.17. Die gespeicherte Koeffizientenmatrix kann, wenn man oben auf den Dateinamen einen Doppelklick ausuebt, in eine Fenster geladen und angeschaut werden. Da sie neben den Regressionskoeffizienten und Effekten noch weitere Informationen enthaelt, ist sie nur fuer den Almo-Spezialisten interpretierbar.

### P45.15.3 Ursachen für die Zielvariable: Zielvariable ist quantitativ

Wir verwenden wieder unsere Beispieldaten „Datmin2.fre“ bzw. „...dir“, die wir in P45.15.2 vorgestellt haben.

Als quantitative Zielvariable setzen wir die Variable „Rueckrate“, also die Höhe der Rückzahlungsrate für den Kredit ein. Die Frage lautet: Welche Variable bestimmen die Rückzahlungsrate und wie stark bestimmen sie diese? Wir führen mit Prog45mf eine Analyse durch, bei der wir alle anderen Variablen aus unseren Beispieldaten als ursächliche Variable einführen.

Die einzelnen Eingabe-Boxen von Prog45mf wurden bereits in Abschnitt P45.15.1.2 erläutert.

Die beiden Eingabe-Boxen für die Zielvariable und die ursächlichen Variablen aus Prog45mf haben nun mehr folgendes Aussehen.

Zielvariable 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 Zielvariable

Hilfe

---

ordinale Zielvariable Hilfe

---

nominale Zielvariable Hilfe

Ursächliche Variable

ursächliche nominale Variable

Wohnort, Geschlecht, Hausbesitz, Produkt

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

paarweise Vergleiche (Kontraste) für die  
ursächlichen nominalen Variablen rechnen

---

ursächliche quantitative Variable

Einkommen, Alter, Bildung

---

ursächliche ordinale Variable

Die Variable „Einkommen“ wird in der Umkodierungsbox durch 10 000 dividiert, damit sie ungefähr den gleichen Wertebereich überdeckt (von 0 bis 5) wie die anderen ursächlichen Variablen.

Lösche wieder diese Sub-Box

Eingabefelder für Umkodierungen und Kein-Wert-Angaben

Einkommen = Einkommen / 10000;

erzeuge zusätzliche Felder für Umkodierungen / Kein\_Wert-Angaben

Wir wollen die Prognosewerte sehen, die Almo für die Zielvariable errechnet. Dazu klicken wir auf die Eingabe-Box „Optionen: Prognosewerte und Residuen“ und geben dann in der eingeblendeten Eingabe-Box im Eingabefeld 1 eine 1 ein.

In der Eingabe-Box „Ausgabe der Ergebnisse“ wird eine 2 eingetragen um einen stark verkürzten Ergebnis-Output zu erhalten.

### P45.15.3.1 Ausgabe

Die Ausgabe entspricht weitgehend der für die dichotome Zielvariable, die wir in P45.15.1.3 bis P45.15.1.5 gezeigt und erläutert haben. Also liefert folgendes Ergebnis (stark verkürzt):

Zusammenfassung

Streuungsquelle	Korrel Koeff.	Signifikanz	
		p	(1-p)100
-----			
alle unabh. Var. zusammen	0.6286	0.0000	99.9995
quant./ordin. Var. zusammen	0.4980	0.0000	99.9995
nominale Variable zusammen	0.5346	0.0000	99.9995
V3 Einkommen	0.4445	0.0000	99.9995
V7 Alter	-0.2613	0.0000	99.9995
V8 Bildung	0.1770	0.0000	99.9995
V1 Wohnort	0.3253	0.0000	99.9995
V2 Geschlecht	0.4311	0.0000	99.9995
V4 Hausbesitz	0.1313	0.0000	99.9995
V6 Produkt	0.3569	0.0000	99.9995

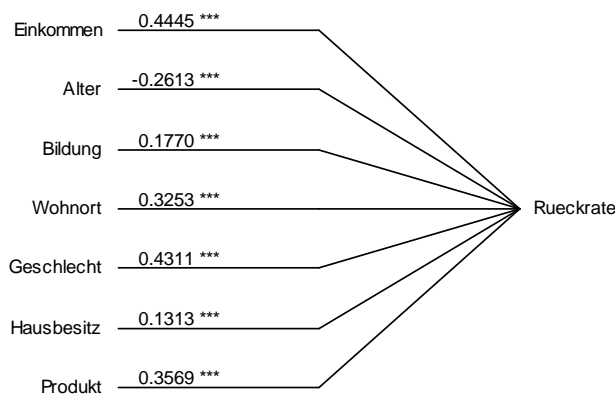
#### \*\*\*\*\* Erläuterung:

Alle ursächlichen Variablen haben einen hoch signifikanten Einfluß auf die Zielvariable der Rückzahlungsrate. Zusammen haben sie einen multiplen Korrelationskoeffizienten hinsichtlich der Zieldimension von 0.6286. Das ist nicht überragend, aber ordentlich. Den stärksten Einfluß hat das Einkommen mit einer partiellen Korrelation von 0.4445.

Zum Begriff des partiellen Korrelationskoeffizienten siehe P45.15.1.3.1.

Also liefert folgendes Flußdiagramm der partiellen Korrelationen

Partielle  
Korrelationskoeffizienten



- 1 Stern hinter dem Koeffizienten bedeutet: Ist signifikant mit 95 %
- 2 Sterne hinter dem Koeffizienten bedeutet: Ist signifikant mit 99 %
- 3 Sterne hinter dem Koeffizienten bedeutet: Ist signifikant mit 99.9 %

Die den Sternen zugeordneten Signifikanzwerte können im Grafikeditor auf der rechten Seite beliebig eingestellt werden.

Zusammenfassung: Effekte und Regressionskoeffizienten  
und ihre Signifikanzen  
hinsichtlich der abhaengigen Variablen  
Rueckrate

	Effekte Regress.koeff	Signifikanz (1-p)*100
A1 Stadt	36.708196	99.995000
A2 Land	-36.708196	99.995000
B1 m	54.602913	99.995000
B2 w	-54.602913	99.995000
C1 kein Haus	-15.344888	99.995000
C2 hat Haus	15.344888	99.995000
D1 Kleidung	75.616991	99.995000
D2 Möbel	-4.844614	65.952617
D3 Technik	-70.772376	99.995000
Einkommen	86.503095	99.999500
Alter	-3.962640	99.999500
Bildung	29.914546	99.999500

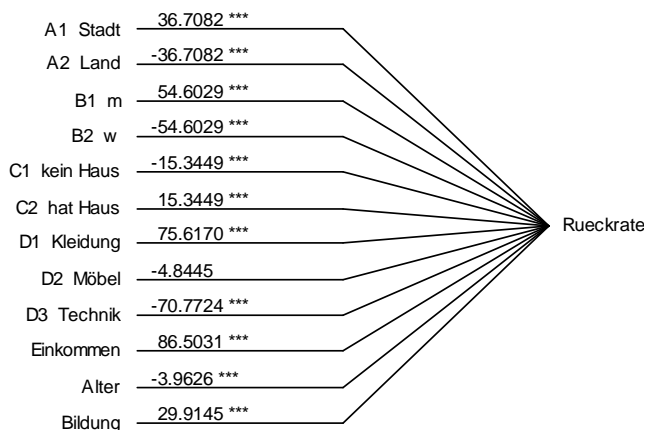
\*\*\*\*\* **Erläuterung:**

Die Effekte sind die Regressionskoeffizienten der Ausprägungen (=Dummies) der ursächlichen nominalen Variablen. Sie sind - am Beispiel des Geschlechts gezeigt - in folgender Weise zu interpretieren: Im Vergleich zur Durchschnittsperson aus Männern und Frauen zahlen Männer 54.6... Geldeinheiten mehr zurück – während Frauen 54.6... weniger zurückzahlen.

Die Regressionskoeffizienten der ursächlichen quantitativen Variablen – am Beispiel des Einkommens gezeigt - sind so zu interpretieren: Nimmt das Einkommen um 1 Einheit zu dann erhöht sich die Rückzahlung um 86.5... Geldeinheiten. Dabei ist zu berücksichtigen, dass das Einkommen in der Umkodierungsbox um 10 000 dividiert wurde. Wird also das Einkommen von 10 000 Geldeinheiten erhöht, dann erhöht sich die Rückzahlungsrate um 86.5 Geldeinheiten.

Almo liefert noch folgendes Flußdiagramm der Effekte und Regressionskoeffizienten

Effekte und Regressionskoeffizienten  
A Wohnort: A1=Stadt A2=Land  
B Geschlecht: B1=m B2=w  
C Hausbesitz: C1=kein Haus C2=hat Haus  
D Produkt: D1=Kleidung D2=Möbel D3=Technik



1 Stern hinter dem Koeffizienten bedeutet: Ist signifikant mit 95 %

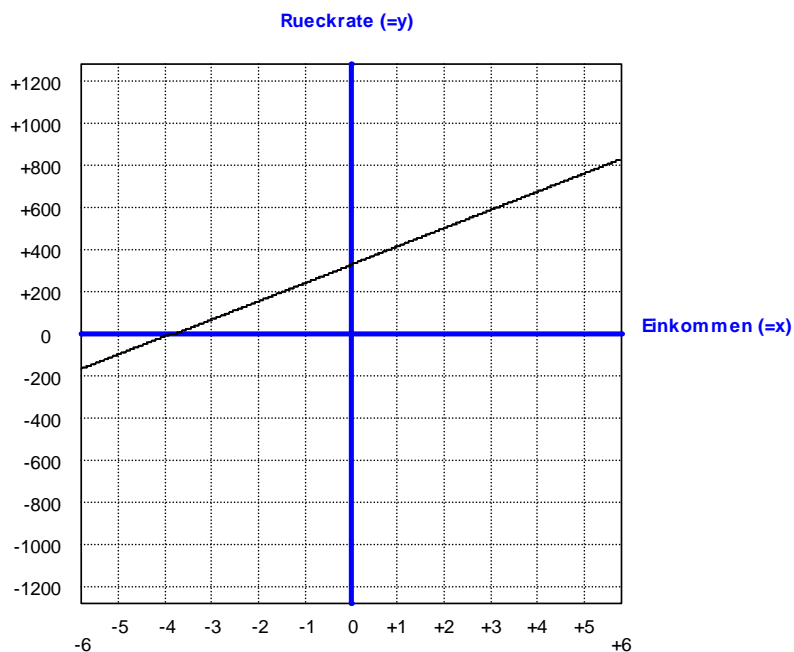
2 Sterne hinter dem Koeffizienten bedeutet: Ist signifikant mit 99 %  
3 Sterne hinter dem Koeffizienten bedeutet: Ist signifikant mit 99.9 %

Die den Sternen zugeordneten Signifikanzwerte können im Grafikeditor auf der rechten Seite beliebig eingestellt werden.

Almo zeichnet nun für jede ursächliche quantitative Variable die lineare Funktion hinsichtlich der Zielvariablen. Wir zeigen das hier nur für das Einkommen

Lineare Funktion für  
abhängige Variable: V5 Rueckrate  
unabhängige Variable: V3 Einkommen

Lineare Funktion  
 $Y = 86.503 * X + 331.7$



Im Grafik-Editor haben wir auf der rechten Bedienungsleiste folgende Einstellungen gewählt.

Hilfe			
Wände: grau/weiß			
Wände: da/weg			
<input checked="" type="checkbox"/>	Achsen: da/weg		
<input checked="" type="checkbox"/>	Kurve da/weg		
<input checked="" type="checkbox"/>	Maßzahl horizontal		
<input checked="" type="checkbox"/>	Gitterlinien da/weg		
<input checked="" type="checkbox"/>	Name an Punkt		
Kommastellen x-Achse			
0	+	-	OK Hilfe
Kommastellen y-Achse			
0	+	-	OK Hilfe
Schrift für Namen			
Schrift für Maßzahlen			
Schrift für Titel			
Schrift für Freie Texte			
Schrittweite			Hilfe
für Maßzahlen			
1	OK	x-Achse	
200	OK	y-Achse	

Die abgebildete Gerade zeigt den Zusammenhang zwischen Einkommen und Rückzahlungsrate - wobei alle anderen ursächlichen Variablen auf ihren Mittelwert gesetzt sind. Siehe die ausführliche Erläuterung zu dieser Art der grafischen Darstellung in Abschnitt P45.15.1.4.

Man kann z.B. ablesen, dass bei einem Einkommen von 3 Einheiten (also 30 000 Geldeinheiten) die Rückzahlungsrate etwas über 600 Geldeinheiten beträgt.

Wurde die Eingabe-Box "Option: Prognosewerte und Residuen" geöffnet, dann ermittelt Almo die Prognosewerte für alle 1000 Personen. Wir erhalten folgende Ausgabe:

Datensatz	tatsaechlicher Wert in der abaengigen Variablen V5 Rueckra	prognostizierter Wert in der abaengigen Variablen V5 Rueckra	Residuen (Differenz) V5 Rueckra
1	819.000	561.742	257.258
2	627.000	686.912	-59.912
3	536.000	535.567	0.43319
4	541.000	668.677	-127.68
5	824.000	542.744	281.256
6	594.000	689.282	-95.282
7	458.000	555.015	-97.015
8	656.000	647.652	8.34795
9	674.000	589.918	84.0819
10	792.000	665.511	126.489
.	.	.	.
.	.	.	.
.	.	.	.
995	764.000	619.215	144.785
996	318.000	444.185	-126.18
997	606.000	545.827	60.1726
998	507.000	538.336	-31.336
999	221.000	490.487	-269.49
1000	469.000	545.598	-76.598

Mittelwert und Standardabweichung der Residuen

	Mittelwert	Standardabweichung
V5 Rueckrate	-0.000311059	106.529

\*\*\*\*\* **Erläuterung:**

Betrachten wir den 1. Datensatz. Diese Person hat eine Rückzahlungsrate von 819 Geldeinheiten Unser Modell "prognostiziert" für sie eine Rückzahlungsrate von 561.742 Geldeinheiten. Es hat also um die Differenz von 257.258 "daneben getroffen". Die Differenzen (auch "Residuen" genannt) haben einen Mittelwert von -0.0003 (also 0). Das ist modellbedingt. Die Effekte und Regressionskoeffizienten sind so gewählt, dass der Mittelwert der Residuen gegen 0 strebt. Die Standardabweichung der Residuen mit 106.529 Geldeinheiten kann als ein Maß für die Güte unseres Modells betrachtet werden. Je kleiner diese Standardabweichung ist umso besser ist unser Modell.

## **P45.15.4 Weiterführende Hinweise**

Das Allgemeine Lineare Modell (ALM) ist als Prog20 in Almo mit einer Vielzahl von Varianten enthalten. Es wird im Almo-Handbuch „Kurt Holm: P20, Allgemeines Lineares Modell“ in aller Ausführlichkeit dargestellt. Auf die Problematik des ALM mit nominalen Zielvariablen gehen insbesondere ein: Urban (1993) und Aldrich/Nelson (1984).

## **P45.16 Schritt 11c: Alternative wenn Zielvariable nominal (dichotom oder polytom): Die Logit-Analyse**

### **P45.16.0 Einführung**

Wir haben in Abschnitt P45.15.1.0 darauf hingewiesen, daß beim Allgemeinen Linearen Modell, wie wir es mit Prog45mf rechnen, für die abhängige Variable „Rückzahlung: nein, ja“ Wahrscheinlichkeiten außerhalb des Bereichs 0 – 1 auftreten können. Beim Logit-Modell wird dieser Defekt durch die Unterstellung verhindert, der Zusammenhang zwischen ursächlicher und Zielvariable sei logistisch und nicht linear.

Wir wollen den Unterschied an einem konstruierten sehr einfachen Beispiel erläutern.

Es soll untersucht werden, wie das Einkommen die Kreditrückzahlung (nein, ja) bestimmt. Dieses Beispiel ist deswegen ein einfaches, weil nur eine unabhängige Variable verwendet wird.

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

Die abhängige Variable Kreditrückzahlung ist nominal. Kreditrückzahlung besitzt 2 Ausprägungen: ja und nein. Wir verwenden dafür in unseren Testdaten die Variable 10.

### **Rechnen mit dem ALM**

Wir rechnen zuerst mit Prog45mf (Abschnitt P45.15.1.1) ein Allgemeines Lineares Modell. Almo liefert uns folgende Ergebnisse (gekürzt):

```
-----  
Koeffizienten fuer quantitative Variable aus univariater Analyse  
hinsichtlich der abhaengigen Variablen      V10-0  Kreditrueckzahlung: ja  
  
                                     95%  
                                     Konfidenz-  
                                     bereich  
                                     nach  
Variable      Regr.   Standard   oben   erklarte   part.  F-Wert   Signifikanz   df1  df2   Test-  
               koeff.   fehler    u.unten  Streuung  Korrel.   p       (1-p)100     stärke  
-----  
V5 Einkommen  0.0640   0.0330   0.0659   0.9092   0.245    3.767   0.057   94.31  1    59    0.4806  
-----  
Koeffizienten fuer Konstante:      0.210491
```

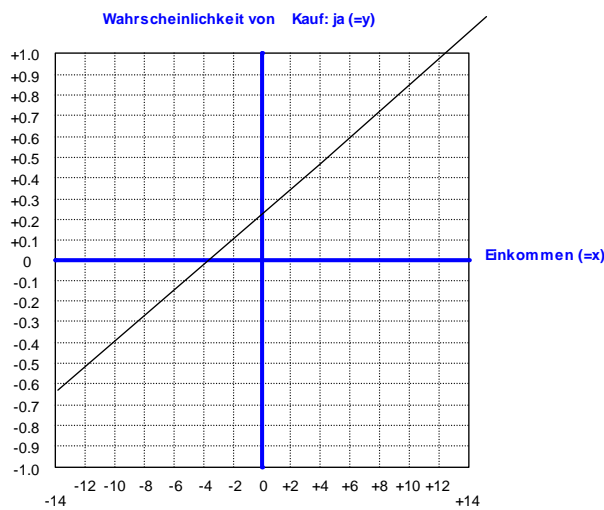
Die wesentlichen Ergebnisse sind also: Der Regressionskoeffizient beträgt 0.064. Er ist mit  $(1-p)100 = 94.31\%$  signifikant. Die Konstante hat einen Wert von 0.2105

Wir können also die lineare Gleichung schreiben:

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

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

Wir wollen die Gleichung als Gerade zeichnen:



Die Wahrscheinlichkeit der Rückzahlung ist z.B. bei einem Einkommen

$$\text{von 4 Einheiten: } p = 0.064 * 4 + 0.2105 = 0.4665$$

$$\text{von 8 Einheiten: } p = 0.064 * 8 + 0.2105 = 0.7225$$

$$\text{von 10 Einheiten: } p = 0.064 * 10 + 0.2105 = 0.8505$$

Der höchste Einkommenswert in unseren Daten ist 10.

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

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

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

### Rechnen mit der Logitanalyse

Die Logit-Analyse besitzt diese Schwäche nicht. Wir wollen mit denselben Daten eine Logit-Analyse rechnen.

Da bei der Logit-Analyse in Almo im Unterschied zum Allgemeinen Linearen Modell die 1. Dummy der abhängigen Variablen eliminiert wird, müssen wir, um die Ergebnisse vergleichen zu können, die Variable Rückzahlung umkodieren: Aus 0 wird 1 und aus 1 wird 0.

Almo liefert folgende Ergebnisse (gekürzt):

```
-----
Ergebnisse für 2. Auspraegung ja der abh. Var. V10 Rueckzahlung
unabh.      Regress.   "Risiko"   Stand.- z-Wert  Signifik.  partielle
Variab.     Koeffiz.    exp(Regr.- Fehler  (1-p)*100  Korrelat.
            koeffiz.)
-----
Konstante  -1.21869    -           0.62553    1.95       94.86      -
Einkommen  0.27058    1.31072    0.14594    1.85       93.63      0.13069
-----
```

Die Logit-Analyse verwendet die logistische Funktion. Deren Gleichung ist:

$$p = \frac{1}{1 + e^{-(c+\beta x)}}$$

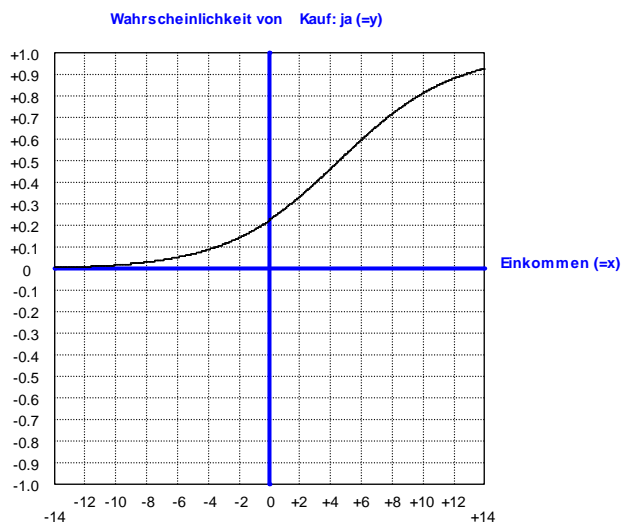
wobei  $c$  = Konstante und  $\beta$  = Regressionskoeffizient

Für unser Beispiel lautet die Gleichung:

$$p = \frac{1}{1 + e^{-(-1.21869+0.27058 \cdot \text{Einkommen})}}$$

Almo liefert folgende Grafik:

Logistische Funktion  
 $Y = 1/(1+e^{*-(-1.2+0.27*X)})$



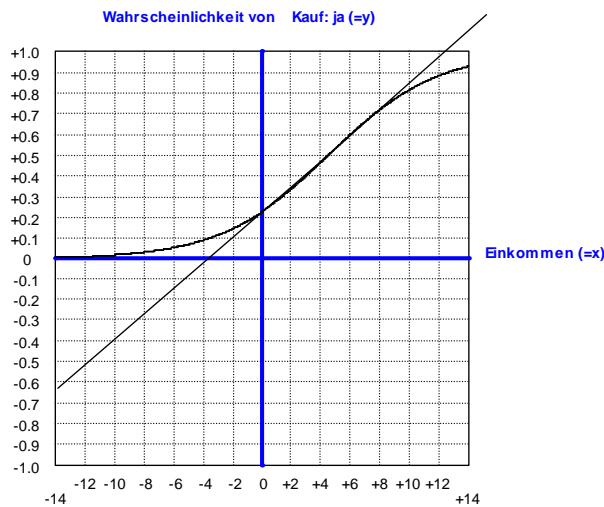
Eigenschaften der logistischen Funktion:

1. Die logistische Funktion nähert sich asymptotisch den Werten  $p = 0$  und  $p = 1$
2. Die Konstante  $c$  bestimmt die horizontale Lage der Kurve. Je grösser  $c$  umso weiter links liegt die Kurve - bei positivem  $\beta$ . Ist  $\beta$  negativ dann umgekehrt.
3. Der Regressionskoeffizient  $\beta$  bestimmt die Steilheit. Je grösser  $\beta$  absolut ist umso steiler. Bei positivem Vorzeichen wächst die Kurve von links nach rechts, bei negativem Vorzeichen umgekehrt

Die Wahrscheinlichkeit der Rückzahlung ist bei einem Einkommen

Einkommen	Logit-Analyse	Allgemeines Lineares Modell
von 4 Einheiten:	p = 0.4660	0.4665
von 8 Einheiten:	p = 0.7203	0.7225
von 10 Einheiten:	p = 0.8156	0.8505
von 14 Einheiten:	p = 0.9289	1.1950

Nun wollen wir die Gerade und die logistische Funktion in einer gemeinsamen Grafik zeigen:



Man erkennt sehr deutlich, daß die Gerade und die logistische Funktion von Einkommen = 0 bis zu einem Einkommen von ca. 9 Einheiten sich decken. Erst dann gehen die beiden auseinander.

## P45.16.1 Logitanalyse mit dichotomer Zielvariablen

### P45.16.1.1 Eingabe in Programm

Prog45m9.Msk

Wirkungsstärke der ursächlichen Variablen  
hinsichtlich einer nominalen Zielvariablen

Es wird eine Logit-Analyse gerechnet  
(alternativ kann auch eine Probit-Analyse gerechnet werden)

Almo-Struktur -->   
Bedienung -->

- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8
- 9
- 10

Vereinbare Variable=  ;

Option: Weitere Vereinbarungen - nur wenn Almo dazu auffordert

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

     zeige = Namensdatei in Output zeigen  
leer = nicht

erzeuge zusätzliche Namensfelder

"C:\Almo7\TESTDAT\DatMin.dir"

ursächliche nominale Variable

---

ursächliche quantitative Variable

Option: Alternative: Probitanalyse

Option: Auflösung der unabhäng. nominalen Variab. in Dummies

Option: Ein- und Ausschliessen von Untersuchungseinheiten

11

↓ Loesche wieder diese Box

**Umkodierungen und Kein-Wert-Angaben**

Umkodierungen   
Kein\_Wert-Angabe

↔ ↓ Einkommen = Einkommen / 10000;  
↔ ↓  
↔ ↓

erzeuge zusätzliche Felder für Umkodierungen / Kein\_Wert-Angaben

---

Kontrollieren, ob Umkodierung so erfolgt wie gewünscht

diese Variablen ...

↔ □ □ Einkommen  
↔ 1:50

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

12

↓ Option: Prognosewerte ermitteln

13

↓ Option: Wertemuster

14

↓ Option: Die errechneten Koeffizienten in eine Datei speichern

15

↓ Grafik-Optionen

16

**Ausgabe der Ergebnisse**

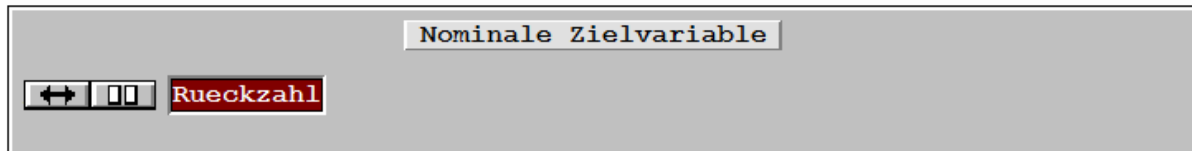
↑↓ 1  
↑↓ 0

0= Ergebnisse in voller Länge ausgeben  
1= Ergebnisse verkürzt ausgeben  
  
1= Basisstatistiken ausgeben  
0= nicht

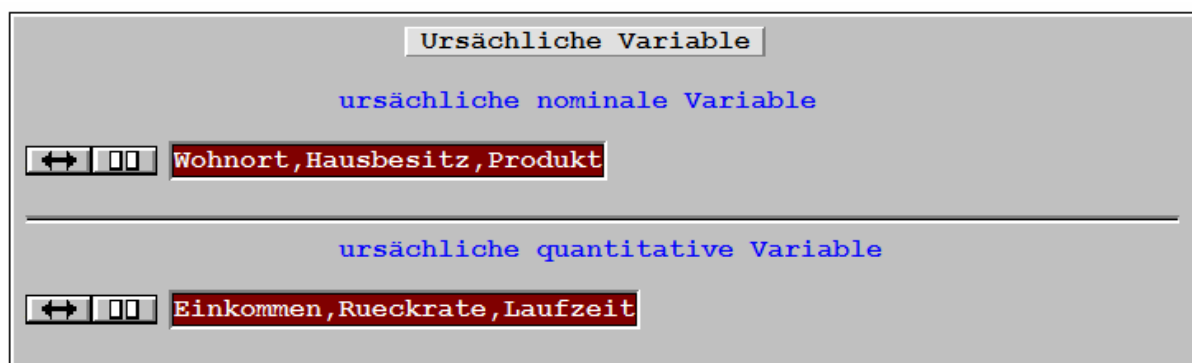
## P45.16.1.2 Erläuterungen zu den Eingabe-Boxen

**Eingabe-Box 1 bis 5:** entsprechen den Eingabe-Boxen aus Prog45mf  
Siehe dazu auch "Arbeiten mit Almo-Datenanalyse-System", Abschnitt P0.1 bis P0.4.

**Eingabe-Box 6:** Zielvariable

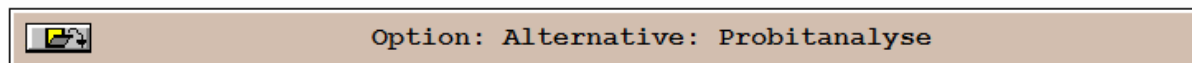


**Eingabe-Box 7:** Ursächliche Variable

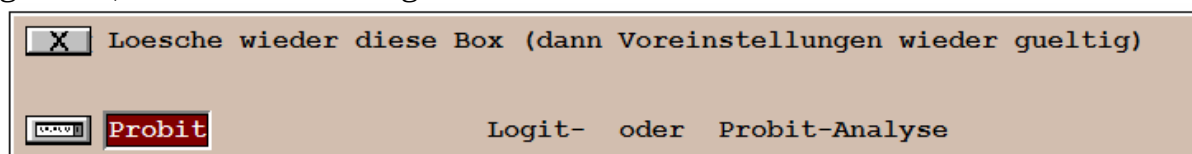


Die Variablen sind wieder dieselben, wie bei Prog45mf (Abschnitt P45.15.1.2), so daß wir hierzu keine Erläuterung benötigen.

**Eingabe-Box 8:** Option: Alternative Probit-Analyse



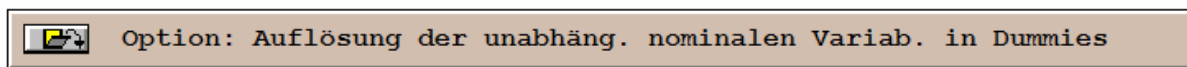
Wird die Optionsbox durch Klick auf den Knopf mit nach unten weisenden Pfeil geöffnet, dann sieht man folgendes



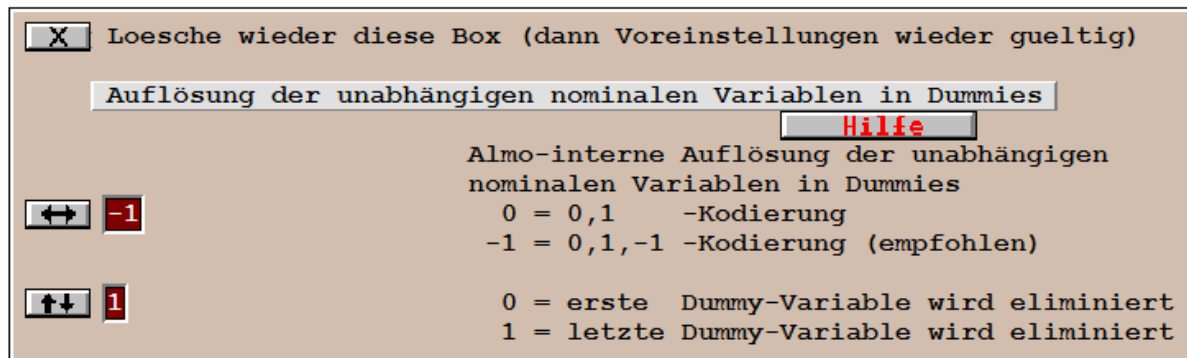
Anstelle der Logit-Analyse wird eine Probit-Analyse gerechnet. Die Probit-Analyse verwendet anstelle der logistischen Funktion die Ogive. Das Kurvenbild der beiden Funktionen ist fast identisch. Die Probit-Analyse hat jedoch Nachteile gegenüber der Logit-Analyse. Es existiert kein Koeffizient der analog zum "Risiko"-Koeffizienten aus der Logit-Analyse interpretierbar ist. Außerdem ist die Probit-Analyse nur auf abhängige dichotome Variable, nicht auf abhängige polytom-nominale Variable, anwendbar. Polytome Variable können bei der Probit-Analyse nur analysiert werden, wenn sie als ordinale betrachtet werden.

Siehe unsere ausführliche Darstellung der Probit-Analyse im Handbuch, Teil 4, P22.

**Eingabe-Box 9:** Option: Auflösung der unabhängigen nominalen Variablen in Dummies



Wird die Optionsbox durch Klick auf den Knopf geöffnet, dann sieht man folgendes.



Wir möchten vorweg empfehlen, die angebotenen Optionen nur dann zu verwenden, wenn man dafür einen Grund hat.

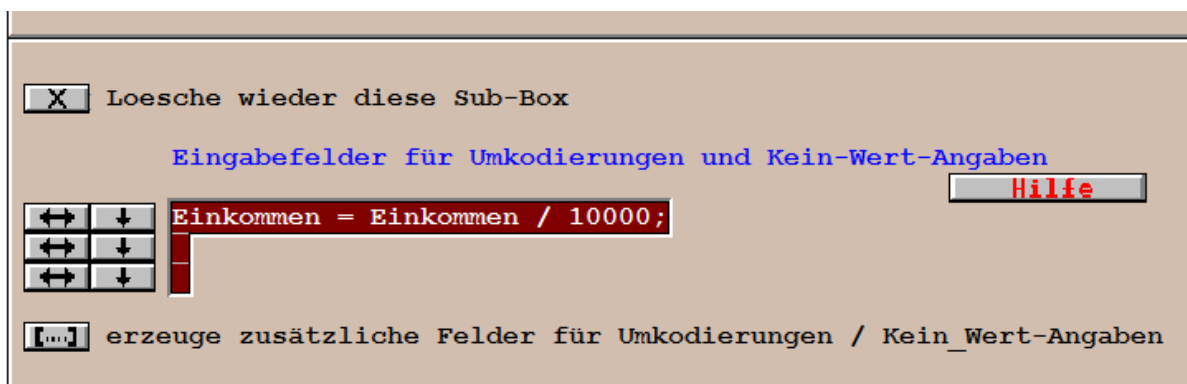
*Eingabefeld 1:* Die unabhängigen nominalen Variablen werden in Dummies aufgelöst. Dabei kann die 0,1 - Kodierungsmethode oder die 0,1, -1 - Kodierungsmethode verwendet werden. Siehe dazu auch Almo-Handbuch zu P20, Abschnitt P20.3. Wir werden den Unterschied bei der Besprechung der Ergebnisse aus der Logitanalyse erklären.

*Eingabefeld 2:* Zur Vermeidung von Redundanz muß eine Dummy eliminiert werden.

Wir die Optionsbox nicht geöffnet, dann rechnet Almo mit der 0,1, -1 -Kodierung und eliminiert die letzte Dummy.

**Eingabe-Box 10:** Ein- und Ausschließen von Untersuchungseinheiten  
Siehe "Arbeiten mit Almo-Datenanalyse-System", Abschnitt P0.7.

**Eingabe-Box 11:** Kein-Wert-Angabe und Umkodierung

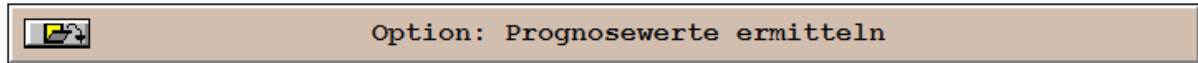


Da die Variable des Einkommens sehr hohe Werte annimmt, dividieren wir sie durch 10 000. Wir werden die Auswirkung dieser Umkodierung bei der Besprechung der Ergebnisse in Abschnitt P45.16.1.3 darstellen. Die Eingabe-Box wird im "Arbeiten mit Almo-Datenanalyse-System", Abschnitt P0.5 ausführlich erläutert.

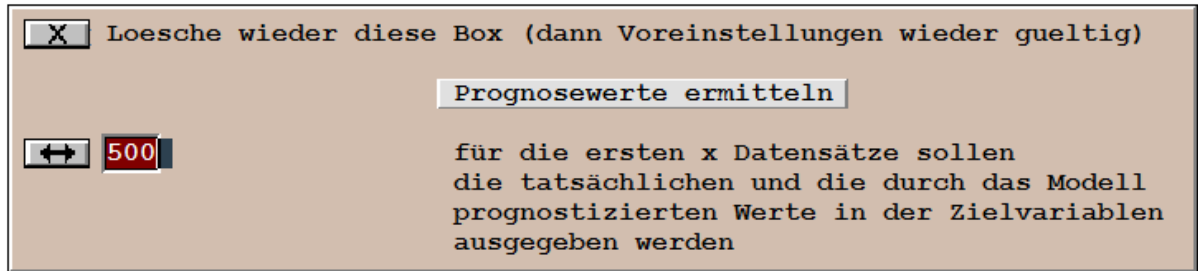
### Eingabe-Box: Ausreisser vom Typ 1 identifizieren

Siehe die entsprechenden Eingabe-Boxen bei Prog45mf in Abschnitt P45.15.1.2.

### Eingabe-Box 12: Prognosewerte ermitteln

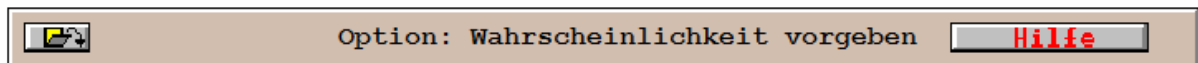


Wird auf den Knopf dieser Eingabe-Box geklickt, dann wird folgende Option angeboten.

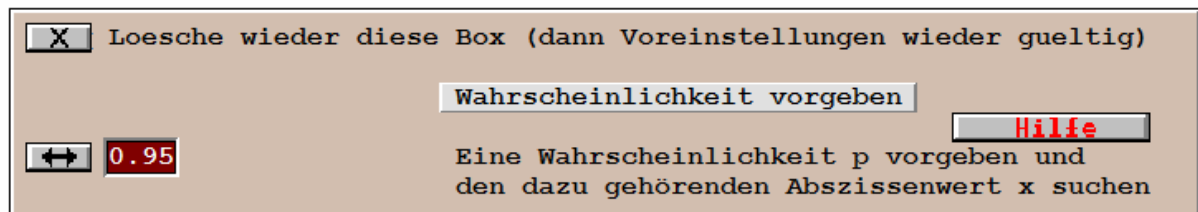


Wenn der Benutzer beispielsweise 500 eingibt, dann wird für die ersten 500 Personen der Datei angegeben, welchen Wert sie in der Zielvariablen (Rückzahlung: nein, ja) tatsächlich haben und welchen das Logit-Modell prognostiziert.

### Eingabe-Box: Wahrscheinlichkeit vorgeben



geöffnet:



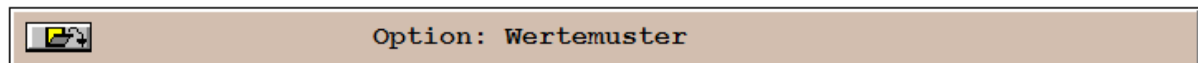
Beispiel:

Sie wollen wissen, welcher Abszissen-Wert (=x-Wert) der Logit- bzw. Probitfunktion hat eine Wahrscheinlichkeit (=y-Wert) von 95 %

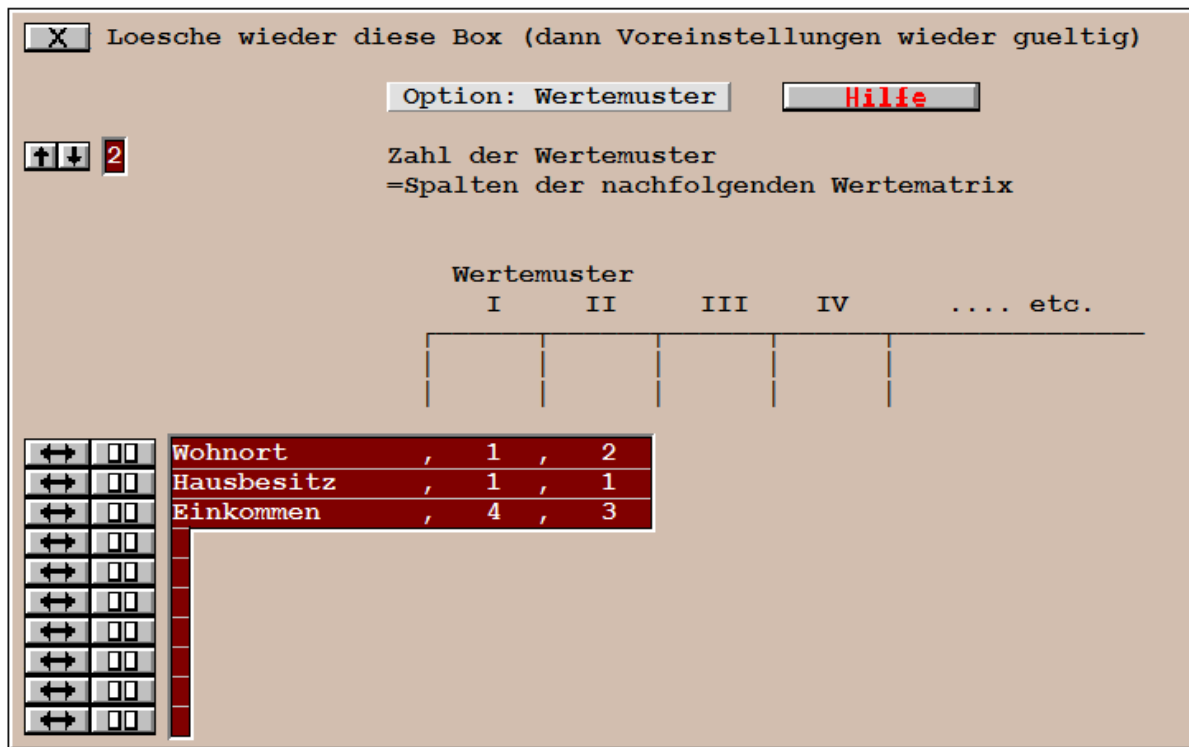
Dazu geben Sie in das Eingabefeld ein: 0.95

Almo ermittelt dann den x-Wert, über dem die Logit- bzw. Probit-Kurve eine Höhe von  $y=0.95$  besitzt.

### Eingabe-Box 13: Option: Wertemuster



Wird die Optionsbox durch Klick auf den Knopf mit nach unten weisenden Pfeil geöffnet, dann sieht man folgendes



Der Benutzer kann sich von Almo berechnen lassen, welche Wahrscheinlichkeit der "Rückzahlung:Nein" bzw der "Rückzahlung:Ja" eine Person mit bestimmten Werten in einer oder mehreren oder allen unabhängigen Variablen hat.

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

Der Benutzer will nun die Wahrscheinlichkeit der Rückzahlung prognostizieren für

1. Städter, die kein Haus besitzen und ein Einkommen von 4 besitzen
2. Landbewohner, die kein Haus besitzen und ein Einkommen von 3 besitzen

(Einkommen wurde in der Umkodierungsbox mit 10 000 dividiert)

Geben Sie als Zahl der Wertemuster = 2 an und schreiben Sie in die Eingabefelder der Wertemustermatrix.

	Wertemuster				
	I	II	III	IV	.... etc.
[ Wohnort	, 1	, 2			
[ Hausbesitz	, 1	, 1			
[ Einkommen	, 4	, 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.

### WICHTIG:

Als Trennzeichen innerhalb eines Eingabefeldes muss ein Beistrich geschrieben werden, auch hinter dem Variablennamen (bzw. Variablennummer). Am Zeilenende wird kein Beistrich geschrieben.

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. Almo 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:

	Wertemuster				
	I	II	III	IV	.... etc.
[Geschlecht	, 1	, 2			
[Alter	, 48	, 58			
[Einkommen	, 4	, kw			

└────────────────── kw eingesetzt

Der Benutzer will 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

Almo setzt dann beim 2. Wertemuster für das Einkommen dessen Mittelwert ein.

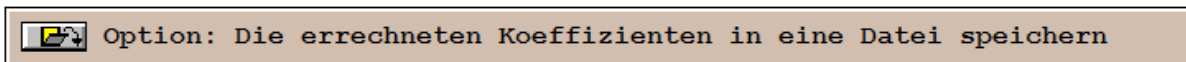
### Hinweis:

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.

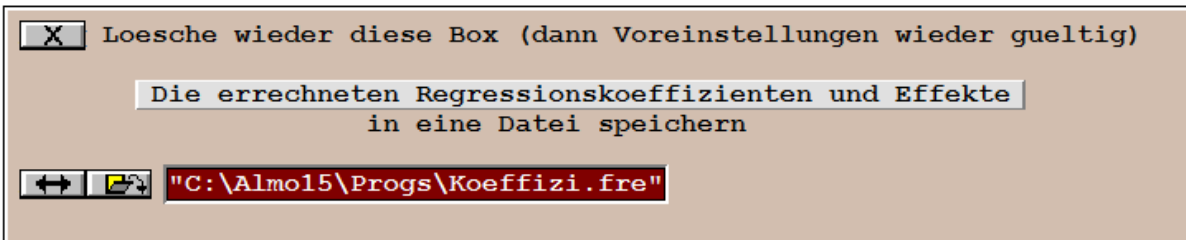
	Wertemuster				
	I	II	III	IV	.... etc.
[Geschlecht	, 1	, 2	, Alter	,48,58	]
[Einkommen	, 7200	, 3500	, Bildung,	5, 3	]

Sie schreiben in ein Eingabefeld 2 oder sogar mehrere Variable mit ihren Werten. BEACHTTE: Alle Zahlenwerte und Variablennamen werden durch Beistrich getrennt. Am Schluß des Eingabefeldes wird kein Beistrich geschrieben. Die Überschrift und die Rahmen dienen nur der "Schönheit". Sie haben keine Bedeutung für Almo.

**Eingabe-Box 14:** Option: Die errechneten Koeffizienten in eine Datei speichern

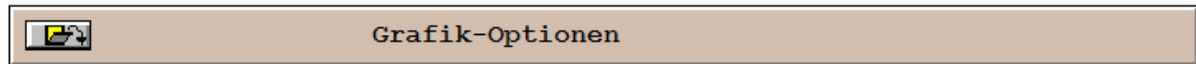


Wird die Optionsbox durch Klick auf den Knopf mit nach unten weisenden Pfeil geöffnet, dann sieht man folgendes.

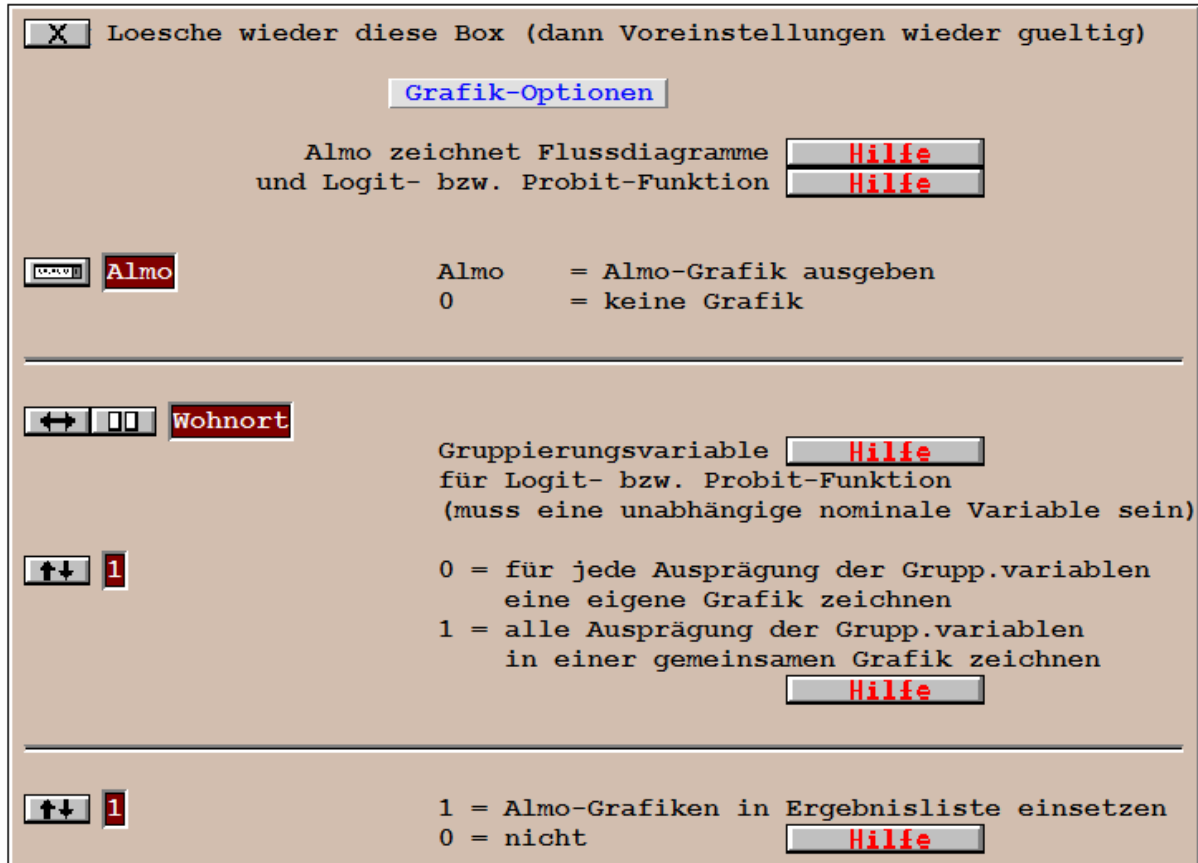


Die errechneten Koeffizienten werden mit einigen Zusatzinformationen in eine Datei gespeichert. Es besteht dann die Möglichkeit mit Prog45mt eine Prognose für die Personen einer anderen Datei zu leisten. Wir werden darauf ausführlich in Abschnitt P45.18 kommen.

## Eingabe-Box 15: Grafik-Optionen



Wird die Optionsbox durch Klick auf den Knopf mit nach unten weisenden Pfeil geöffnet, dann sieht man folgendes.



Almo zeichnet - standardmäßig, auch ohne daß diese Optionsbox aktiviert wurde - ein Flußdiagramm, in dem die Risikokoeffizienten der ursächlichen Variablen hinsichtlich der Zielvariablen eingetragen sind. Außerdem zeichnet Almo je eine logistische Funktion (bei der Probitanalyse: die Ogive) für die ursächlichen quantitativen Variablen. Dabei wird die ursächliche quantitative Variable an die x-Achse geschrieben und die Zielvariable (in unserem Beispiel: "Wahrscheinlichkeit für Rückzahlung: Ja" an die y-Achse.

Die jeweils anderen ursächlichen quantitativen Variablen werden dabei auf ihren Mittelwert gesetzt. Auch die Dummies der ursächlichen nominalen Variablen werden auf ihre Mittelwerte (d.h. Anteilswerte) gesetzt.

Bei der Interpretation der Almo-Ergebnisse in Abschnitt P45.16.1.3 werden wir ausführlich die von Almo erzeugten Kurvendiagramme besprechen und nochmals ausführlich die Eingabe in die Grafik-Optionen-Box erläutern.

*Eingabefeld 1:* Der Eintrag "Almo" bedeutet: Es werden, wie oben ausgeführt, Flußdiagramme und Kurvendiagramme erzeugt. Dies geschieht auch standardmäßig, ohne daß diese Optionsbox aktiviert wurde. Der Eintrag "0" (Null) bedeutet: Es wird keine Almo-Grafik erzeugt.

*Eingabefeld 2:* Es kann eine oder mehrere Gruppierungsvariable angegeben werden. In unserem Beispiel wird "Wohnort" als Gruppierungsvariable angegeben. Almo zeichnet dann die logistischen Funktionen (so wie oben beschrieben) für die beiden Ausprägungen "Stadt" und "Land". Es können auch Variable kombiniert werden. Die Eingabe lautet dann z.B.: „Wohnort MIT Hausbesitz“.

Beachte: Als Gruppierungsvariable können nur Variable verwendet werden, die in der Eingabe-Box "Ursächliche Variable" als nominale Variable angegeben wurden.

*Eingabefeld 3:* Betrachten wir ein Beispiel: Als Gruppierungsvariable wurde beispielsweise "Wohnort" mit den beiden Ausprägungen "Stadt" und "Land" eingesetzt.

Wenn Sie in das Eingabefeld 1 einsetzen, dann zeichnet Almo eine einzige Grafik, in der sich zwei Kurven befinden, eine für die Stadtbewohner und eine für die Landbewohner. Sie erkennen dann sehr gut den Unterschied zwischen den beiden Bewohnern.

Wenn Sie in das Eingabefeld 0 einsetzen, dann zeichnet Almo zwei Grafiken mit je einer Kurve, eine für die Stadtbewohner und eine für die Landbewohner. Da Sie dann 2 getrennte Grafiken besitzen, ist der Unterschied zwischen den Bewohnern nicht so leicht zu erkennen.

Wenn Sie eine Gruppierungsvariable angegeben haben, die viel Ausprägungen besitzt, beispielsweise 10, dann werden bei Eingabe von 1 alle 10 Kurven in einer Grafik dargestellt. Die Kurven können dann so dicht beieinander liegen, dass sie nicht mehr unterscheidbar sind. In einem solchen Fall ist es besser eine 0 in das Eingabefeld einzusetzen. Dieser Fall tritt vor allem dann auf, wenn sie mehrere Gruppierungsvariable durch das Wort MIT miteinander kombinieren. Es können dann sehr viele Ausprägungskombinationen entstehen, für die Almo je eine Kurve zeichnet.

*Eingabefeld 4:* Wenn Sie '1' eingeben, dann werden die Almo-Grafiken direkt in die Ergebnisliste eingesetzt.

Wenn Sie durch die Ergebnisliste blättern oder scrollen, dann werden Ihnen (anschliessend an Tabellen und Matrizen) auch die Almo-Grafiken gezeigt.

Wenn Sie '0' eingeben, dann können die Almo-Grafiken nur im Grafik-Editor angeschaut werden.

In der Ergebnisliste ist dann (anschliessend an Tabellen und Matrizen) nur ein Grafikknopf enthalten. Durch Klick auf diesen Knopf gelangen Sie in den Grafik-Editor, wo Ihnen die Grafik gezeigt wird.

Der Grafikknopf ist auch vorhanden, wenn Sie '1' eingeben, wenn also die Grafiken in die Ergebnisliste eingesetzt werden. Durch Klick auf den Grafikknopf in der Ergebnisliste können die Grafiken dann im Grafik-Editor bearbeitet werden und von dort durch Klick auf den Knopf "Einsetzen" in der veränderten Form wieder in die Ergebnisliste übergeben werden. Eine Bearbeitung der Grafik wird häufig notwendig sein. Man möchte beispielsweise die Balken in einem Balkendiagramm schlanker abgebildet haben, als dies Almo standardmässig tut. Oder man möchte mehr Perspektive in die Grafik bringen etc. Oder man möchte noch zusätzliche Beschriftungen einfügen.

### P45.16.1.3 Ausgabe

Ausgabe verkürzt: Almo liefert folgende Ausgabe (die wir hier nochmals etwas verkürzen).

Ergebnisse fuer 2. Auspraegung "ja" der abhaengigen Variablen V10 Rueckzahl  
(als Referenz wird die 1. Auspraegung "nein" verwendet)

unabhaengige Variable			Regress. koeff.ß	Risiko epx(ß)	relatives Risiko	Signifikanz (1-p)*100	partielle Korrelation
A1	Wohnort:	Stadt	-0.43493	0.64731	-35.26902	100.00	-0.13168
A2	Wohnort:	Land	0.43493	1.54486	54.48553	100.00	0.13168
B1	Hausbesi:	kein Hau	-0.74569	0.47440	-52.55955	100.00	-0.15825
B2	Hausbesi:	hat Haus	0.74569	2.10791	110.79059	100.00	0.15825
C1	Produkt:	Kleidung	-0.55559	0.57373	-42.62689	100.00	-0.10907
C2	Produkt:	Möbel	-0.08107	0.92213	-7.78698	50.33	-0.03584
C3	Produkt:	Technik	0.63666	1.89016	89.01633	100.00	0.14141
V4	Einkommen		0.68943	1.99257	99.25744	100.00	0.25486
V7	Rueckrate		-0.00077	0.99923	-0.07689	100.00	-0.25619
V8	Laufzeit		0.04562	1.04667	4.66727	99.22	0.06526

#### \*\*\*\*\* Erläuterung zum Regressionskoeffizient:

Bevor wir den für die inhaltliche Interpretation der Ergebnisse wichtigsten Begriff des "Risikos" erläutern, wollen wir den Begriff des "Regressionskoeffizienten" erklären.

Betrachten wir die beiden Regressionskoeffizienten für den Wohnort

A1	Wohnort:	Stadt	-0.43493
A2	Wohnort:	Land	0.43493

Das Logit-Modell lautet

$$(0) p_1 = \frac{1}{1 + e^{-(c + a(i) + b(j) + \beta_1 \cdot E + \beta_2 \cdot R + \beta_3 \cdot L)}}$$

Diese Gleichung kann so umgewandelt werden, daß auf der rechten Seite ein linearer Ausdruck steht

$$(1) \ln(p_1/p_2) = c + a(i) + b(j) + \beta_1 \cdot E + \beta_2 \cdot R + \beta_3 \cdot L$$

p<sub>1</sub>=Wahrscheinlichkeit für Kreditkauf: ja  
p<sub>2</sub>=Wahrscheinlichkeit für Kreditkauf: nein  
Natürlich gilt: p<sub>2</sub> = 1-p<sub>1</sub>  
e =e-Zahl 2.718  
c =Konstante

a(i) bezeichnet die Regressionskoeffizienten für die 2  
Dummy-Variable des Wohnorts  
b(j) bezeichnet die Regressionskoeffizienten für die 2  
Dummy-Variable des Hausbesitz

es ist also:

a<sub>1</sub>=Regressionskoeffizient für "Stadt"  
a<sub>2</sub>=Regressionskoeffizient für "Land"

E =Einkommen  
ß<sub>1</sub>=Regressionskoeffizient für Einkommen

R =Rueckrate  
ß<sub>2</sub>=Regressionskoeffizient für Rueckrate



Almo liefert folgendes Ergebnis:

Ergebnisse für 2. Ausprägung "ja" der abhängigen Variablen "Rückzahlung"  
(die Ausprägung "nein" wird als Referenzkategorie verwendet)

unabhängige Variable	Regress. Koeffiz.	"Risiko" exp(Regr.- koeffiz.)	relatives Risiko in %
c	Konstante	1.88227	-
a1	Beruf:Arbeiter	1.37706	3.96324
a2	Beruf:Angestellte	-0.92524	0.39644
a3	Beruf:Selbständige	-0.45182	0.63647
X	Einkommen	-0.37586	0.68670

Die Logit-Modell-Gleichung ist folgende:

$$(0) \quad p_1 = \frac{1}{1 + e^{-(c+a(i)+\beta \cdot x)}}$$

Man beachte:

$p_1$  ist die Wahrscheinlichkeit für die 2. Ausprägung "ja" der Zielvariablen "Rückzahlung". Mit  $p_2$  werden wir die Wahrscheinlichkeit für die Referenzkategorie "nein" bezeichnen

Diese Gleichung kann so umgewandelt werden, daß auf der rechten Seite ein linearer Ausdruck steht.

$$(1) \quad \ln(p_1/p_2) = c + a(i) + \beta X$$

$p_1$ =Wahrscheinlichkeit für Rückzahlung: ja  
 $p_2$ =Wahrscheinlichkeit für Rückzahlung: nein  
 Natürlich gilt:  $p_2 = 1-p_1$   
 $c$  =Konstante

$a(i)$  bezeichnet die Regressionskoeffizient für die 3  
 Dummy-Variablen des Berufs (die den 3 Ausprägungen entsprechen)

es ist also:

$a_1$ =Regressionskoeffizient für "Arbeiter"  
 $a_2$ =Regressionskoeffizient für "Angestellter"  
 $a_3$ =Regressionskoeffizient für "Selbständiger"

$X$  =Einkommen  
 $\beta$  =Regressionskoeffizient für Einkommen

Für einen Arbeiter in der Einkommensklasse  $X=4$  lautet also die Gleichung

$$(1a) \quad \ln(p_1/p_2) = c + a_1 + \beta X \\ = 1.88 + 1.38 - 0.38 \cdot 4$$

Gleichung 1 bzw. 1a kann so transformiert werden, daß der auf der linken Gleichungsseite stehende Logarithmus verschwindet.

$$(2) \quad p_1/p_2 = \exp(c) * \exp(a(i)) * \exp(\beta X)$$

$\exp(\dots)$  = Exponentialfunktion von ...

Für unseren Arbeiter mit Einkommen  $X=4$

$$\begin{aligned}
 (2a) \quad p_1/p_2 &= \exp(c) \quad * \quad \exp(a_1) \quad * \quad \exp(\beta * X) \\
 &= \exp(1.88) \quad * \quad \exp(1.38) \quad * \quad \exp(-0.38 * 4) \\
 &= 6.62 \quad * \quad 3.96 \quad * \quad 0.22 \\
 &= 5.7886
 \end{aligned}$$

Zuerst ist festzuhalten, daß sich die Interpretation auf die 2. Ausprägung der Zielvariablen also auf "Rückzahlung:Ja" bezieht.

$p_1$  ist also die Wahrscheinlichkeit für Rückzahlung: ja  
 $p_2$  ist also die Wahrscheinlichkeit für Rückzahlung: nein

Das Wahrscheinlichkeits-Verhältnis  $p_1/p_2$  wird in der angelsächsischen Literatur "odds" genannt.

Wenn man  $p_1$  als Gewinn-Wahrscheinlichkeit und  $p_2$  als Verlust-Wahrscheinlichkeit interpretiert, dann könnte man  $p_1/p_2$  als "Gewinn-zu-Verlust-Verhältnis" bezeichnen.

Ist die Zielvariable, wie in unserem Beispiel, dichotom, dann gilt

$$p_2 = 1 - p_1$$

Ist  $p_1=0.5$  dann ist  $p_2$  auch  $=0.5$ . Dann ist  $p_1/p_2=1$ . Das "Gewinn-zu-Verlust-Verhältnis" ist also ausgeglichen.

Ist  $p_1=0.6666..$  dann ist  $p_2=0.33333..$  Dann ist  $p_1/p_2 = 2$ . Die Gewinn-Chance ist 2 mal besser als die Verlust-Chance

In unserem Beispiel ist  $p_1/p_2=5.7886$ . Für unseren Arbeiter mit einem Einkommen von 4 gilt also, dass seine Wahrscheinlichkeit den Kredit zurückzuzahlen 5.7886 mal größer ist als ihn nicht zurückzuzahlen.

Wie groß ist dann  $p_1$  ?

Hier gilt die allgemeine Formel:

$$\begin{aligned}
 p_1 &= f / (1+f) \\
 &= 5.7886 / (1+5.7886) \\
 &= 0.853
 \end{aligned}$$

wobei  $f=p_1/p_2$

Die Wahrscheinlichkeit unseres Arbeiters mit Einkommen 4 den Kredit zurückzuzahlen ist also  $p_1=0.853$ .

Betrachten wir einige Werte von  $p_1$

$p_1$	dann ist $p_2= 1-p_1$	"Gewinn-zu-Verlust-Verhältnis" $p_1/p_2$
----	-----	-----
0.1	0.9	0.111
0.2	0.8	0.250
0.3	0.7	0.429
0.4	0.6	0.667
0.5	0.5	1
0.6	0.4	1.500
0.7	0.3	2.333
0.8	0.2	4
0.9	0.1	9

Betrachten wir nun wieder Gleichung 2 bzw. 2a. Alle Arbeiter haben - im Vergleich zum Durchschnitt aller Untersuchungspersonen - eine um den Faktor  $\exp(a_1)$

=3.96 erhöhtes Wahrscheinlichkeits-Verhältnis  $p_1/p_2$ , d.h. ihre Wahrscheinlichkeit den Kredit zurückzuzahlen ist erhöht.

Dieser Faktor wird in der Literatur gelegentlich "Risiko" genannt. Auch der Begriff "Effekt-Koeffizient" wird gelegentlich gebraucht (so bei D. Urban: Logit-Analyse, 1993, S. 40).

Wäre  $\exp(a_1)=1$ , dann würden sich die Arbeiter so verhalten wie der Durchschnitt.

Wir definieren nun als

$$\text{relatives Risiko} = (\exp(a(i)) - 1) * 100$$

Für die Arbeiter finden wir dann

$$\begin{aligned} \text{relatives Risik} &= (\exp(a_1) - 1) * 100 \\ &= (3.96 - 1) * 100 \\ &= 296 \end{aligned}$$

Wir können jetzt formulieren: Arbeiter haben ein um 296 % höheres Risiko einen Kredit zurückzuzahlen als die durchschnittliche Untersuchungsperson.

Zu beachten ist, daß die Bezugskategorie der Durchschnitt aller Untersuchungspersonen ist. Dies ist in Almo der Fall, wenn die 0,1,-1 - Kodierung der Dummies der unabhängigen nominalen Variablen verwendet wird. Dies ist die Voreinstellung in Almo (siehe Abschnitt P45.16.1.2, Eingabe-Box 9).

Wird die 0,1 - Kodierung verwendet, dann wird (standardmäßig) die letzte Dummy, in unserem Beispiel die Selbständigen, auf 0 gesetzt. Sie erscheint dann auch gar nicht in der Ergebnis-Ausgabe.

Almo liefert folgendes Ergebnis (verkürzt):

Ergebnisse für 2. Auspräg. "ja" der abhäng. Variablen "Rückzahlung"

unabhängige Variable	Regress. Koeffiz.	"Risiko" exp(Regr.-koeffiz.)	relatives Risiko
c Konstante	1.43044	-	-
a1 Beruf:Arbeiter	1.82889	6.22695	522.69462
a2 Beruf:Angestellte	-0.47341	0.62287	-37.71264
X Einkommen	-0.37586	0.68670	-31.33039

Die Selbständigen sind jetzt die Bezugskategorie. Die Arbeiter haben im Vergleich zu den Selbständigen eine um 522 % erhöhte Wahrscheinlichkeit den Kredit zurückzuzahlen und die Angestellten eine um 37.7 % reduzierte Wahrscheinlichkeit.

In Almo ist es bei der 0,1 - Kodierung möglich, entweder die erste oder die letzte Dummy zu eliminieren.

Allgemein gilt:

- Bei der 0,1 - Kodierung ist die Bezugskategorie die eliminierte Dummy.
- Bei der 0,1,-1 - Kodierung ist die Bezugskategorie der Durchschnitt aller Untersuchungspersonen.

## Risiko bei quantitativen Variablen

Betrachten wir nochmals obige Gleichung (2)

$$(2) \quad p_1/p_2 = \exp(c) * \exp(a(i)) * \exp(\beta * X)$$

Das Einkommen unseres Arbeiters ist  $X=4$ .

Der Ausdruck  $\exp(\beta * X)$  ist also  $\exp(-0.37586 * 4) = 0.22236$

Wenn sich das Einkommen dieser Person um 1 Einheit erhöht, dann ist der Ausdruck  $\exp(\beta * X) = \exp(-0.37586 * 5) = 0.15270$

Wenn wir für  $X=5$  obige Gleichung (2) für unsere Person ausrechnen, dann erhalten wir

$$p_1/p_2 = 3.9750$$

Für  $X=4$  haben wir oben errechnet

$$p_1/p_2 = 5.7886$$

So hat sich also  $p_1/p_2$  um den multiplikativen Faktor

$$3.9750 / 5.7886 = 0.68670$$

verringert. Und das ist genau das in obiger Tabelle angegebene Risiko  $\exp(\beta)$ .

Risiko-Werte unter 1 führen zu einer Verringerung von  $p_1/p_2$ . D.h.  $p_1$  wird kleiner und  $p_2$  wird größer.

Risiko-Werte über 1 führen zu einer Erhöhung von  $p_1/p_2$ . D.h.  $p_1$  wird größer und  $p_2$  wird kleiner.

Wir können nun den Begriff "Risiko" ( $=\exp(\beta)$ ) bei ursächlichen quantitativen Variablen allgemein definieren.

Nimmt die ursächliche quantitative Variable  $X$  um 1 Einheit zu, dann nimmt das Wahrscheinlichkeits-Verhältnis  $p_1/p_2$  um den multiplikativen Faktor  $\exp(\beta)$  zu.

Wir können diese Zunahme bzw. Abnahme auch in Prozentwerten ausdrücken. Sie beträgt dann  $100(\exp(\beta)-1)$ . Das ist das relative Risiko.

Betrachten wir für Arbeiter die Werte, die sich gemäß Gleichung 2 für Einkommenswerte  $X$  von 0 bis 6 ergeben.

X	$p_1/p_2$	Multiplikator
0	26.0326	
1	17.8765	0.6867
2	12.2758	0.6867
3	8.4298	0.6867
4	5.7886	0.6867
5	3.9750	0.6867
6	2.7297	0.6867

Das Wahrscheinlichkeits-Verhältnis  $p_1/p_2$  einer nachfolgenden Einkommensstufe entsteht durch Multiplikation mit  $\exp(\beta)=0.6867$  des Wahrscheinlichkeits-Verhältnis  $p_1/p_2$  der vorhergehenden Einkommensstufe.

**\*\*\*\*\* Erläuterung: Signifikanz**

Alle Dummies, mit Ausnahme von "Möbel" haben eine signifikante, d.h. überzufällige Wirkung auf die abhängige Variable Rückzahlung: ja.

**\*\*\*\*\* Erläuterung: Partielle Korrelation**

Die partiellen Korrelationskoeffizienten ermöglichen es, die Wirkungsstärke der unabhängigen Dummies und der unabhängigen quantitativen Variablen zu vergleichen.

Wir sehen, daß die Rückzahlungsrate und das Einkommen, die am stärksten wirkenden Variablen sind. Hingegen sind "Möbel" und "Laufzeit" die am schwächsten wirkenden Variablen.

Die partiellen Korrelationskoeffizienten sind unabhängig von der für die Variablen jeweils gewählten Maßeinheit.

Beobachtete und durch das Modell reproduzierte (prognostizierte) Wahrscheinlichkeiten (in %)

die unabhaengigen nominalen Variablen sind

A = V1 Wohnort

B = V6 Hausbesitz

C = V9 Produkt

ihre Auspraegungen werden mit 1,2,3,... durchnummeriert

die unabhaengigen quantitativen Variablen sind

quant1 = V4 Einkommen

quant2 = V7 Rueckrate

quant3 = V8 Laufzeit

beo1 ... = Auspraegung 1 der abhaengigen Variablen

1=aufgetreten 0=nicht aufgetreten

repl ... = reproduzierte (prognostizierte) Wahrscheinlichkeit fuer das Auftreten der Auspraegung 1 in der abhaeng. Variablen

Nr.	A	B	C	quant1	quant2	quant3	beo1	beo2	repl	rep2
1	1	1	1	3.238	5211.00	20.000	0	1	65.5	34.5
2	2	1	2	3.477	4236.00	14.000	0	1	20.7	79.3
3	2	1	1	2.565	5545.00	14.000	1	0	68.3	31.7
4	2	1	3	4.164	4748.00	20.000	0	1	8.2	91.8
5	2	1	3	1.864	1568.00	10.000	0	1	5.6	94.4
6	2	1	2	2.723	3772.00	11.000	1	0	26.0	74.0
7	2	2	3	0.000	4901.00	21.000	0	1	27.6	72.4
8	1	1	3	2.839	5411.00	13.000	0	1	54.9	45.1
9	2	1	1	4.965	2409.00	15.000	0	1	3.4	96.6
10	2	1	2	5.000	3124.00	21.000	0	1	2.7	97.3
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.

**\*\*\*\*\* Erläuterung: Reproduzierte (prognostizierte) Wahrscheinlichkeiten**

Wurde in Eingabe-Box 11 (Option: Prognosewerte ermitteln) angegeben, daß für x Personen Prognosewerte ermittelt werden sollen, so werden diese nun hier ausgegeben.

Zuerst wird die laufende Nummer angegeben. "1" bezeichnet also die Person 1 etc.

In den Spalten A, B, C werden dann die Werte dieser Person in den ursächlichen nominalen Variablen angegeben, also in Wohnort, Hausbesitz, Produkt.

In den Spalten "quant1" bis "quant3" werden die Werte der Person in den ursächlichen quantitativen Variablen angegeben, also in Einkommen, Rückrate, Laufzeit.

Die Spalte "beo1" bezeichnet dann die tatsächliche 1. Ausprägung der Zielvariablen, also Rückzahl: nein.

Die Spalte "beo2" bezeichnet dann die tatsächliche 2. Ausprägung der Zielvariablen, also Rückzahl: ja.

Die 1. Person hat ihren Kredit zurückgezahlt, also hat sie in beo1 den Wert 0 und in beo2 den Wert 1

Mit "rep1" wird die von der Logitanalyse prognostizierte Wahrscheinlichkeit angegeben, daß sich die Person in der Ausprägung 1 der Zielvariablen (also Rückzahl: nein) befindet

und mit "rep2" die Wahrscheinlichkeit, daß sich die Person in der Ausprägung 2 der Zielvariablen (also Rückzahl: ja) befindet.

Bei der 1. Person irrt die Logitanalyse. Sie gibt mit 65.5% eine höhere Wahrscheinlichkeit für "Rückzahl: nein" an. Tatsächlich befindet sich aber Person 1 in der Ausprägung "Rückzahl: ja".

Trefferhäufigkeiten bei Individualdaten  
fuer abhaengige Variable V10 Rueckzahl

		tatsaechlich		prognostiziert absolut	
		1	2	1	2
		nein	ja	nein	ja
nein	1	286	0	155	131
ja	2	0	714	74	640

		prognostiziert relativ		erwartet Zufall	
		1	2	1	2
		nein	ja	nein	ja
nein	1	152.2	133.8	81.8	204.2
ja	2	133.8	580.2	204.2	509.8

absolut: Chi-Quadrat(1) =208.032      Signifikanz 100\*(1-p) = 100.000  
relativ: Chi-Quadrat(1) =118.848      Signifikanz 100\*(1-p) = 100.000

#### \*\*\*\*\* Erläuterung: Trefferhäufigkeit

Im Verlauf der Logit-Analyse wird für jede Person die Wahrscheinlichkeit prognostiziert, daß sie der Gruppe der Nicht-Rückzahler bzw. der Gruppe der Rückzahler angehört. Ist die Wahrscheinlichkeit den Nicht-Rückzahlern anzugehören größer als die für die Rückzahler, dann wird sie der Gruppe der Nicht-Rückzahler zugerechnet – und entsprechend umgekehrt.

In der 1. Tabelle, überschrieben mit "tatsächlich", erkennen wir, daß 286 Personen Nicht-Rückzahler sind und 714 Rückzahler.

In der 2. Tabelle, überschrieben mit "prognostiziert absolut", sehen wir, daß 155 Nicht-Rückzahler vom Logit-Modell richtig identifiziert und 131 falsch identifiziert wurden. Von den 714 Rückzahler werden 640 richtig und 74 falsch identifiziert.

Wir können nun diese Ergebnis vergleichen mit dem, das wir aus dem Allgemeinen Linearen Modell mit Prog45mf erhalten haben. Siehe Abschnitt P45.15.1.6.

		tatsaechlich	davon richtig prognostiziert	
			Allg. lin. Mod.	Logit-Analyse
Gruppe 1	Rückzahlung nein	286	148 (51.7 %)	155 (54.2 %)
Gruppe 2	ja	714	650 (91.0 %)	640 (89.6 %)

Die beiden Ergebnisse sind fast gleich. Es ist nicht möglich zu sagen, dass die eine Methode besser sei als die andere.

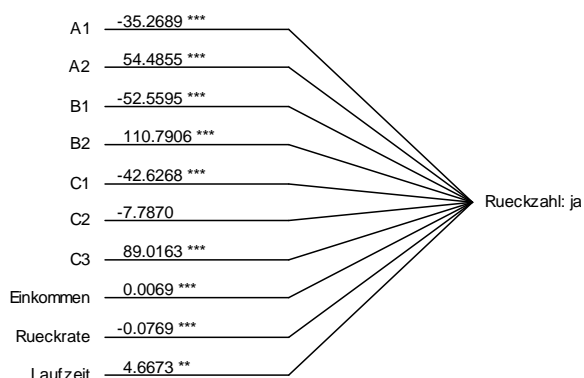
In der 4. Tabelle, überschrieben mit "erwartet Zufall" wird uns gezeigt, wie die Prognose wäre, wenn wir zufällig aus den 1000 Personen 286 Nicht-Rückzahler auswählen würden. Dann wären davon nur 81.8 (gerundet 82) Personen richtig getroffen worden. Die Trefferquote wäre nur  $100 \cdot 81.8 / 286 = 28.6$  %. Das Logit-Modell hat jedoch eine beinahe doppelt so große Trefferquote von 54.2 %.

Almo gibt uns nun auch noch aus, ob die Logit-Prognose im Vergleich zur "Zufalls-Prognose" signifikant verschieden ist. Es wird ein Chi-Quadrat-Wert von 208.032 gefunden, der mit 100 % signifikant ist. Diese 100 % sind durch Runden entstanden. Der tatsächliche Signifikanzwert ist 99.99999....

Die 3. Tabelle, überschrieben mit "prognostiziert relativ" und der dazu gehörende Chi-Quadrat-Wert, bezeichnet mit "relativ: Chi-Quadrat(1) = 118.848 ...." wird hier nicht erläutert. Siehe dazu Almo-Handbuch Teil 4, P22.

Almo erzeugt nun noch ein Flußdiagramm der relativen Risiko-Koeffizienten. Es zeigt uns nochmals die Zusammenhänge zwischen den unabhängigen und der abhängigen Variablen.

relative Risikoeffizienten  
 fuer unabhangige Variable  
 A Wohnort: A1=Stadt A2=Land  
 B Hausbesitz: B1=kein Haus B2=hat Haus  
 C Produkt: C1=Kleidung C2=Mobel C3=Technik



Auf den Strichen stehen die relativen Risikoeffizienten. Die Sterne hinter den Koeffizienten symbolisieren die Signifikanz 1 (p-100).

3 Sterne = ist mit 99.9% signifikant  
 2 Sterne = ist mit 99.0% signifikant

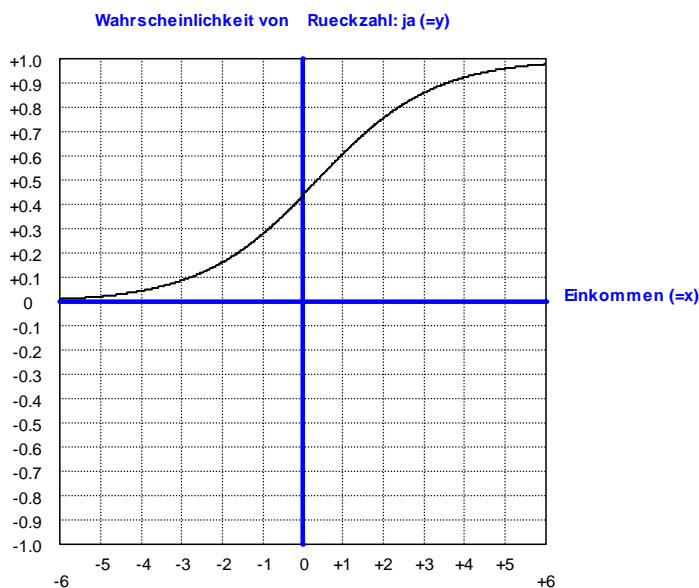
1 Stern = ist mit 95.0% signifikant  
Kein Stern = Signifikanz unter 95.

Die den Sternen zugeordneten Signifikanzwerte können im Grafikeditor auf der rechten Seite beliebig gesetzt werden.

Almo zeichnet nun je eine logistische Funktion für die 3 ursächlichen quantitativen Variablen. Dabei wird die ursächliche quantitative Variable an die x-Achse geschrieben und die abhängige Variable "Wahrscheinlichkeit für Rückzahlung:Ja" an die y-Achse.

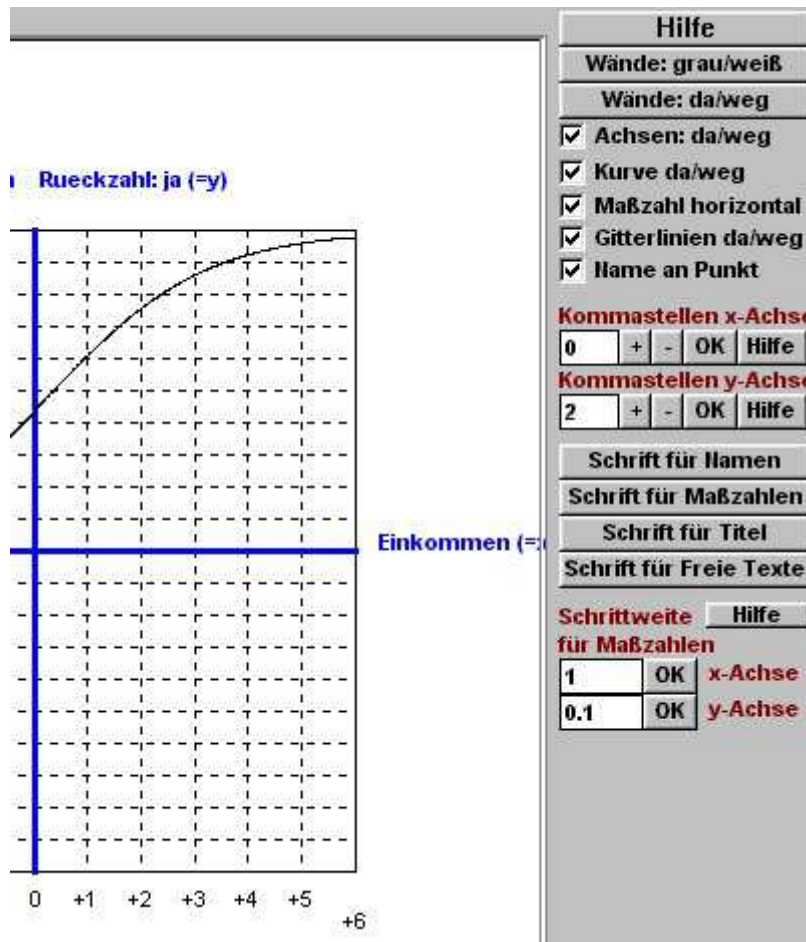
Zuerst wird der logistische Zusammenhang zwischen "Einkommen" (x-Achse) und "Wahrscheinlichkeit für Rückzahlung:Ja" (y-Achse) gezeichnet.

Logistische Funktion  
 $Y = 1/(1+e^{*(-0.24+0.69*x)})$



Zu beachten ist, daß das Einkommen in der Umkodierungsbox mit 10000 dividiert wurde, so daß an der x-Achse jetzt die Werte 1 bis 6 stehen.

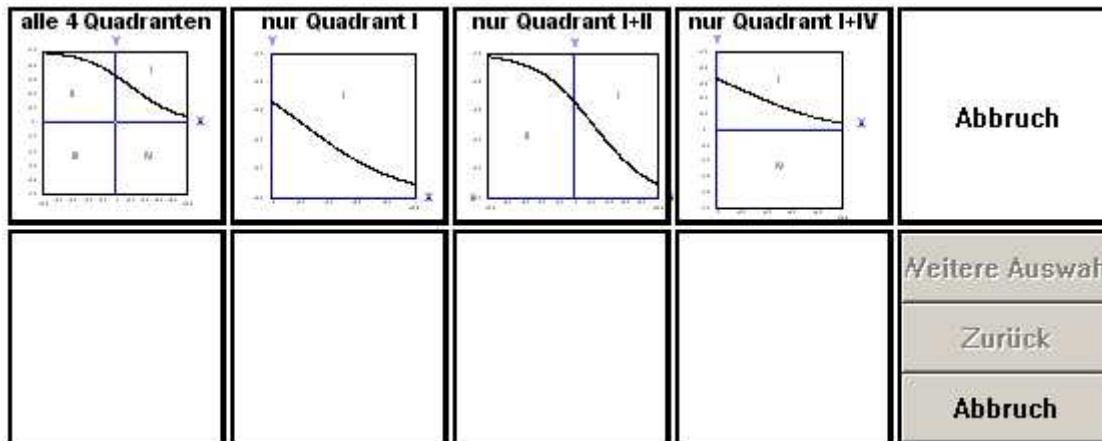
Wir haben diese Grafik im Almo-Grafik-Fenster etwas "verschönert". Wir zeigen einen Ausschnitt von der rechten Hälfte des Grafikfensters.



Wir haben folgende Aktionen vorgenommen:

1. Zuerst haben wir auf den Knopf "Wände: grau / weiß" geklickt. Dadurch wurde der Hintergrund weiß
2. Die Checkbox "Maßzahl horizontal" wurde selektiert. Dadurch werden alle Maßzahlen horizontal geschrieben.
3. Die "Kommastellen an der x-Achse" wurden auf 0 gesetzt und an der y-Achse auf 1. Dadurch werden an die x-Achse Ganzzahlwerte geschrieben und an die y-Achse Kommazahlen mit einer Stelle.
4. Bei der "Schrittweite für Maßzahlen" haben wir für die x-Achse "1" und für die y-Achse "0.1" eingesetzt. An der x-Achse stehen dann die Ziffern 1, 2, .... 6 und an der y-Achse 0.1, 0.2, .... 1.0

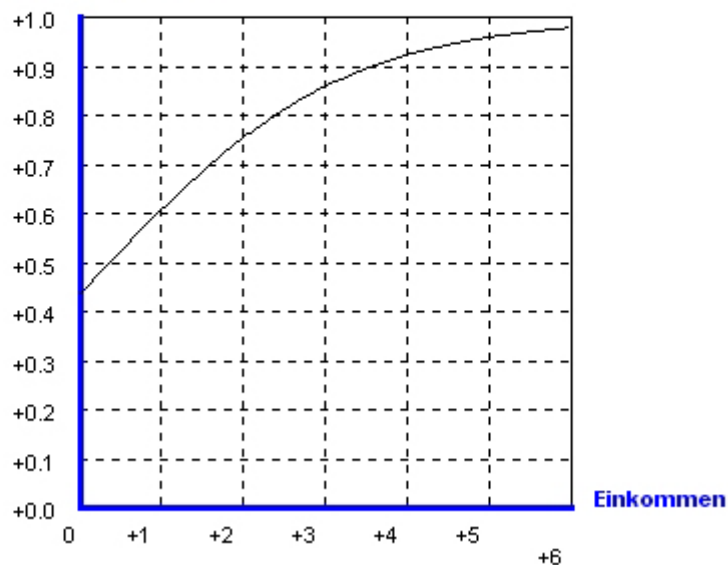
Der jeweils negative Ast der x- und y-Achse ist natürlich irrelevant. Wenn Sie im Grafik-Editor auf der linken Seite auf den Knopf „Diverse Positionen“ klicken, dann bietet Ihnen Also folgende Auswahl an:



Sie können sich nur den positiven Ast der Kurve zeigen lassen. Klicken Sie auf das 2. kleine Fenster. Also beschränkt dann die Grafik auf den 1. Quadranten des Koordinatensystems. Sie sehen dann folgende Grafik:

Logistische Funktion  
 $Y = 1 / (1 + e^{-( -0.24341 + 0.68943 * X )})$

Wahrscheinlichkeit von Rueckzahl: ja



Wir erkennen, daß beispielsweise bei einem Einkommen von 10000 (1 Einheit) die Wahrscheinlichkeit, daß der Kredit zurück bezahlt wird, ca 0.6 ist.

Die jeweils anderen ursächlichen quantitativen Variablen werden dabei auf ihren Mittelwert gesetzt. Auch die Dummies der ursächlichen nominalen Variablen werden auf ihre Mittelwert gesetzt. Dieser entspricht dem Anteilswert der Ausprägungen.

Die Logit-Modell-Gleichung für unser Beispiel ist nachfolgend in (1) angegeben, die Gleichung, die Almo zeichnet, in (2)

Gleichung (1)	Gleichung (2)		
$p=1/(1+\exp(-V))$	$p=1/(1+\exp(-V))$		
wobei	wobei		
$V= \beta_1 * E$	$V= \beta_1 * E$	Einkommen	Variable an der x-Achse
$+ \beta_2 * R$ $+ \beta_3 * L$	$+ \beta_2 * MR$ $+ \beta_3 * ML$	Rückrate Laufzeit	
$+ \beta_4 * Ws$ $+ \beta_5 * Wl$	$+ \beta_4 * MWS$ $+ \beta_5 * MWL$	Wohnort: Stadt Wohnort: Land	alle anderen werden auf ihren Mittelwert gesetzt
$+ \beta_6 * Hk$ $+ \beta_7 * Hh$	$+ \beta_6 * MHk$ $+ \beta_7 * MHh$	Hausbesitz: kein Haus Hausbesitz: hat Haus	
$+ \beta_8 * Pk$ $+ \beta_9 * Pm$ $+ \beta_{10} * Pt$	$+ \beta_8 * MPk$ $+ \beta_9 * MPm$ $+ \beta_{10} * MPt$	Produkt: Kleidung Produkt: Möbel Produkt: Technik	
+const	+const	Konstante	Konstante

$\exp(-V)$  =das ist "e hoch -V"

p = "Wahrscheinlichkeit für Rückzahlung:Ja"

MR, ML =Mittelwert aus Rückrate, Laufzeit

MWs, MWL =Mittelwert für Wohnort: Stadt bzw. Land

MHk, MHh =Mittelwert für Hausbesitz: kein Haus bzw. hat Haus

MPk, MPm, MPt =Mittelwert für Produkt: Kleidung bzw. Möbel bzw. Technik

$\beta_1$  bis  $\beta_3$  =Regressionskoeffizient der ursächlichen quantitativen Variablen

$\beta_4$  bis  $\beta_{10}$  =Regressionskoeffizient der Dummies der ursächlichen nominalen Variablen

const =Konstante

Für die ursächlichen quantitativen Variablen Rückrate und Laufzeit ist in (2) deren Mittelwert eingesetzt worden. Ebenso für die Dummies der unabhängigen nominalen Variablen. Das entspricht der Einsetzung einer "Durchschnittsperson".

Wir können also etwas verkürzt formulieren:

In der Almo-Grafik wird für die "Durchschnittsperson" der logistische Zusammenhang zwischen Einkommen und "Wahrscheinlichkeit für Rückzahlung:Ja" gezeichnet.

Im Titel der Almo-Graphik wird Gleichung (2) angegeben. Dabei wird der Gleichungsteil

$\beta_2 * MR$ $+ \beta_3 * ML$	
$+ \beta_4 * MWS$ $+ \beta_5 * MWL$	
$+ \beta_6 * MHk$ $+ \beta_7 * MHh$	die anderen Variablen die auf ihren Mittelwert gesetzt wurden
$+ \beta_8 * MPk$ $+ \beta_9 * MPm$	

+β10*MPT		Konstante
+const		

aus obiger Gleichung in den Zahlenwert 0.24 zusammengefaßt. So entsteht für "Einkommen" (x-Achse) versus "Wahrscheinlichkeit für Rückzahlung:Ja" (y-Achse) folgende Gleichung

$$y = 1 / (1 + e^{*-(-0.24 + 0.69*x)})$$

Betrachten wir die allgemeine Formel für die logistische Funktion

$$y = 1 / (1/a + e^{*-(b + c*x)})$$

Bei der Logit-Analyse ist a=1. Die Gleichung vereinfacht sich also zu

$$y = 1 / (1 + e^{*-(b + c*x)})$$

Das negative Vorzeichen vor dem Exponenten  $-(b+c*x)$  kann auch weggelassen werden. Die logistische Funktion ist dann einfach um die Senkrechte durch ihren Wendepunkt gedreht.

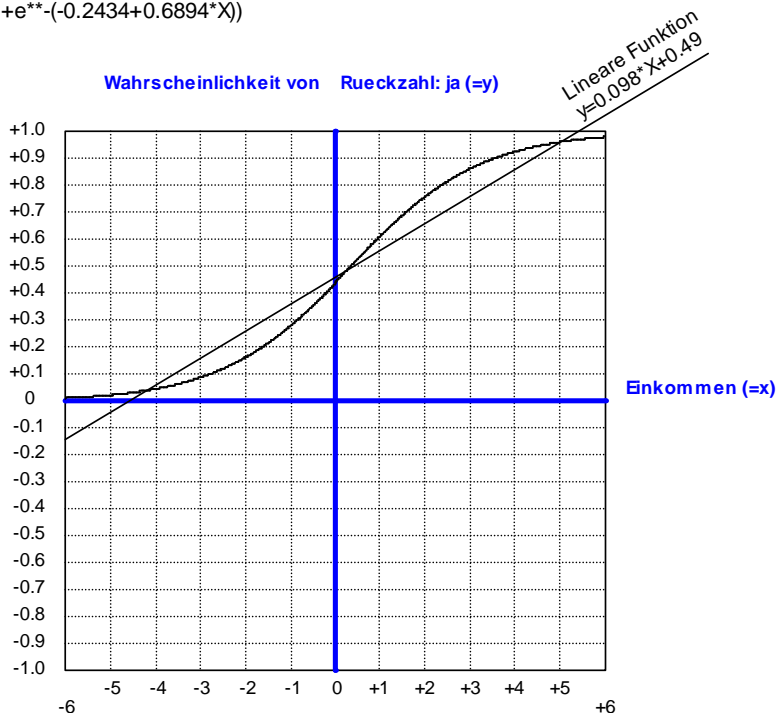
Der Parameter a bestimmt die Obergrenze, der sich die logistische Funktion annähert. Bei der Logit-Analyse ist dies 1.0.

Der Parameter b bestimmt die horizontale Lage der Kurve. Je größer b umso weiter links liegt die Kurve - bei positivem c. Ist c negativ dann umgekehrt. Eine Änderung von b bewirkt also eine links-rechts -Verschiebung der logistischen Kurve.

Der Parameter c bestimmt die Steilheit. Je größer c absolut ist umso steiler. Wenn das Vorzeichen von c positiv ist, dann wächst die Kurve von links nach rechts, ist es negativ, dann umgekehrt. c entspricht der Steigung bei der linearen Funktion.

Nicht uninteressant ist es, die logistische Kurve mit der linearen Funktion zu vergleichen, die das Allgemeine Lineare Modell aus Prog45mf liefert.

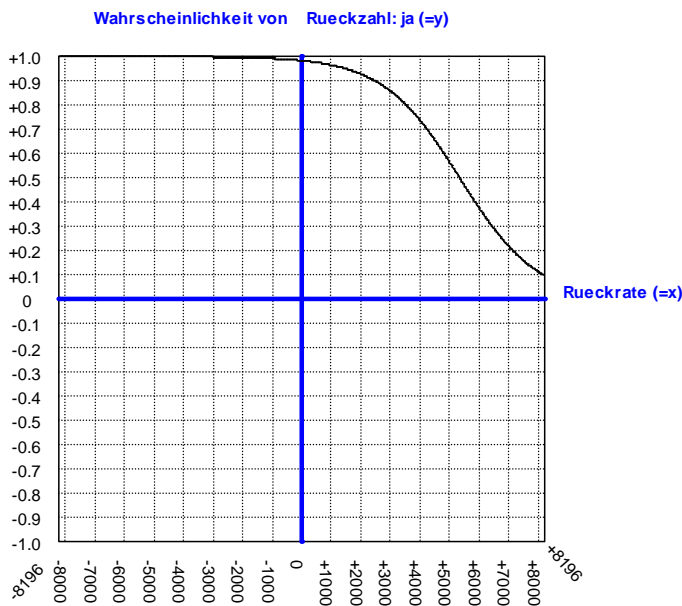
Logistische Funktion  
 $Y = 1/(1+e^{*-(-0.2434+0.6894*X)})$



Der Kurvenverlauf der beiden Kurven ist sehr ähnlich. Die "Überlegenheit" der logistischen Kurve, Wahrscheinlichkeiten größer 1, zu vermeiden, wird offenkundig. Andererseits wird auch einsichtig, dass es plausibel ist bei der linearen Funktion Wahrscheinlichkeiten größer 1 auf 1 zu setzen.

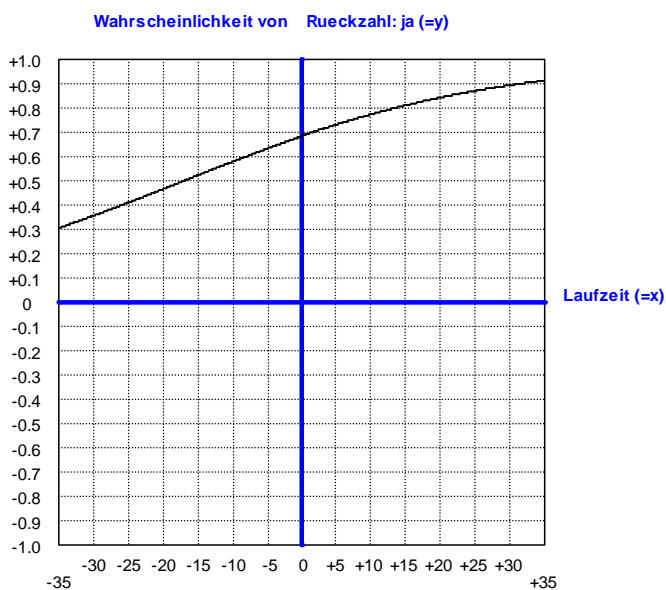
Almo zeichnet auch noch die logistische Kurve für die Rückrate und die Laufzeit hinsichtlich der "Wahrscheinlichkeit für Rückzahlung: Ja"

Logistische Funktion  
 $Y = 1/(1+e^{-(4.1-0.00077 \cdot X)})$



Wir erkennen, daß mit wachsender Rückzahlungsrate die "Wahrscheinlichkeit für Rückzahlung:Ja" abnimmt

Logistische Funktion  
 $Y = 1/(1+e^{-(0.79+0.046 \cdot X)})$



Wir erkennen, daß mit wachsender Laufzeit die "Wahrscheinlichkeit für Rückzahlung:Ja" zunimmt. Der Kurvenverlauf ist allerdings sehr flach, d.h. der Zusammenhang zwischen diesen beiden Variablen ist schwach.

### Gruppierungsvariable für logistische Kurve

Nun besteht die Möglichkeit die logistische Kurve wiederholt für die Ausprägungen einer oder mehrerer Gruppierungsvariable zu zeichnen.

Beachte: Als Gruppierungsvariable können nur Variable verwendet werden, die als unabhängige nominale Variable angegeben wurden.

Es wird beispielsweise der "Wohnort" als Gruppierungsvariable angegeben. In die Eingabe-Box "Grafik-Optionen" wird dann eingetragen.

X Loesche wieder diese Box (dann Voreinstellungen wieder gueltig)

**Grafik-Optionen**

Almo zeichnet Flussdiagramme und Logit- bzw. Probit-Funktion

Almo = Almo-Grafik ausgeben  
0 = keine Grafik

---

Gruppierungsvariable   
für Logit- bzw. Probit-Funktion  
(muss eine unabhängige nominale Variable sein)

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 dann die logistischen Funktionen (so wie oben beschrieben) für die beiden Ausprägungen des Wohnorts. Es werden also folgende Kurven gezeichnet:

1. Einkommen (x-Achse) mit "Wahrscheinlichkeit für Rückzahlung:Nein" (y-Achse) für die Städter und die Landbewohner
2. Rückrate (x-Achse) mit "Wahrscheinlichkeit für Rückzahlung:Nein" (y-Achse) für die Städter und die Landbewohner
3. Laufzeit (x-Achse) mit "Wahrscheinlichkeit für Rückzahlung:Nein" (y-Achse) für die Städter und die Landbewohner

Die jeweils anderen ursächlichen Variablen sind dabei auf ihren Mittelwert gesetzt. Bei (1) wird also

Hausbesitz:keinHaus  
 Hausbesitz:hatHaus  
 Produkt:Kleidung  
 Produkt:Möbel  
 Produkt:Technik

Rückrate  
 Laufzeit

auf den Mittelwert bzw. Anteilswert gesetzt.

Wir wollen nur die erste Kurve betrachten

**Logistische Funktion fuer**  
 abhaengige Variable: U10 Rueckzahl: 2.Auspraegung: ja  
 unabhengige Variable: U4 Einkommen

**Gruppierungsvariable:**  
 1: gruene Linie: U1 Wohnort 1.Auspraegung: Stadt  
 2: blaue Linie: U1 Wohnort 2.Auspraegung: Land

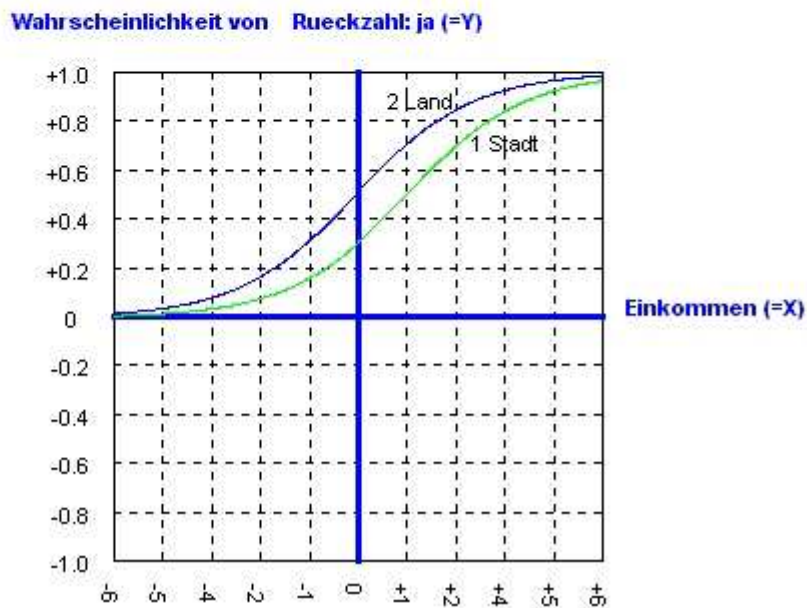
Hilfe

Logistische Funktion Grafik 02 1084090121 Grafik erzeugen und bearbeiten

Hilfe Grafik

← Grafik loeschen

Logistische Funktion  
 1: gruene Linie: Stadt  $Y = 1 / (1 + e^{*-(-0.81665 + 0.68943 * X)})$   
 2: blaue Linie: Land  $Y = 1 / (1 + e^{*-(0.053209 + 0.68943 * X)})$



Auf dem Bildschirm ist die Kurve der Landbewohner blau und die der Städter grün.

Es ist deutlich zu erkennen, daß die Kurve bei den Landbewohnern höher liegt. Sie schneidet die y-Achse bei ca. 0.5; die Kurve der Städter hingegen bei ca. 0.3. Bei jeweils gleichem Einkommen ist also die Wahrscheinlichkeit daß der Kredit zurückgezahlt wird bei den Landbewohnern deutlich höher als bei den Städtern.

Betrachten wir nochmals die Formel für die Logit-Funktion

$$y = 1 / (1 + e^{*-(b+c*x)})$$

- a) b) bestimmt die horizontale Lage der Kurve. Je grösser b umso weiter links liegt die Kurve - bei positivem c. Ist c negativ dann umgekehrt.
- b) bestimmt die Steilheit. Je grösser c absolut ist umso steiler. Vorzeichen positiv, dann wächst die Kurve von links nach rechts Vorzeichen negativ, dann umgekehrt.

Betrachten wir nun die Gleichungen der beiden oben abgebildeten Kurven

Für die Städter:  $p = 1/(1+e^{*-(-0.820 + 0.69*X)})$   
 Für die Landbewohner:  $p = 1/(1+e^{*(+0.053 + 0.69*X)})$

Der die Steilheit bestimmende Parameter c ist bei beiden Gleichungen gleich groß. Der Parameter b ist mit +0.053 bei den Landbewohnern erheblich größer als bei den Städtern mit -0.820 (wo er sogar negativ ist). Die Kurve ist also lediglich horizontal parallel verschoben.

### Kombinierte Gruppierungsvariable für logistische Kurve

Zwei oder mehrere Gruppierungsvariable können auch kombiniert werden. Dazu wird die MIT-Anweisung aus der Almo-Programmiersprache verwendet. Siehe dazu Handbuch Teil 2, Abschnitt 23. Betrachten wir ein Beispiel:

Es soll Wohnort mit Hausbesitz kombiniert werden.

In die Eingabe-Box "Grafik-Optionen" schreiben Sie in das Eingabefeld für die Gruppierungs-variable beispielsweise

Wohnort MIT Hausbesitz

BEACHTTE: Es können maximal 4 Variable durch MIT kombiniert werden.



Wir bilden nur den oberen Teil der Grafik-Optionsbox ab.

Almo erzeugt dann folgende Kombinationen in folgender Reihenfolge

- Stadt mit hat kein Haus
- Stadt mit hat Haus
- Land mit hat kein Haus
- Land mit hat Haus

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:

Wohnort, Hausbesitz

Almo zeichnet dann für die ursächlichen quantitativen Variablen (Einkommen, Rückrate, Laufzeit) je 4 Kurven, eine für die Städter, eine für die Landbewohner, eine für die Hausbesitzer und eine für die Nicht-Hausbesitzer.

Mehrere einzelne Gruppierungsvariable und mehrere durch MIT kombinierte Gruppierungsvariable können angegeben werden.

Beispiel:

Produkt, Wohnort mit Hausbesitz

Betrachten wir folgende aus den vielen Kurven, die Almo zu zeichnen hat:

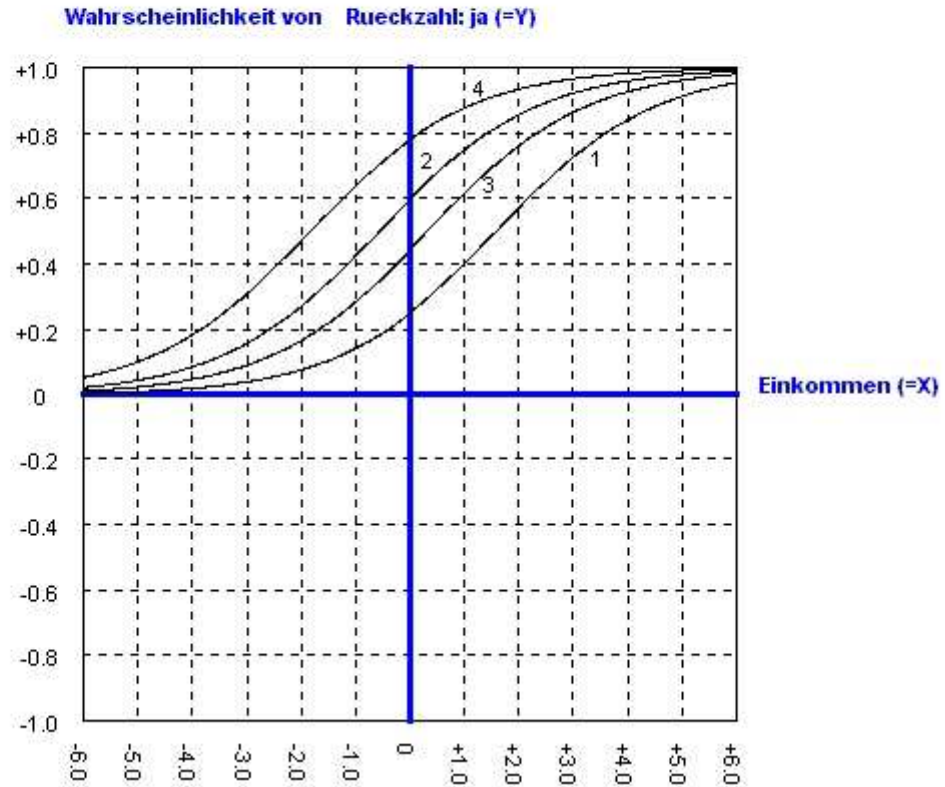
**Logistische Funktion fuer**  
abhaengige Variable: U10 Rueckzahl: 2.Auspraegung: ja  
unabhaengige Variable: U4 Einkommen

**Gruppierungsvariable:**

1: gruene	Linie:	U1 Wohnort 1.Auspraegung: Stadt mit U6 Hausbesitz 1.Auspraegung: kein Haus
2: blaue	Linie:	U1 Wohnort 1.Auspraegung: Stadt mit U6 Hausbesitz 2.Auspraegung: hat Haus
3: gelbe	Linie:	U1 Wohnort 2.Auspraegung: Land mit U6 Hausbesitz 1.Auspraegung: kein Haus
4: tuerkis	Linie:	U1 Wohnort 2.Auspraegung: Land mit U6 Hausbesitz 2.Auspraegung: hat Haus

Logistische Funktion

- 1: grüne Linie:  $Y = 1 / (1 + e^{*-(-1.0941 + 0.68943 * X)})$
- 2: blaue Linie:  $Y = 1 / (1 + e^{*-(0.39734 + 0.68943 * X)})$
- 3: gelbe Linie:  $Y = 1 / (1 + e^{*-(0.22419 + 0.68943 * X)})$
- 4: türkise Linie:  $Y = 1 / (1 + e^{*-(1.2672 + 0.68943 * X)})$



Auf dem Bildschirm sind die Kurven farblich gezeichnet.

Der Unterschied in der Lage der Kurve ist beträchtlich. Die Landbewohner mit Hausbesitz (Kurve 4) haben bei gleichem Einkommen die eindeutig bessere Rückzahlungsmoral als die Städter ohne Hausbesitz (Kurve 1).

## Wertemuster

Wenn die Wertemuster-Box aktiviert wurde, dann gibt Almo noch die Prognosewerte für die Wertemuster aus:

Prognosewerte fuer Wertemuster fuer abhaengige Variable V10 Rueckzahl

-----  
Wertemuster 1

Werte der unabhaengigen Variablen

V1 Wohnort	1 Stadt	1	
V1 Wohnort	2 Land	0	
V6 Hausbesitz	1 kein Haus	1	
V6 Hausbesitz	2 hat Haus	0	
V9 Produkt	1 Kleidung	0.204	(=Anteilswert)
V9 Produkt	2 Möbel	0.387	(=Anteilswert)
V9 Produkt	3 Technik	0.409	(=Anteilswert)
V4 Einkommen		4	
V7 Rueckrate		3459.78	(=Mittelwert)
V8 Laufzeit		14.771	(=Mittelwert)

Prognosewerte fuer abhaengige Variable

p-Wert fuer V10 Rueckzahl	1 nein	0.159271
p-Wert fuer V10 Rueckzahl	2 ja	0.840729

-----  
Wertemuster 2

Werte der unabhaengigen Variablen

V1 Wohnort	1 Stadt	0	
V1 Wohnort	2 Land	1	
V6 Hausbesitz	1 kein Haus	1	
V6 Hausbesitz	2 hat Haus	0	
V9 Produkt	1 Kleidung	0.204	(=Anteilswert)
V9 Produkt	2 Möbel	0.387	(=Anteilswert)
V9 Produkt	3 Technik	0.409	(=Anteilswert)
V4 Einkommen		3	
V7 Rueckrate		3459.78	(=Mittelwert)
V8 Laufzeit		14.771	(=Mittelwert)

Prognosewerte fuer abhaengige Variable

p-Wert fuer V10 Rueckzahl	1 nein	0.136568
p-Wert fuer V10 Rueckzahl	2 ja	0.863432

-----  
Almo gibt zuerst die Werte der ursächlichen Variablen aus, die der Benutzer eingeben hat und danach die errechneten Prognosewerte für die Zielvariable.

## P45.16.2 Logit-Analyse mit polytomer Zielvariabler

Wir wollen nun den Fall der polytomen Zielvariablen betrachten. Die Zielvariable ist also nominal und sie besitzt drei oder mehr Ausprägungen.

Wir rechnen mit den Beispieldaten „Datmin2“.

Als Zielvariable verwenden wir auf Kredit gekauftes „Produkt“. Die Ausprägungen sind (1) Kleidung, (2) Möbel, (3) Technik.

Wir setzen in das Programm Prog45m9 ein.

### Eingabe

Die Eingabe entspricht derjenigen die wir bereits in P45.16.1.1 gezeigt und erläutert haben. Wir wollen deswegen hier nur die Eingabe-Boxen abbilden und erläutern, die anders ausgefüllt sind.

The screenshot shows three input panels from the Prog45m9 software. The first panel is titled "Datei aus der gelesen wird" and contains a file path: "C:\Almo15\TESTDAT\DatMin2.dir". A "Hilfe" button is visible in the top right. The second panel is titled "Nominale Zielvariable" and contains the variable name "Produkt". The third panel is titled "Ursächliche Variable" and is divided into two sections. The top section is labeled "ursächliche nominale Variable" and contains the variables "Wohnort, Geschlecht, Hausbesitz". The bottom section is labeled "ursächliche quantitative Variable" and contains the variables "Einkommen, Alter, Bildung, Rueckrate". Each input field has a small icon to its left, likely for file selection or variable lookup.

Wir lesen Daten aus der Datei "DatMin2.dir" ein. Die Zielvariable ist "Produkt". Das ist eine polytome Variable. Sie besitzt die 3 Ausprägungen: Kleidung, Möbel, Technik.

Als ursächliche Variable verwenden wir 3 nominale und 4 quantitative Variable.

In der Eingabe-Box „Kein-Wert-Angabe und Umkodierungen“ dividieren wir das Einkommen mit 10 000, damit ein Regressionskoeffizient mit noch „sichtbarer Größe“ entsteht.

### **P45.16.2.1 Ausgabe (verkürzt)**

Almo liefert folgende Ergebnisse

Modellspezifikation:

mehrdimensionales Logit-Modell

Analysevariablen:

-----

unabhaengige nominale Variablen:

-----

V1	Wohnort	Werte-Untergrenze = 1	Obergrenze = 2
V2	Geschlecht	Werte-Untergrenze = 1	Obergrenze = 2
V4	Hausbesitz	Werte-Untergrenze = 1	Obergrenze = 2

Beachte: Für die unabhängige nominale Variable wird die 0,1,-1 Dummy-Kodierung verwendet.

#### **\*\*\*\*\* Erläuterung:**

Standardmäßig verwendet Almo für die ursächlichen nominalen Variablen die 0,1,-1 Dummy-Kodierung. In diesem Fall ist die Bezugskategorie der Durchschnitt aller Untersuchungspersonen in der ursächlichen nominalen Variablen. Wir werden noch darauf zurückkommen.

In der Eingabe-Box "Option: Auflösung der unabhängigen nominalen Variablen in Dummies" kann auch die 0,1 - Kodierung eingestellt werden. Dann wird (standardmäßig) die letzte Dummy auf 0 gesetzt. Sie erscheint dann auch gar nicht in der Ergebnis-Ausgabe In diesem Fall ist die Bezugskategorie die letzte Ausprägung der ursächlichen nominalen Variablen. Wir werden noch darauf zurückkommen.

unabhaengige quantitative Variablen:

-----

V3	Einkommen
V7	Alter
V8	Bildung
V5	Rueckrate

abhaengige nominale Variable:

-----

V6	Produkt	Werte-Untergrenze = 1	Obergrenze = 3
----	---------	-----------------------	----------------

Beachte: Zur Schaetzung wird die 1. Auspraegung der abhaeng. Variablen als Referenz verwendet.

#### **\*\*\*\*\* Erläuterung:**

Für die Interpretation der Ergebnisse benötigen wir also eine Bezugsgruppe auf Seiten der Zielvariablen und eine Bezugsgruppe je ursächliche nominale Variable. Wir werden das noch ausführlich darstellen.

Zahl der eingelesenen Datensätze = 1000

Zahl der in Analyse einbezogenen Datensätze = 1000

Maximum-Likelihood-Schätzer der Koeffizienten:

Ergebnisse fuer 2.Auspraeg. Möbel der abh. Var. V6 Produkt

unabhaengige Variable	Risiko epx( $\beta$ )	relatives Risiko	Signifikanz (1-p)*100	partielle Korrelation
A1 Wohnort: Stadt	1.39691	39.69119	99.81	0.06077
A2 Wohnort: Land	0.71586	-28.41352	99.81	-0.06077
B1 Geschlecht: m	3.85574	285.57397	100.00	0.21640
B2 Geschlecht: w	0.25935	-74.06464	100.00	-0.21640
C1 Hausbesi:kein Haus	0.37220	-62.78015	100.00	-0.18368
C2 Hausbesi: hat Haus	2.68674	168.67386	100.00	0.18368
V3 Einkommen	1.69472	69.47174	99.62	0.05509
V7 Alter	0.79091	-20.90868	100.00	-0.27341
V8 Bildung	7.88930	688.92999	100.00	0.24111
V5 Rueckrate	0.99225	-0.77459	100.00	-0.16188

Ergebnisse fuer 3.Auspraeg. Technik der abh. Var. V6 Produkt

unabhaengige Variable	Risiko epx( $\beta$ )	relatives Risiko	Signifikanz (1-p)*100	partielle Korrelation
A1 Wohnort: Stadt	1.66955	66.95524	99.96	0.07116
A2 Wohnort: Land	0.59896	-40.10371	99.96	-0.07116
B1 Geschlecht: m	12.36139	1136.13864	100.00	0.29236
B2 Geschlecht: w	0.08090	-91.91029	100.00	-0.29236
C1 Hausbesi:kein Haus	0.13604	-86.39589	100.00	-0.26543
C2 Hausbesi: hat Haus	7.35072	635.07212	100.00	0.26543
V3 Einkommen	3.44555	244.55463	100.00	0.10044
V7 Alter	0.63296	-36.70364	100.00	-0.38810
V8 Bildung	47.81511	4681.51142	100.00	0.32828
V5 Rueckrate	0.98494	-1.50572	100.00	-0.23515

Almo gibt weiter unten in der Ergebnisliste noch ein Flußdiagramm mit den relativen Risikokoeffizienten für "Möbel" und für "Technik" aus.

\*\*\*\*\* **Erläuterung: Risiko bei polytomer Zielvariabler**

Der Begriff „Risiko“, wie er im Almo verwendet wird, ist nicht durchgängig in der Literatur anzutreffen. Gelegentlich wird auch von „Effekt“ gesprochen (so bei Urban, 1993, S. 40) oder einfach von „exp( $\beta$ )“.

Zuerst ist festzuhalten, daß sich die von Almo gelieferten Ergebnisse auf die 2. und 3. Ausprägung der Zielvariablen "Produkt", also auf "Möbel" und "Technik" beziehen. Die 1. Ausprägung "Kleidung" ist die Bezugskategorie.

## Risiko bei ursächlichen nominalen Variablen

Betrachten wir die beiden obersten Zeilen

unabhaengige Variable			Risiko epx( $\beta$ )	relatives Risiko	Signifikanz (1-p)*100	partielle Korrelation
A1	Wohnort:	Stadt	1.39691	39.69119	99.81	0.06077
A2	Wohnort:	Land	0.71586	-28.41352	99.81	-0.06077

Die Stadtbewohner haben - im Vergleich zum Durchschnitt aller Personen - eine um 39.69119 % erhöhte Wahrscheinlichkeit Möbel auf Kredit zu kaufen. Die Landbewohner eine um 28.41352 % reduzierte Wahrscheinlichkeit. Ein negatives Vorzeichen beim relativen Risiko bedeutet also - im Vergleich zum Durchschnitt - eine reduzierte Wahrscheinlichkeit, ein positives eine erhöhte Wahrscheinlichkeit.

Entsprechend sind auch die relativen Risikowerte für "Geschlecht" und "Hausbesitz" zu interpretieren. Auffällig ist die hohe positive Wirkung von "Geschlecht: männlich".

Diese Interpretation gilt im Falle der 0,1,-1 - Kodierung der Dummies der ursächlichen nominalen Variablen. Dies ist die Voreinstellung in Almo. Die Bezugsgruppe ist also die "Durchschnitts-Person".

Wird die 0,1 - Kodierung verwendet, dann wird die letzte Dummy, beim Wohnort beispielsweise "Land", auf 0 gesetzt. Sie ist dann die Bezugsgruppe, mit der die Stadtbewohner verglichen werden.

Wir erhalten in diesem Fall folgendes Ergebnis (gekürzt):

Ergebnisse fuer 2. Auspraegung: Möbel der abh. Var. V6 Produkt

unabhaengige Variable			relatives Risiko
A1	Wohnort:	Stadt	95.13628
B1	Geschlecht:	m	1386.67283
C1	Hausbesi:	kein Haus	-86.14683
V3	Einkommen		69.47174
V7	Alter		-20.90868
V8	Bildung		688.92999
V5	Rueckrate		-0.77459

Die Bezugsgruppe erscheint nicht in der Ergebnis-Ausgabe. Die Stadtbewohner haben - im Vergleich zu den Landbewohnern eine um 95.13628 % erhöhte Wahrscheinlichkeit Möbel auf Kredit zu kaufen.

Nun tritt ein Interpretationsproblem auf, das nur im Falle der polytomen Zielvariablen erkennbar wird. Betrachten wir nochmals die Stadtbewohner im Vergleich zum Durchschnitt aller Personen. Unsere Interpretation muß, wenn sie vollständig und korrekt sein soll, folgendermaßen lauten:

Die Stadtbewohner haben - im Vergleich zum Durchschnitt aller Personen - eine um 39.69119 % erhöhte Wahrscheinlichkeit eher "Möbel" auf Kredit als "Kleidung" auf Kredit zu kaufen.

Die Landbewohner haben - im Vergleich zum Durchschnitt aller Personen - eine um 28.41352 % reduzierte Wahrscheinlichkeit eher "Möbel" auf Kredit als "Kleidung" auf Kredit zu kaufen.

Wir haben zwei Bezugsgruppen:

- (1) Je eine auf Seiten der ursächlichen nominalen Variablen
- (2) und eine auf Seiten der nominalen Zielvariablen.

Die erstere ist in unserem Beispiel der Durchschnitt aller Personen (bzw. bei der 0,1-Kodierung die letzte Ausprägung). Die zweite ist die erste Ausprägung der Zielvariablen, in unserem Beispiel also die "Kleidung".

Betrachten wir die ursächliche Variable des Geschlechts. Hier müssen wir interpretieren:

Männer haben - im Vergleich zum Durchschnitt aller Personen - eine um 285.57397 % erhöhte Wahrscheinlichkeit eher "Möbel" auf Kredit als "Kleidung" auf Kredit zu kaufen.

Entsprechend sind auch die Ergebnisse für Frauen, Hausbesitzer und Nicht-Hausbesitzer zu interpretieren.

Betrachten wir nun die Ergebnisse für die Ausprägung "Technik" der Zielvariablen. Für die Stadtbewohner müssen wir interpretieren:

Die Stadtbewohner haben - im Vergleich zum Durchschnitt aller Personen - eine um 66.95524 % erhöhte Wahrscheinlichkeit eher "Technik" auf Kredit als "Kleidung" auf Kredit zu kaufen.

Almo verwendet prinzipiell die erste Ausprägung als Bezugsgruppe der Zielvariablen. Die Ergebnisse für diese Bezugsgruppe, in unserem Beispiel der Kauf von "Kleidung" auf Kredit werden nicht ausgegeben. Natürlich interessieren auch diese. Wir müssen um diese Ergebnisse zu bekommen eine 2. Analyse rechnen, bei der wir beispielsweise die "Möbel" zur ersten Ausprägung und damit zur Bezugsgruppe machen. Wir erreichen dies dadurch, daß wir die Zielvariable Produkt in der Umkodierungsbox umkodieren

Loesche wieder diese Sub-Box

Eingabefelder für Umkodierungen und Kein-Wert-Angaben

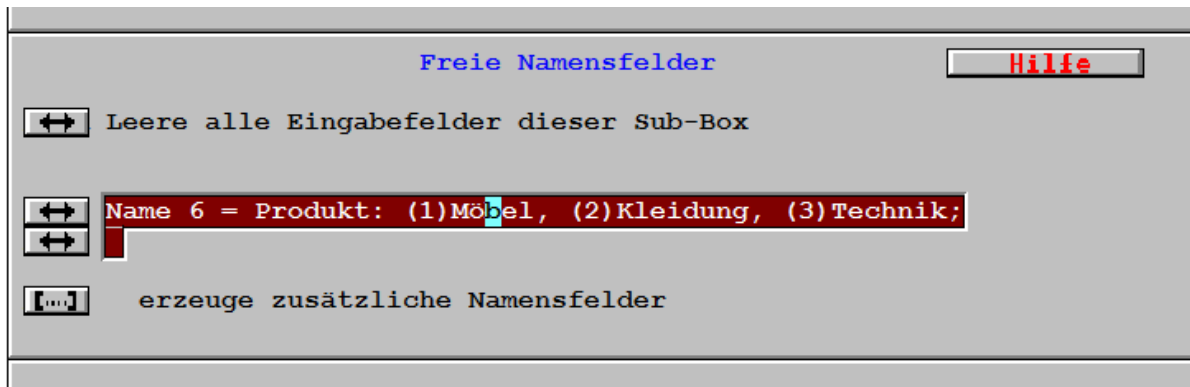
Einkommen = Einkommen / 10000;  
  Produkt (2=1; 1=2; 3=3)

erzeuge zusätzliche Felder für Umkodierungen / Kein\_Wert-Angaben

Die Ausprägung 2 (also Möbel) wird zu 1 und die seitherige Ausprägung 1 (also Kleidung) wird zu 2. Die Ausprägung 3 (also Technik) bleibt 3. Die Angabe 3=3 ist natürlich redundant und kann deswegen entfallen, so daß der Eintrag auch lauten könnte:

Produkt ( 2=1; 1=2 )

Damit die den Codeziffern zugeordneten Ausprägungsnamen wieder stimmen, schreiben wir in die Eingabe-Box "Freie Namensfelder" folgende neue Namensgebung.



Almo liefert folgendes Ergebnis für diese 2. Analyse (gekürzt)

Bezugsgruppe der Zielvariablen: Möbel

Ergebnisse fuer 2.Auspraeg. Kleidung der abh. Var. V6 Produkt

unabhaengige Variable	Risiko epx(ß)	relatives Risiko	Signifikanz (1-p)*100	partielle Korrelation
A1 Wohnort: Stadt	0.71463	-28.53748	99.82	-0.06110
A2 Wohnort: Land	1.39933	39.93348	99.82	0.06110
B1 Geschlec: m	0.25908	-74.09226	100.00	-0.21652
B2 Geschlec: w	3.85985	285.98510	100.00	0.21652
C1 Hausbesi:kein Hau	2.68735	168.73544	100.00	0.18368
C2 Hausbesi:hat Haus	0.37211	-62.78868	100.00	-0.18368
V3 Einkommen	0.58980	-41.02002	99.62	-0.05508
V7 Alter	1.26482	26.48186	100.00	0.27402
V8 Bildung	0.12654	-87.34556	100.00	-0.24132
V5 Rueckrate	1.00781	0.78143	100.00	0.16200

Eine 2. Analyse ist eigentlich nicht notwendig. Denn selbstverständlich besteht ein eindeutiger Zusammenhang. Das bedeutet, daß die Ergebnisse der 2. Analyse aus denen der 1. Analyse leicht errechnet werden können.

In der 1. Analyse ist "Kleidung" die Bezugsgruppe. Almo liefert uns dann für "Möbel" die Ergebnisse.

In der 2. Analyse ist "Möbel" die Bezugsgruppe. Almo liefert uns dann für "Kleidung" die Ergebnisse.

Wir betrachten wieder die Stadtbewohner.

Wir verwenden folgende Notation

RM = Risiko epx(ß) für Stadtbewohner aus 1. Analyse. Es ist 1.39691

RK = Risiko epx(ß) für Stadtbewohner aus 2. Analyse. Es ist 0.71463

rRK= relatives Risiko für Stadtbewohner aus 2. Analyse. Es ist -28.53748

RK kann nun leicht aus RM errechnet werden. Es ist der Kehrwert von RM

$$\begin{aligned}
 \text{RK} &= 1 / \text{RM} \\
 &= 1 / 1.39691 \\
 &= 0.71463
 \end{aligned}$$

(Da wir hier nur 5 Kommastellen verwenden, ist das Ergebnis nicht ganz exakt gleich 0.71463)

Das relative Risiko aus der 2. Analyse ist dann sehr einfach

$$\text{rRK} = 100 * ( \text{RK} - 1 )$$

$$= 100 * ( 0.71463 - 1 )$$

$$= -28.53748$$

### **Risiko bei den ursächlichen quantitativen Variablen**

Bei den ursächliche quantitativen Variablen fällt die Interpretation leichter.

Betrachten wir das Einkommen.

Nimmt das Einkommen um 1 Einheit zu, dann erhöht sich die Wahrscheinlichkeit, eher "Möbel" auf Kredit zu kaufen als "Kleidung" um 69.47174 %. Wir haben hier also nur eine Bezugsgruppe auf Seiten der nominalen Zielvariablen.

Dabei ist es nun natürlich ausschlaggebend, in welchen Maßeinheiten das Einkommen gemessen wurde. Wir haben die Variable des Einkommens mit 10 000 dividiert. Damit ist unsere Maßeinheit DM 10 000.-. Wir können also sagen, erhöht sich das Einkommen um DM 10 000.-, dann nimmt die Bereitschaft, eher "Möbel" auf Kredit zu kaufen als "Kleidung" um 69.47174 %.

Entsprechend sind auch die Ergebnisse für die anderen ursächlichen quantitativen Variablen zu interpretieren. Etwa: Nimmt die Bildung um 1 Einheit zu, dann erhöht sich die Wahrscheinlichkeit, eher "Möbel" auf Kredit zu kaufen als "Kleidung" um 688.92999 %.

#### **\*\*\*\*\* Erläuterung: Signifikanz**

Alle Dummies der ursächlichen nominalen Variablen und alle ursächlichen quantitativen Variablen haben eine signifikante, d.h. überzufällige Wirkung auf die abhängige Variable "Möbel im Verhältnis zur Kleidung".

#### **\*\*\*\*\* Erläuterung: Partielle Korrelation**

Die partiellen Korrelationskoeffizienten ermöglichen es, die Wirkungsstärke der ursächlichen Dummies und der ursächlichen quantitativen Variablen zu vergleichen. Sie sind unabhängig von der Maßeinheit mit der die jeweilige ursächliche Variable gemessen wurde.

Wir sehen, daß das Alter mit -0.27341 die am stärksten wirkende Variable ist. Hingegen sind die beiden Dummies des Wohnorts die am schwächsten wirkenden Variablen.

Beobachtete und durch das Modell reproduzierte (prognostizierte) Wahrscheinlichkeiten (in %)

die unabhangigen nominalen Variablen sind  
A = V1 Wohnort  
B = V2 Geschlecht  
C = V4 Hausbesitz  
ihre Auspraegungen werden mit 1,2,3,... durchnummeriert

die unabhangigen quantitativen Variablen sind  
quant1 = V3 Einkommen  
quant2 = V7 Alter  
quant3 = V8 Bildung  
quant4 = V5 Rueckrate

beo1 ... = Auspraegung 1 der abhaengigen Variablen  
1=aufgetreten 0=nicht aufgetreten  
repl ... = reproduzierte (prognostizierte) Wahrscheinlichkeit fuer das Auftreten der Auspraegung 1 in der abhaeng. Variablen

Nr.	A	B	C	quant1	quant2	quant3	quant4	beo1	beo2	beo3	repl	rep2	rep3
1	1	2	1	3.186	81.000	4.000	819.000	1	0	0	100.0	0.0	0.0
2	1	2	1	3.879	57.000	3.000	627.000	1	0	0	95.0	5.0	0.0
3	2	1	2	2.595	47.000	3.000	536.000	0	0	1	0.7	22.7	76.6
4	1	2	1	3.918	70.000	4.000	541.000	1	0	0	96.2	3.7	0.0
5	2	1	2	2.028	57.000	4.000	824.000	0	1	0	31.4	57.9	10.7
6	2	1	2	3.234	44.000	1.000	594.000	1	0	0	48.3	43.3	8.4
7	2	1	2	3.186	55.000	3.000	458.000	0	0	1	3.2	37.8	59.1
8	2	1	1	2.966	56.000	3.000	656.000	1	0	0	79.6	19.7	0.7
9	2	2	2	2.977	51.000	3.000	674.000	1	0	0	73.3	25.0	1.7
10	2	1	1	3.147	63.000	4.000	792.000	1	0	0	87.1	12.6	0.2
.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.

**\*\*\*\*\* Erluterung: Reproduzierte (prognostizierte) Wahrscheinlichkeiten**

Wurde in Eingabe-Box 11 (Option: Prognosewerte ermitteln) angegeben, da fur x Personen Prognosewerte ermittelt werden sollen, so werden diese nun hier ausgegeben.

Zuerst wird die laufende Nummer angegeben. "1" bezeichnet also die Person 1 etc.

In den Spalten A, B, C werden dann die Werte dieser Person in den ursachlichen nominalen Variablen angegeben, also in Wohnort, Geschlecht und Hausbesitz.

In den Spalten "quant1" bis "quant4" werden die Werte der Person in den ursachlichen quantitativen Variablen angegeben, also in Einkommen, Alter, Bildung und Ruckrate.

Die Spalte "beo1" bezeichnet dann die beobachtete, d.h. die tatsachliche 1. Auspraegung der Zielvariablen, also "Kauf von Kleidung auf Kredit".

Die Spalte "beo2" bezeichnet dann die tatsachliche 2. Auspraegung der Zielvariablen, also "Kauf von Mobel auf Kredit".

Die Spalte "beo3" bezeichnet dann die tatsachliche 3. Auspraegung der Zielvariablen, also "Kauf von Technik auf Kredit".

Die 1. Person hat Kleidung gekauft, also hat sie in beo1 den Wert 1 und in beo2 und beo3 den Wert 0

Mit "rep1" wird die von der Logitanalyse reproduzierte Wahrscheinlichkeit angegeben, daß sich die Person in der Ausprägung 1 der Zielvariablen (also "Kauf von Kleidung auf Kredit") befindet.

Mit "rep2" wird die Wahrscheinlichkeit angegeben, daß sich die Person in der Ausprägung 2 der Zielvariablen (also "Kauf von Möbel auf Kredit") befindet.

und mit "rep3" die Wahrscheinlichkeit, daß sich die Person in der Ausprägung 3 der Zielvariablen (also "Kauf von Technik auf Kredit") befindet.

Bei den ersten 10 Personen, die wir oben ausgegeben haben, stimmt die Prognose der Logitanalyse. Die maximale Wahrscheinlichkeit steht immer an der richtigen Stelle.

Trefferhaeufigkeiten bei Individualdaten  
fuer abhaengige Variable V6 Produkt

		tatsaechlich			prognostiziert absolut		
		1	2	3	1	2	3
		Kleidu	Möbel	Techni	Kleidu	Möbel	Techni
Kleidung	1	460	0	0	391	65	4
Möbel	2	0	285	0	70	163	52
Technik	3	0	0	255	8	56	191

		prognostiziert relativ			erwartet Zufall		
		1	2	3	1	2	3
		Kleidu	Möbel	Techni	Kleidu	Möbel	Techni
Kleidung	1	357.1	87.4	15.5	211.6	131.1	117.3
Möbel	2	86.9	125.6	72.5	131.1	81.2	72.7
Technik	3	15.9	72.6	166.4	117.3	72.7	65.0

absolut: Chi-Quadrat(4) =761.277      Signifikanz 100\*(1-p) = 100.000  
relativ: Chi-Quadrat(4) =487.740      Signifikanz 100\*(1-p) = 100.000

\*\*\*\*\* **Erläuterung: Trefferhäufigkeit**

Im Verlauf der Logit-Analyse wird für jede Person die Wahrscheinlichkeit prognostiziert, daß sie einer der 3 Gruppe "Kleidung", "Möbel", "Technik" angehört. Eine Person wird der Gruppe zugeordnet, für die sie die maximale Wahrscheinlichkeit besitzt.

In der 1. Tabelle, überschrieben mit "tatsächlich", erkennen wir, daß 460 Personen Kleidung, 285 Möbel und 255 Technik auf Kredit gekauft haben.

In der 2. Tabelle, überschrieben mit "prognostiziert absolut", sehen wir, daß von den Kleidungskäufern 391 Personen richtig identifiziert wurden. 65 wurden fälschlicherweise als Möbelkäufer und 4 fälschlicherweise als Technikkäufer prognostiziert.

Entsprechen ist auch die 2. und 3. Zeile dieser Tabelle zu interpretieren.

Tabelle 1 und 2 lassen sich auch vereinfacht so darstellen:

		tatsaechlich	davon richtig prognostiziert
		-----	-----
Gruppe 1	Kleidung auf Kredit	460	391 (=85 %)
Gruppe 2	Möbel auf Kredit	285	163 (=57 %)
Gruppe 3	Technik auf Kredit	255	191 (=75 %)

Wir können nun dieses Ergebnis vergleichen mit dem, das wir aus dem Allgemeinen Linearen Modell mit Prog45mf (in Abschnitt P45.15.2.1) erhalten haben.

		tatsaechlich	davon richtig prognostiziert		zufaellig richtig
		-----	Logit	ALM	-----
Gruppe 1	Kleidung auf Kredit	460	391 (85 %)	435 (95 %)	212
Gruppe 2	Möbel auf Kredit	285	163 (57 %)	6 ( 2 %)	81
Gruppe 3	Technik auf Kredit	255	191 (75 %)	224 (88 %)	65
		-----			
		1000			

Das Allgemeine Lineare Modell bringt hier zwar bei Gruppe 1 und 3 eine bessere Prognose als das Logit-Modell. Seine Prognose hinsichtlich der Gruppe 2 ist jedoch geradezu katastrophal schlecht. Sie liegt sogar weit unter der zufälligen Prognose.

In der 4. Tabelle, überschrieben mit "erwartet Zufall" wird uns gezeigt, wie die Prognose wäre, wenn wir zufällig aus den 1000 Personen 460 Personen auswählen würden, in der Hoffnung, daß sie Kleidungskäufer wären. Dann wären davon nur 211.6 (gerundet 212) Personen richtig getroffen worden.

Die Trefferquote wäre nur 46 %. Das Logit-Modell hat jedoch eine erheblich bessere Trefferquote von 85 % erbracht.

Zufällig könnten wir also folgende Trefferquoten erzielen:

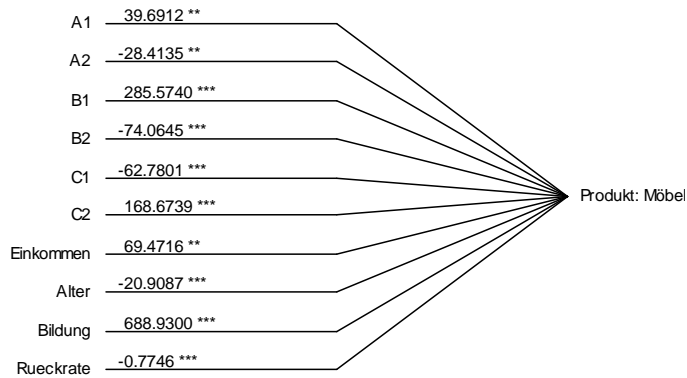
Gruppe 1	Kleidung auf Kredit	212	(=46 %)
Gruppe 2	Möbel auf Kredit	81	(=28 %)
Gruppe 3	Technik auf Kredit	65	(=25 %)

Almo gibt uns nun auch noch aus, ob die Logit-Prognose im Vergleich zur "Zufalls-Prognose" signifikant verschieden ist. Es wird ein Chi-Quadrat-Wert von 208.032 gefunden, der mit 100 % signifikant ist. Diese 100 % sind durch Runden entstanden. Der tatsächliche Signifikanzwert ist 99.99999....

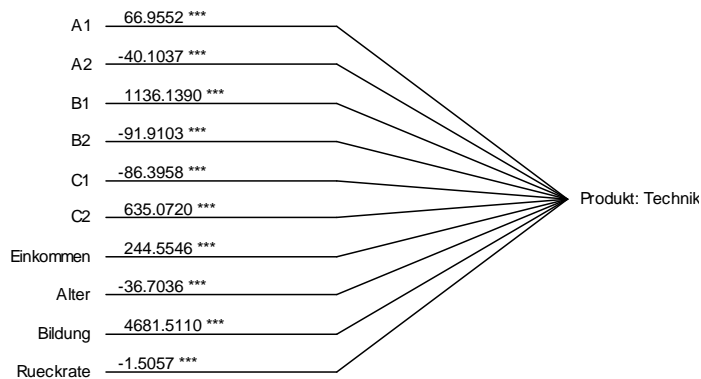
Die 3. Tabelle, überschrieben mit "prognostiziert relativ" und der dazu gehörende Chi-Quadrat-Wert, bezeichnet mit "relativ: Chi-Quadrat ...." wird hier nicht erläutert. Siehe dazu Almo-Handbuch zu P22.

Almo gibt schlußendlich noch 2 Flußdiagramme mit den relativen Risikokoeffizienten aus.

relative Risikoeffizienten  
 fuer unabhangige Variable  
 A Wohnort: A1=Stadt A2=Land  
 B Geschlecht: B1=m B2=w  
 C Hausbesitz: C1=kein Haus C2=hat Haus



relative Risikoeffizienten  
 fuer unabhangige Variable  
 A Wohnort: A1=Stadt A2=Land  
 B Geschlecht: B1=m B2=w  
 C Hausbesitz: C1=kein Haus C2=hat Haus



Auf den Strichen stehen die relativen Risikoeffizienten. Die Sterne hinter den Koeffizienten symbolisieren die Signifikanz 1 (p-100).

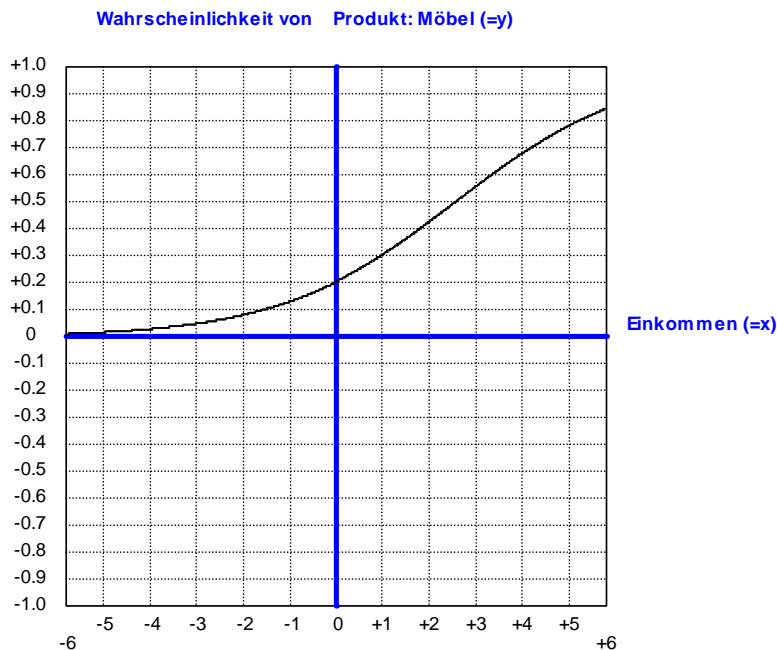
- 3 Sterne = ist mit 99.9% signifikant
- 2 Sterne = ist mit 99.0% signifikant
- 1 Stern = ist mit 95.0% signifikant
- Kein Stern = Signifikanz unter 95.

Die den Sternen zugeordneten Signifikanzwerte konnen im Grafikeditor auf der rechten Seite beliebig gesetzt werden.

Almo zeichnet nun je eine logistische Funktion fur die 3 quantitativen ursachlichen Variablen. Wir zeigen hier nur diejenige des Einkommens.

Logistische Funktion

$$Y = 1 / (1 + e^{*-(-1.3457 + 0.52752 * X)})$$



Wir haben diese Grafik im Grafikeditor "verschönert" (so wie bereits in Abschnitt P45.16.1.3 gezeigt).

Die logistische Kurve verläuft relativ flach. D.h. dass der Zusammenhang zwischen Einkommen und der Wahrscheinlichkeit "Möbel auf Kredit zu kaufen" schwach ist.

Betrachten wir eine "Durchschnitts-Person" mit einem Einkommen von 4 (also von 40 000 Geldeinheiten). Für sie müssen wir folgendermaßen interpretieren: Mit einer Wahrscheinlichkeit von ca. 0.68 (=68 %) wird diese Person eher Möbel als Kleidung (die Referenzgruppe) kaufen.

Wir haben diese Person als "Durchschnitts-Person" bezeichnet. Damit meinen wir, daß diese Person in allen ursächlichen Variablen den Mittelwert (bzw. den Anteilswert bei den nominalen Variablen) besitzt. Nur in der ursächlichen Variablen des Einkommens besitzt diese Person nicht den Mittelwert, sondern den Wert 4.

### P45.16.3 Weiterführende Hinweise

Die Logitanalyse ist als Prog22 in Almo enthalten. Prog22 enthält auch die „Probit-Analyse“, die jedoch gegenüber der Logitanalyse einige Nachteile besitzt. Die beiden Verfahren wurden von Heinrich Potuschak für Almo programmiert und von Johann Bacher und Kurt Holm an Almo adaptiert und ergänzt. Die beiden Verfahren sind im Almo-Handbuch „Teil 4, Fortgeschrittene Verfahren“, dargestellt. Zur Literatur: Eine ausführliche Darstellung der Logitanalyse ist enthalten in Urban (1993).

## Kapitel 9: Prognose leisten

Bei neuen Käufern möchten wir gerne wissen, ob sie möglicherweise ihren Kredit nicht zurückzahlen werden. Wenn wir die Werte der neuen Käufer in den ursächlichen Variablen Wohnort, Hausbesitz .. etc. kennen, dann können wir die Wahrscheinlichkeit, daß sie ihren Kredit nicht zurückzahlen, prognostizieren. Wir haben ja mit den alten Käufern (von denen wir wissen, ob sie ihren Kredit zurückbezahlt haben oder nicht) eine Analyse gerechnet, die uns die relevanten ursächlichen Variablen und ihre Koeffizienten mitgeteilt hat. Diese Koeffizienten können wir nun für die Prognose verwenden. Die Art und Weise wie diese Prognose berechnet wird, wollen wir an einem Beispiel mit einer dichotomen Zielvariable illustrieren. Selbstverständlich können in entsprechender Weise auch die Werte einer nominal-polytomen oder einer quantitativen Zielvariablen prognostiziert werden.

### ***P45.17 Schritt 12a: Werte für Zielvariable mit ALM prognostizieren***

#### **Prognose mit dem Allgemeinen Linearen Modell**

Almo liefert aus der Analyse der alten Käufer u.a. folgende Ergebnisse (siehe dazu Abschnitt P45.15.1.4 „Ausgabe etwas gekürzt“).

Effekte und Regressionskoeffizienten  
hinsichtlich der abhaengigen Variablen  
Rueckzahl: nein

	Effekte Regress.koeff -----
A1 Stadt	0.061133
A2 Land	-0.061133
B1 kein Haus	0.083115
B2 hat Haus	-0.083115
C1 Kleidung	0.087351
C2 Möbel	-0.000120
C3 Technik	-0.087232
Einkommen	-0.098402
Rueckrate	0.000103
Laufzeit	-0.007303
Konstante	0.264560

Diese Koeffizienten stammen aus einem (ungewichteten) ALM mit einer dichotomen Zielvariablen. In Abschnitt P45.15.1.0 haben wir darauf hingewiesen, daß das auf dichotome Zielvariable angewandte ALM erhebliche Defizite aufweist – wenn es auch in der Praxis in seiner Prognosefähigkeit regelmäßig nicht schlechter als jene Verfahren abschneidet, die als korrekte Alternativen empfohlen werden.

Eine Alternative ist das gewichtete ALM. Wir haben es in Abschnitt P45.15.1.7 (für den dichotomen Fall) und in P45.15.2.2 (für den polytomen Fall) dargestellt.

Die Effekte und Regressionskoeffizienten, die wir aus dem gewichteten ALM erhalten, können somit anstelle obiger Koeffizienten verwendet werden.

Das gilt auch für die Koeffizienten, die wir aus einem gewichteten ALM mit nominal-polytomer Zielvariablen erhalten. Wir werden darauf später in Abschnitt P45.17.4 eingehen.

Für die Prognose nominaler Zielvariablen mit Koeffizienten, die aus der Logitanalyse kommen, haben wir ein eigenes Programm entwickelt. Wir werden es im Abschnitt P45.18 darstellen.

Betrachten wir einen neuen Käufer mit folgenden Ausprägungen in den ursächlichen Variablen:

Wohnort: Stadt  
 Hausbesitz: nein  
 Produkt: Kleidung  
 Einkommen: 3.2384\*  
 Rückrate: 5211  
 Laufzeit: 20

\*Beachte: Wir haben das Einkommen mit 10 000 dividiert.

Mit den oben angegebenen Effekten und Regressionskoeffizienten können wir nun die Wahrscheinlichkeit berechnen, daß diese Person den Kredit nicht zurückbezahlt.

		Wert	Koeffizient	Wert * Koeffizient
		----	-----	-----
Wohnort	Stadt	1	0.061133	0.061133
	Land	0	-0.061133	0
Hausbesitz	nein	1	0.083115	0.083115
	ja	0	-0.083115	0
Produkt	Kleidung	1	0.087351	0.087351
	Möbel	0	-0.000120	0
	Technik	0	-0.087232	0
Einkommen		3.2384	-0.098402	-0.318665
Rueckrate		5211	0.000103	0.536733
Laufzeit		20	-0.007303	-0.14606
Konstante		1	0.264560	0.264560
-----				
Summe (=Wahrscheinlichkeit)				0.568167

Dieser neue Käufer hat eine Wahrscheinlichkeit von 0.568167 den Kredit nicht zurück zu zahlen.

Mit nachfolgend dargestelltem Programm können wir für neue Käufer, von denen noch nicht bekannt ist, ob sie ihren Kredit zurückzahlen oder nicht, die Wahrscheinlichkeit der Rückzahlung bzw. Nicht-Rückzahlung prognostizieren:

## P45.17.1 Eingabe

**Prog45mp.Msk**  
**Prognose leisten**  
mit dem Allgemeinen Linearen Modell  
Werte der Zielvariablen prognostizieren

Das Programm setzt voraus,

1. daß zuvor mit Prog45mf (Allgemeines Lineares Modell) eine Analyse gerechnet wurde, bei der die errechneten Koeffizienten in eine Datei gespeichert wurden
2. dass in das vorliegende Prognoseprogramm Prog45mp dieselben ursächlichen Variablen eingesetzt werden können, wie jene, mit denen Prog45mf gerechnet wurde

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

1    
Vereinbare Variable=  ;

2  Option: Weitere Vereinbarungen - nur wenn Almo dazu auffordert

3    
  "C:\Almo7\TESTDAT\DMProgn. nam"  
        zeige = Namensdatei in Output zeigen  
leer = nicht

4    
    
 erzeuge zusätzliche Namensfelder

5   
Name  =Zielvariable;  
|  
|      Geben Sie hier eine freie Variablennummer ein

6    
 "C:\Almo7\TESTDAT\DMProgn. dir"

7   
 "C:\Almo7\Testdat\Koeffiz. fre"

8

**Ursächliche Variable**

ursächliche nominale Variable

**Wohnort, Hausbesitz, Produkt**

Zahl der Ausprägungen

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

---

ursächliche quantitative Variable

**Einkommen, Rueckrate, Laufzeit**

9

**Die zu prognostizierende Zielvariable**

Die Zielvariable kann quantitativ oder (exklusiv) nominal sein  
 Klicken Sie auf den Knopf mit dem nach unten weisenden Pfeil

Zielvariable ist quantitativ  
 es sind mehrere möglich  
 siehe dazu ----->

**Zielvariable** Zielvariable ist nominal (nur eine möglich)  
   Zahl ihrer Ausprägungen

10

Option: Ein- und Ausschliessen von Untersuchungseinheiten

11

Loesche wieder diese Box

**Umkodierungen und Kein-Wert-Angaben**

Umkodierungen   
 Kein\_Wert-Angabe

**Einkommen = Einkommen / 10000;**

erzeuge zusätzliche Felder für Umkodierungen / Kein\_Wert-Angaben

---

Kontrollieren, ob Umkodierung so erfolgt wie gewünscht

diese Variablen ...

**Einkommen**

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

12

Option: Prognosewerte in Datei schreiben

13

Wir haben eine Datei mit 20 neuen Kunden, die bei uns auf Kredit kaufen wollen. Es soll prognostiziert werden, ob sie den Kredit zurückbezahlen werden oder nicht. Von dieser Prognose machen wir es abhängig, ob diese neuen Kunden beliefert werden oder nicht.

Die Datei der neuen Kunden (mit dem Namen "DMPrognos.dir") ist folgende:

V1 Haus- besitz	V2 Rueck- rate	V3 Lauf- zeit	V4 Produkt	V5 Wohn- ort	V6 Geschlecht	V7 Beruf	V8 Ein- kommen	V9 Kinder- zahl
1	5211	20	1	1	2	2	32384	5
1	4236	14	2	2	1	2	34770	1
1	5545	14	1	2	2	2	25653	1
1	4748	20	3	2	1	2	41643	0
1	1568	10	3	2	1	2	18641	1
1	3772	11	2	2	2	1	27226	2
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.

Die Datei enthält jene Variable, die in Prog45mf (Ursachen für Zielvariable) als ursächliche Variable eingesetzt worden sind. Das ist notwendig, da sonst keine korrekte Prognose geleistet werden kann.

### 1. Bedingung:

Wir wollen die 1. Bedingung formulieren, die erfüllt sein muß, damit eine Prognose geleistet werden kann:

Die Datei, für deren Untersuchungseinheiten eine Prognose abgegeben werden soll, muß alle jene Variablen enthalten, die in Prog45mf als ursächliche Variable eingesetzt worden sind (um die Koeffizienten für die Rückzahlungswahrscheinlichkeit zu erhalten). In unserem Beispiel ist diese Bedingung erfüllt.

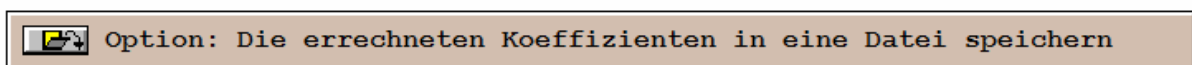
Zwar ist die Reihenfolge der Variablen in der Datei "DMPrognos.dir" in Prog45mp (das wir zur Prognose verwenden) eine andere als in der Datei "DatMining.dir" in Prog45mf (das wir zur Ermittlung der Koeffizienten verwendet haben). Aber das ist bedeutungslos.

Wir können obige Bedingung auch umgekehrt formulieren: Mit Prog45mf (Ursachen für Zielvariable) wird eine Analyse gerechnet, die jene Variablen als ursächliche enthält, die in der Datei der neuen Kunden vorhanden sind.

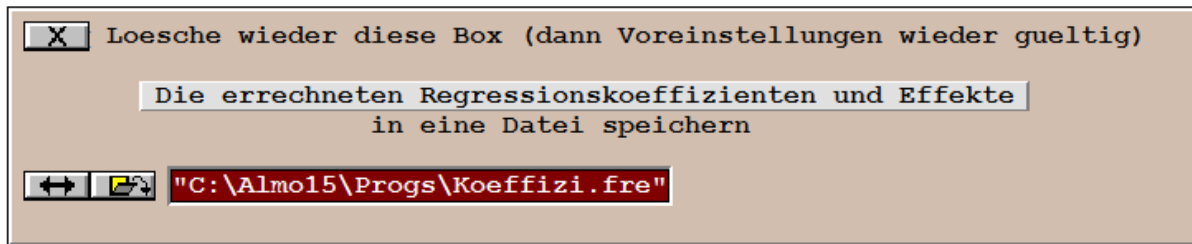
Die **2. Bedingung** ist folgende:

Die mit Prog45mf errechneten Koeffizienten müssen in eine Datei gespeichert worden sein.

In unserem Beispiel haben wir in Prog45mf (Abschnitt P45.15.1.2) die Koeffizienten in die Datei "Koeffiz.fre" gespeichert, so daß auch diese Bedingung erfüllt ist. Wir haben zu diesem Zweck in Prog45mf die Optionsbox



aktiviert und in folgender Weise ausgefüllt:



Die **3. Bedingung** ist folgende:

Im Prognoseprogramm Prog45mp müssen die Daten in genau derselben Weise manipuliert werden wie dies im Prog45mf (Ursachen für Zielvariable) geschehen ist. Es müssen also gleich sein:

1. das Ein- und Ausschließen von Untersuchungseinheiten in der Eingabe-Box "Option: Ein- und Ausschließen von Untersuchungseinheiten"
2. die Umkodierungen der Variablen in der Eingabe-Box "Kein-Wert-Angabe und Umkodierungen"
3. die Gewichtungen in der Eingabe-Box "Option: Untersuchungseinheiten gewichten"

Die **4. Bedingung** ist folgende:

Die ursächlichen Variablen, die in der nachfolgend beschriebenen Eingabe-Box 8 „Ursächliche Variable“ eingetragen sind, müssen in genau derselben Reihenfolge hintereinander stehen, wie im Prog45mf in der entsprechenden Eingabe-Box. Siehe Abschnitt P45.15.1.2.

Wir wollen nun die einzelnen Eingabe-Boxen des Prognoseprogramms Prog45mp erläutern.

## P45.17.2 Erläuterungen zu den Eingabe-Boxen von Prog45mp

### **Eingabe-Box 1:** Speicher für x Variable

Siehe "Arbeiten mit Almo-Datenanalyse-System", Abschnitt P0.1.

### **Eingabe-Box 2:** Option: Weitere Vereinbarungen

Siehe "Arbeiten mit Almo-Datenanalyse-System", Abschnitt P0.2.

### **Eingabe-Box 3:** Datei der Variablennamen

Siehe "Arbeiten mit Almo-Datenanalyse-System", Abschnitt P0.3.

Wir stellen die Namensdatei für das hier zu besprechende Prognoseprogramm Prog45mp der Namensdatei für das ALM-Programm Prog45mf gegenüber, mit dem wir die Koeffizienten ursprünglich errechnet hatten. Siehe Abschnitt P45.15.1.1.

Neue Kunden	Alte Kunden
Namensdatei für	Namensdatei für
Prognoseprogramm Prog45mp	ALM-Programm Prog45mf
-----	-----
Name1=Hausbesitz:kein Haus,hat Haus;	Name1=Wohnort:Stadt,Land;
Name2=Rueckrate;	Name2=Geschlecht:m,w;
Name3=Laufzeit;	Name3=Beruf:Selbst,Unselbst;
Name4=Produkt:Kleidung,Möbel,Technik;	Name4=Einkommen:bis 10,10-20,20 bis 30,
Name5=Wohnort:Stadt,Land;	30 bis 40,40-50;
Name6=Geschlecht:m,w;	Name5=Kinderzahl;
Name7=Beruf:Selbst,Unselbst;	Name6=Hausbesitz:kein Haus,hat Haus;
Name8=Einkommen:bis 10,10-20,20 bis 30,	Name7=Rueckrate;
30 bis 40,40-50;	Name8=Laufzeit;
Name9=Kinderzahl;	Name9=Produkt:Kleidung,Möbel,Technik;
	Name10=Rueckzahl:nein,ja;

Es ist ersichtlich, dass die beiden Dateien nicht übereinstimmen. Die Datei in Prog45mp umfasst nur 9 Variable, während die Datei in Prog45mf 10 Variable umfasst. Mehrere Variable stimmen jedoch überein. Allerdings haben sie verschiedene Variablennummern in den beiden Dateien.

Entscheidend ist nun, dass die ursächlichen Variablen, mit denen ursprünglich im ALM-Programm Prog45mf gerechnet wurden, alle auch für das Prognoseprogramm Prog45mp vorhanden sind. Das ist in unserem Beispiel auch der Fall. Dass sie verschiedene Variablennummern haben ist belanglos.

### **Eingabe-Box 4:** Freie Namensfelder

Siehe "Arbeiten mit Almo-Datenanalyse-System", Abschnitt P0.3.

### **Eingabe-Box 5:** Namen für die zu prognostizierende Zielvariable

Namen für die zu prognostizierende Zielvariable

Name 10 =Zielvariable

\_\_\_\_\_ Geben Sie hier eine freie Variablennummer ein

Hilfe

Setzen Sie in das Eingabefeld eine Variablennummern ein, die frei ist. Am besten eine Nummern, die höher ist als die Nummer der letzten eingelesenen Variablen, aber niedriger als die Zahl der vereinbarten Variablen. In unserem Beispiel umfaßt ein Datensatz 9 Variable. Also verwenden wir die freie Nummer 10.

**Eingabe-Box 6:** Datei aus der gelesen wird

The dialog box has a title bar 'Datei aus der gelesen wird' and a 'Hilfe' button. Below the title bar is a text input field containing the file path: "C:\Almo15\TESTDAT\DMProgn. dir".

Wenn die Datei der neuen Kunden noch nicht im direkten Format vorliegt (was vermutlich der Fall sein wird), dann muß sie mit Prog45md oder Prog45mh in eine Almo-Arbeitsdatei übertragen werden. Siehe dazu die Ausführungen in Abschnitt P45.1 und P45.2.

**Eingabe-Box 7:** Koeffizienten (die zuvor mit Prog45mf.Msk errechnet wurden) aus Datei einlesen

The dialog box has a title bar 'Koeffizienten (die zuvor mit Prog45mf.Msk errechnet wurden) aus Datei einlesen' and a 'Hilfe' button. Below the title bar is a text input field containing the file path: "C:\Almo15\Testdat\Koeffiz.fre".

Tragen Sie hier den Namen jener Datei ein, in welche Sie die mit Prog45mf errechneten Koeffizienten gespeichert haben. Siehe obige Ausführungen zur 2. Bedingung.

**Eingabe-Box 8:** Ursächliche Variable

The dialog box is titled 'Ursächliche Variable' and has a 'Hilfe' button. It is divided into two main sections:

- ursächliche nominale Variable:** This section contains a list of variables: 'Wohnort, Hausbesitz, Produkt'. Below the list, there are input fields for the number of categories for each variable: '2, 2, 3'. A label 'Zahl der Ausprägungen' is positioned to the right of these fields. Below this, there is a control for interactions, showing a dropdown set to '0'. Text below explains: 'Interaktionen x. Ordnung zwischen den ursächlichen nominalen Variablen bilden oder einige ausgewählte Interaktionen bilden. 0 =keine Interaktionen bilden'. A 'Hilfe' button is located at the bottom right of this section.
- ursächliche quantitative Variable:** This section contains a list of variables: 'Einkommen, Rueckrate, Laufzeit'.

Die ursächlichen nominalen und qualitativen Variablen müssen in exakt derselben Reihenfolge angegeben werden, wie in Prog45mf in Abschnitt P45.15.1.2. Auch die Interaktion x-ter Ordnung muß genau übereinstimmen.

Die Variablennamen und -nummer, die im Prog45mf und im Prognoseprogramm Prog45mp verwendet werden, müssen nicht übereinstimmen. Beispiel: In Prog45mf heißt die Variable V5 „Einkommen“. In Prog45mp entspricht dieser Variablen V8. Sie heißt auch „Einkommen“, könnte aber auch einen anderen Namen besitzen, etwa „Verdienst“. In diesem Fall würde die Eingabe in das 3. Eingabefeld lauten: Verdienst, Rueckrate, Laufzeit. Siehe hierzu auch die 1. Erläuterung im nachfolgenden Abschnitt P45.17.3.

**Eingabe-Box 9:** Die zu prognostizierende Zielvariable

Die zu prognostizierende Zielvariable

Die Zielvariable kann quantitativ oder (exklusiv) nominal sein  
 Klicken Sie auf den Knopf mit dem nach unten weisenden Pfeil

<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<p>Zielvariable ist quantitativ              es sind mehrere möglich              siehe dazu -----&gt; <span style="border: 1px solid black; padding: 2px;">Hilfe</span></p>
<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<p><span style="border: 1px solid black; padding: 2px;">Zielvariable</span>  <span style="border: 1px solid black; padding: 2px;">2</span></p> <p>Zielvariable ist nominal (nur eine möglich)              Zahl ihrer Ausprägungen</p>

In unserem Beispiel ist die zu prognostizierende Zielvariable (Rückzahlung: nein, ja) eine nominale Variable mit 2 Ausprägungen (nein, ja). Wir klicken also im 2. Eingabefeld auf den Knopf mit dem nach unten weisenden Pfeil. Also setzt dann automatisch das Wort "Zielvariable" ein. Im 3. Eingabefeld tragen wir als Ausprägungszahl 2 ein.

Ist die zu prognostizierende Zielvariable quantitativ, dann muß im 1. Eingabefeld auf den Knopf mit dem nach unten weisenden Pfeil geklickt werden. Wenn Sie mehrere quantitative Zielvariable haben, z.B. 3, dann schreiben Sie in das Eingabefeld

Zielvariable, V11, V12

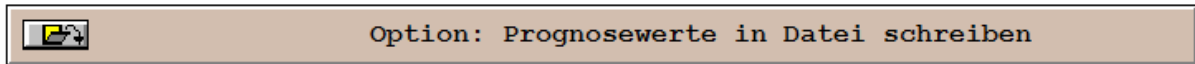
Sie fügen also dem Namen Zielvariable noch zwei Variablennummern hinzu. Diese müssen frei sein, also nicht bereits anderweitig verwendet werden.

**Eingabe-Box 10:** Option: Ein- und Ausschliessen von Untersuchungseinheiten  
 Beachte obige 3. Bedingung. Ansonsten siehe P45.1.2.

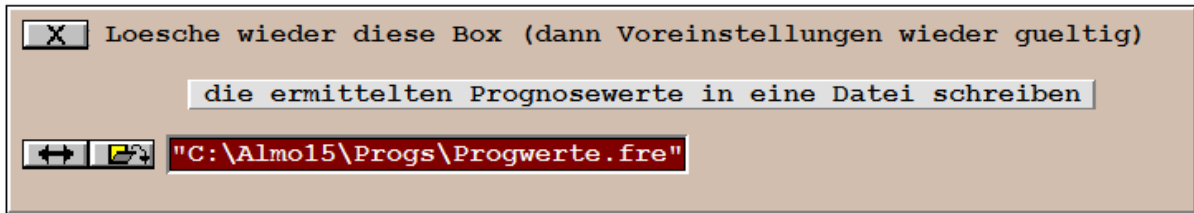
**Eingabe-Box 11:** Kein-Wert-Angabe und Umkodierungen  
 Beachte obige 3. Bedingung. Ansonsten siehe P45.1.2.

**Eingabe-Box 12:** Option: Untersuchungseinheiten gewichten  
 Beachte obige 3. Bedingung. Ansonsten siehe P45.1.2.

**Eingabe-Box 13:** Option: Prognosewerte in Datei schreiben



Wird diese Eingabe-Box aktiviert (durch Klick auf den Pfeilknopf) dann sieht man folgende Eingabe-Box:



Die vom Programm Prog45mp für die neuen Kunden ermittelten Prognosewerte können in eine Datei gespeichert werden. Der Aufbau der Datei wird im Output zu Prog45mp erläutert.

### P45.17.3 Ausgabe aus Prog45mp

Almo liefert folgenden Output:

```
***** WARNUNG
Es koennen prognostizierte Wahrscheinlichkeiten ausserhalb 0-1 auftreten
  Als Alternative ist die Logit-Analyse (Prog22m oder Prog45m9) moeglich
  Bei ihr treten keine Wahrscheinlichkeiten ausserhalb 0-1 auf
```

```
***** WARNUNG
Andere Variablennummern in Koeffizientenmatrix
```

```
Die Koeffizientenmatrix
wurde urspruenglich fuer
folgende Variable abgespeichert
  V1.01 (Dummy)
  V1.02 (Dummy)
  V6.01 (Dummy)
  V6.02 (Dummy)
  V9.01 (Dummy)
  V9.02 (Dummy)
  V9.03 (Dummy)
  V4
  V7
  V8
  V91
  V10.01 (Dummy)
  V10.02 (Dummy)

Jetzt werden folgende
Variable analysiert
  V5.01 (Dummy)
  V5.02 (Dummy)
  V1.01 (Dummy)
  V1.02 (Dummy)
  V4.01 (Dummy)
  V4.02 (Dummy)
  V4.03 (Dummy)
  V8
  V2
  V3
  V91
  V10.01 (Dummy)
  V10.02 (Dummy)

***** verschieden
***** verschieden
***** verschieden
***** verschieden
***** verschieden
***** verschieden
***** verschieden
***** verschieden
***** verschieden
***** verschieden
***** verschieden
***** verschieden
```

ALMO unterstellt, dass die jetzigen Variablennummern den urspruenglichen Var.nummern (so wie sie sich in einer Zeile gegeneinanderstehen) entsprechen

**\*\*\*\*\* Erläuterung:**

In der Datei der neuen Kunden, mit der in Prog45mp gerechnet wurde, stehen die Variablen in einer anderen Reihenfolge hintereinander als in der Datei, mit der in Prog45mf die Koeffizienten berechnet wurden. Die Variablennummern sind also verschieden. Almo bringt aus Sicherheitsgründen diese Warnung und stellt die sich entsprechenden Variablennummern aus Prog45mf und Prog45mp einander gegenüber. Der Benutzer sollte überprüfen, daß die in einer Zeile sich gegenüberstehenden Nummern dieselbe Variable meinen.

Betrachten wir die 1. Zeile. Die Variable „Wohnort“ ist im Prog45mf aus der Datei „DatMin.dir“ als V1 eingelesen worden. Im Prognoseprogramm Prog45mg wird der „Wohnort“ aus der Datei „DMPrognos.dir“ als V5 eingelesen. Die Dummies des „Wohnorts“ haben also einmal die Nummern V1.01 und V1.02 und zum anderen die Nummern V5.01 und V5.02. Sie sind jedoch identisch.

Prognosewerte fuer Variable V10 Zielvariable:

-----

Die Gruppe mit maximaler Wahrscheinlichkeit ist mit \* markiert

Datensatz	prognostizierte Wahrscheinlichkeit der Zugehoerigkeit zu Gruppe		prognostizierte Zugehoerigkeit zu Gruppe
	1	2	
1	0.570*	0.430	1
2	0.280	0.720*	2
3	0.593*	0.407	1
4	0.135	0.865*	2
5	0.105	0.895*	2
6	0.328	0.672*	2
7	0.387	0.613*	2
8	0.507*	0.493	1
9	0.025	0.975*	2
10	-0.036	1.036*	2
11	0.141	0.859*	2
12	0.657*	0.343	1
13	0.516*	0.484	1
14	-0.157	1.157*	2
15	0.493	0.507*	2
16	-0.028	1.028*	2
17	0.121	0.879*	2
18	0.164	0.836*	2
19	0.621*	0.379	1
20	0.054	0.946*	2

Bei 3 Datensatzen (=15.0 %) liegt die prognostizierte Wahrscheinlichkeit ausserhalb des zulaessigen Wertebereichs von 0 bis 1

**\*\*\*\*\* Erläuterung:**

In der 3. Spalte stehen die prognostozierten "Gruppenzugehörigkeiten" Gruppe 1 ist die Gruppe der potentiellen Nicht-Rückzahler.

Wir haben folgenden "Trick" verwendet: Unsere (simulierte) Datei der neuen Kunden umfaßt die ersten 20 Datensätze aus der Datei "DatMin.dir" (bzw. "DatMin.fre"), die wir ja für Prog45mf zur Berechnung der Koeffizienten verwendet haben. Von diesen 20 "neuen" Kunden wissen wir also, ob sie ihren Kredit zurückgezahlt haben oder nicht. Wir können also überprüfen, wie gut unsere Prognose ist. In dieser glücklichen Lage befindet man sich natürlich nicht, wenn man eine echte Datei neuer Kunden mit dem Prognoseprogramm Prog45mp analysiert.

Gruppe 1 = Nicht-Rückzahler  
 Gruppe 2 = Rückzahler

Datensatz	tatsächliche Zugehörigkeit zu Gruppe	prognostizierte Zugehörigkeit zu Gruppe	
1	2	1	*
2	2	2	
3	1	1	+
4	2	2	
5	2	2	
6	1	2	-
7	2	2	
8	2	1	*
9	2	2	
10	2	2	
11	2	2	
12	1	1	+
13	1	1	+
14	2	2	
15	1	2	-
16	2	2	
17	2	2	
18	2	2	
19	1	1	+
20	2	2	

Von den 6 tatsächlichen "Nicht-Rückzahlern" (Datensatz 3,6,12,13,15,19) haben wir 4 richtig identifiziert (durch + gekennzeichnet) und 2 fälschlicherweise als "Rückzahler" prognostiziert (durch ein Minus-Zeichen gekennzeichnet). Das ist ein gutes Ergebnis.

2 tatsächliche "Rückzahler" haben wir als "Nicht-Rückzahler" prognostiziert (durch \* gekennzeichnet) mit der Konsequenz, daß wir sie nicht auf Kredit beliefern werden.

Zu berücksichtigen ist jedoch, daß unser Prognoseprogramm einen neuen Kunden jener Gruppe zuordnet, für die es die maximale Wahrscheinlichkeit errechnet hat. Bei 2 Gruppen heißt das, daß eine Kunde schon dann zur Gruppe der "Nicht-Rückzahler" zugeordnet wird, wenn seine Wahrscheinlichkeit für diese Gruppe minimal über 0.5 liegt. Hier könnte man nun eine andere Entscheidung treffen und beispielsweise festlegen, daß ein Kunde nur dann als "Nicht-Rückzahler" eingestuft wird, wenn seine Wahrscheinlichkeit für diese Gruppe etwa 0.667 ist.

#### **P45.17.4 Prognose im Anschluß an gewichtetes ALM mit nominal-polytomer Zielvariablen**

In Abschnitt P45.15.2.2 haben wir das Programm Prog45gw für ein gewichtetes ALM mit einer nominal-polytomen Zielvariablen vorgestellt. Die nominale Zielvariable war dort das gekaufte Produkt mit den 3 Ausprägungen Kleidung, Möbek, Technik. Bei diesem Programm ist es ebenfalls möglich, die errechneten Regressionskoeffizienten und Effekte der ursächlichen Variablen als Matrix in eine Datei zu speichern. Also können wir diese Matrix auch verwenden, um mit Prog45mp Prognose zu leisten. Die Vorgehensweise entspricht dabei genau der, wie

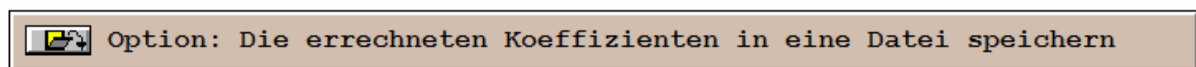
wir sie in den vorausgehenden Abschnitten P45.17.1 bis P45.17.3 dargestellt haben.

### **P45.18 Schritt 12b: Werte für Zielvariable mit Logitanalyse prognostizieren**

Die Zielvariable muss nominal (dichotom oder auch polytom) sein, damit die Logitanalyse zur Prognose eingesetzt werden kann.

Die Vorgehensweise ist, wie bei Prog45mp in Abschnitt P45.17, folgende:

1. Zuerst wird mit Prog45m9 (Logitanalyse) eine Analyse gerechnet. Siehe Abschnitt P45.16. Dabei muss die Eingabe-Box "Option: Die errechneten Koeffizienten in eine Datei speichern" geöffnet werden.



Nachdem diese Eingabe-Box geöffnet wurde sieht man folgendes



2. Wir verfügen nun über eine Datei neuer Kunden, die bei uns auf Kredit kaufen wollen. Ob sie den Kredit zurückzahlen werden oder nicht, das wissen wir noch nicht; das soll prognostiziert werden. Von diesen neuen Kunden besitzen wir die Variablenwerte aller ursächlichen Variablen, mit denen wir für die alten Kunden mit Prog45m9 die Koeffizienten errechnet und abgespeichert haben - wie in obigem Punkt 1 kurz beschrieben.
3. Die 4 Bedingungen, die wir in Abschnitt P45.17.1 formuliert haben, müssen entsprechend erfüllt sein.
4. Dann wird das Prognose-Programm Prog45mt gestartet. Es liest die gespeicherten Koeffizienten ein sowie die Datei der neuen Kunden und errechnet für jeden neuen Kunden einen Prognosewert für die Rückzahlung.

#### **P45.18.1 Eingabe in Prog 45mt**

**Prog45mt.Msk**

Prognose leisten  
mit dem Logit-Modell  
Werte der nominalen Zielvariablen prognostizieren

Das Programm setzt voraus,

1. dass zuvor mit Prog45m9 (Logitanalyse) eine Analyse gerechnet wurde, bei der die errechneten Koeffizienten in eine Datei gespeichert wurden
2. dass in das vorliegende Prognoseprogramm Prog45mt dieselben ursächlichen Variablen eingesetzt werden können, wie jene, mit denen Prog45m9 gerechnet wurde

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

- 1    
Vereinbare Variable=  ;
- 2  Option: Weitere Vereinbarungen - nur wenn Almo dazu auffordert
- 3    
  "C:\Almo7\TESTDAT\DMProgns.nam"  
        zeige = Namensdatei in Output zeigen  
leer = nicht
- 4    
    
 erzeuge zusätzliche Namensfelder
- 5   
Name  =Zielvariable;  
    |  
    └─ Geben Sie hier eine freie Variablennummer ein
- 6    
 "C:\Almo7\TESTDAT\DMProgns.dir"
- 7   
 "C:\Almo7\Testdat\Logkoeff.fre"

8

**Ursächliche Variable**

ursächliche nominale Variable

**Wohnort, Hausbesitz, Produkt**  
 2, 2, 3

Zahl der Ausprägungen

---

ursächliche quantitative Variable

**Einkommen, Rueckrate, Laufzeit**

9

**Die zu prognostizierende Zielvariable**

Die Zielvariable muss nominal sein  
 Klicken Sie auf den Knopf mit dem nach unten weisenden Pfeil

**Zielvariable**      Zielvariable ist nominal  
  **2**                      Zahl ihrer Ausprägungen

10

**Option: Ein- und Ausschliessen von Untersuchungseinheiten**

11

**Loesche wieder diese Box**

**Umkodierungen und Kein-Wert-Angaben**

Umkodierungen   
 Kein\_Wert-Angabe

**Einkommen = Einkommen / 10000;**

**erzeuge zusätzliche Felder für Umkodierungen / Kein\_Wert-Angaben**

---

Kontrollieren, ob Umkodierung so erfolgt wie gewünscht

diese Variablen ...

**Einkommen**

**1:20**                      ... aus diesen Datensätzen  
 vor und nach der Umkodierung  
 zur Kontrolle anzeigen

12

**Option: Prognosewerte in Datei schreiben**

13

**Programmende**

P45.18.2 Erläuterungen zu den Eingabe-Boxen

Prog45mt entspricht weitgehend dem bereits in P45.17.2 erläuterten Prognoseprogramm P45mp. Wir werden deswegen nur jene Eingabe-Boxen erläutern, die anders sind.

**Eingabe-Box 1 bis Eingabe-Box 7:**

Siehe Eingabe-Box 1 bis Eingabe-Box 7 in P45.17.2.

**Eingabe-Box 8:** Ursächliche Variable

Ursächliche Variable

ursächliche nominale Variable Hilfe

↔ □ □ Wohnort, Hausbesitz, Produkt  
↔ 2, 2, 3  
Zahl der Ausprägungen

---

ursächliche quantitative Variable

↔ □ □ Einkommen, Rueckrate, Laufzeit

Interaktionen zwischen den nominalen Variablen wie beim Prog45mp sind nicht zulässig.

**Eingabe-Box 9:** Die zu prognostizierende Zielvariable

Die zu prognostizierende Zielvariable

Die Zielvariable muss nominal sein  
Klicken Sie auf den Knopf mit dem nach unten weisenden Pfeil

↔ ↓ Zielvariable Zielvariable ist nominal  
↔ ↑ ↓ 2 Zahl ihrer Ausprägungen

Die Zielvariable muss nominal sein, dabei darf sie dichotom aber auch polytom sein.

**Eingabe-Box 10 bis Eingabe-Box 13:**

Siehe Eingabe-Box 10 bis Eingabe-Box 13 in P45.17.2.

### P45.18.3 Ausgabe

Almo liefert denselben Output wie Prog45mp. Siehe dazu P45.17.2. Für die 20 neuen Kunden wird folgende Prognose abgegeben.

Die Gruppe mit maximaler Wahrscheinlichkeit ist mit \* markiert

Gruppe 1 = Nicht-Rückzahler

Gruppe 2 = Rückzahler

Datensatz	prognostizierte Wahrscheinlichkeit der Zugehörigkeit zu Gruppe		prognostizierte Zugehörigkeit zu Gruppe
	1	2	
1	0.655*	0.345	1
2	0.207	0.793*	2
3	0.683*	0.317	1
4	0.082	0.918*	2
5	0.056	0.944*	2
6	0.260	0.740*	2
7	0.276	0.724*	2
8	0.549*	0.451	1
9	0.034	0.966*	2
10	0.027	0.973*	2
11	0.069	0.931*	2
12	0.770*	0.230	1
13	0.570*	0.430	1
14	0.007	0.993*	2
15	0.532*	0.468	1
16	0.025	0.975*	2
17	0.053	0.947*	2
18	0.089	0.911*	2
19	0.731*	0.269	1
20	0.047	0.953*	2

Nun ist es interessant zu vergleichen, welche Prognose das Allgemeine Lineare Modell geliefert hat:

Datensatz	Logit-Modell	Allgemeines Lineares Modell	
	----- prognostizierte Zugehoerigkeit zu Gruppe	----- prognostizierte Zugehoerigkeit zu Gruppe	
1	1	1	
2	2	2	
3	1	1	
4	2	2	
5	2	2	
6	2	2	
7	2	2	
8	1	1	
9	2	2	
10	2	2	
11	2	2	
12	1	1	
13	1	1	
14	2	2	
15	1	2	<----- Unterschied
16	2	2	
17	2	2	
18	2	2	
19	1	1	
20	2	2	

Lediglich beim Datensatz 15 liefern Logit-Modell und Allgemeines Lineares Modell eine unterschiedliche Prognose.

# Kapitel 10: Zusammengehörigkeiten zwischen Objekten suchen

## ***P45.19 Schritt 13: Cluster von Objekten bilden: Clusteranalyse***

Betrachten wir folgendes Beispiel. Die Daten sind simuliert.

Junge Menschen im Alter von 16 bis 32 wurden gefragt

1. wie viele Zigaretten sie am Tag rauchen
2. wie häufig sie Bier trinken
3. Wein
4. Schnaps
5. Aufputschgetränke
6. nicht-alkoholische Getränke
7. welche Art von Kleidung sie vorzugsweise tragen:
  - a. konventionelle Kleidung
  - b. unkonventionelle (schlampig, ausgeflippt)
  - c. elegant, modisch

Die Frage, die wir stellen wollen, lautet: Lassen sich die Jugendlichen über diese sieben Variablen in Typen untergliedern. Anders formuliert: Lassen sich die Jugendlichen in „Cluster“ unterteilen.

Um diese Frage zu beantworten rechnen wir eine Clusteranalyse mit Prog45mn. Wir werden drei Cluster unterscheiden können. Ein Cluster bilden beispielsweise die Jugendlichen, die sich modisch-elegant kleiden. Sie rauchen deutlich weniger als die anderen und trinken eher nicht-alkoholische Getränke – und wenn schon Alkohol, dann eher Wein.

Nun wird der Marktforscher noch eine zweite Frage stellen wollen: Lassen sich diese 3 Cluster mit demographischen und sozioökonomischen Variablen wie Geschlecht, Bildungsniveau, Lebensalter beschreiben. Unsere Daten umfassen deswegen noch folgende Variable.

8. Geschlecht
9. Bildungsgrad
10. Alter

Das Clusteranalyse-Programm in Almo wurde von Johann Bacher entwickelt. Siehe Abschnitt P45.19.4.

### **Eingabe in Clusteranalyse-Programm**

Prog77ml.Msk  
Clusteranalyse  
nach dem k-means-Verfahren  
für den allgemeinen Fall  
für beliebige Variablenkonstellation

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

1

Vereinbare Variable=  ;

2

Option: Weitere Vereinbarungen - nur wenn Almo dazu auffordert

3

"C:\Almo7\Testdat\Clustana.nam"  
        zeige = Namensdatei in Output zeigen  
leer = nicht

4

erzeuge zusätzliche Namensfelder

5

"C:\Almo7\Testdat\Clustana.dir"

6

quantitative Variable für die Clusterung  
  Rauchen,Bier,Wein,Schnaps,Aufputschdrinks,nichtalkoh.drinks  
-----  
ordinale Variable für die Clusterung  
(werden wie quantitative behandelt)  
    
-----  
nominale Variable für die Clusterung   
  Kleidung

7  
8  
9  
10  
11  
12  
13  
14

**Deskriptions-Variable** **Hilfe**

quantitative Variable

**Bildungsgrad, Alter**

---

ordinale Variable  
<werden wie quantitative behandelt>

---

nominale Variable **Hilfe**

**Geschlecht**

---

**Option: Ein- und Ausschliessen von Untersuchungseinheiten**

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

**Option: Untersuchungseinheiten gewichten**

---

**Clusterzahl**

**2**  
  **3**

Minimale Zahl von Clustern  
Maximale Zahl von Clustern

---

**Option: Clusterzugehörigkeiten der Objekte in Datei speichern**

---

**Grafik-Optionen**

---

**Ausgabe der Ergebnisse**

**0**

0= Ergebnisse stark verkürzt ausgeben  
1= Ergebnisse mittelstark verkürzt ausgeben  
2= Ergebnisse leicht verkürzt ausgeben  
3= Ergebnisse in voller Länge ausgeben

## P45.19.1 Erläuterung zu den Eingabe-Boxen

**Korrektur:** Die Programm-Nummer des Clusteranalyse-Programms ist Prog45mn und nicht wie oben abgebildet Prog77m1

**Eingabe-Box 1:** Speicher für x Variable

Siehe "Arbeiten mit Almo-Datenanalyse-System", Abschnitt P0.1.

**Eingabe-Box 2:** Option: Weitere Vereinbarungen

Siehe "Arbeiten mit Almo-Datenanalyse-System", Abschnitt P0.2.

**Eingabe-Box 3:** Datei der Variablennamen

Siehe "Arbeiten mit Almo-Datenanalyse-System", Abschnitt P0.3.

In unserem Beispiel umfaßt die Datei der Variablennamen folgende Namensgebungen:

Name 1=Rauchen;

Name 2=Bier;

Name 3=Wein;

Name 4=Schnaps;

Name 5=Aufputschdrinks;

Name 6=nichtalkoh.drinks;

Name 2:6=:nie,selten,ca. 1x Monat,ca. 1x in 14 Tagen, ca. 1x in Woche,mehrmals in Woche,täglich;

Name 7=Kleidung:konventionell,unkonventionell,elegant;

Name 8=Geschlecht:männl,weibl;

Name 9=Bildungsgrad;

Name 10=Alter;

Beachte: Die Variablen V2 bis V6, also "Bier" bis "nichtalkoh.drinks" besitzen alle dieselben Ausprägungsnamen. Almo erlaubt hier dem Benutzer folgende arbeitssparende Schreibweise.

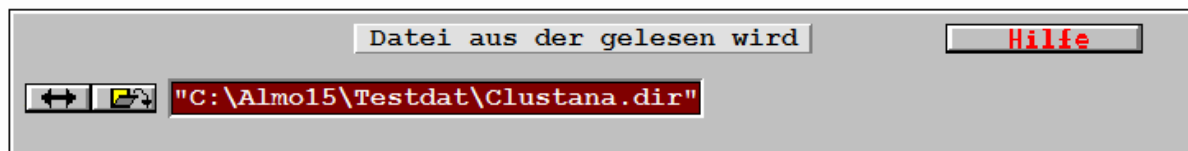
Name 2:6=: .....

Hinter dem Gleichheitszeichen folgt kein Variablennamen, sondern gleich ein Doppelpunkt. Damit wird signalisiert, daß jetzt Ausprägungsnamen folgen.

**Eingabe-Box 4:** Freie Namensfelder

Siehe "Arbeiten mit Almo-Datenanalyse-System", Abschnitt P0.3.

**Eingabe-Box 5:** Datei aus der gelesen wird



Es muß zuerst eine Almo-Arbeitsdatei (im Format DIREKT) erstellt werden. Dazu wird Prog45md oder Prog45mh verwendet. Siehe unsere ausführliche Darstellung in Abschnitt P45.1 und P45.2.

### Eingabe-Box 6: Klassifikations-Variable für die Clusterung

Klassifikations-Variable für die Clusterung

quantitative Variable für die Clusterung

Rauchen,Bier,Wein,Schnaps,Aufputschdrinks,nichtalkoh.drinks

ordinale Variable für die Clusterung  
(werden wie quantitative behandelt)

nominale Variable für die Clusterung Hilfe

Kleidung

Die Klassifikationsvariable sind jene Variable, die verwendet werden, um Cluster zu identifizieren.

*Eingabefeld 1:* Quantitative Variable für die Clusterung

Wir haben in unserem Beispiel als quantitative Variable eingesetzt:

Rauchen,Bier,Wein,Schnaps,Aufputschdrinks,nichtalkoh.drinks

*Eingabefeld 2:* Ordinale Variable

Ordinale Variable werden wie quantitative behandelt. D.h. letztendes: Sie sind nicht möglich. Der Benutzer sollte sie besser als nominale oder als quantitative Variable einführen. Unsere Empfehlung ist, ordinale Variable mit 2 oder 3 Ausprägungen als nominal und solche mit vier und mehr Ausprägungen als quantitative Variable zu deklarieren.

*Eingabefeld 3:* Nominale Variable für die Clusterung

Wir haben als nominale Variable

Kleidung

mit den 3 Ausprägungen konventionell, unkonventionell, elegant eingesetzt.

Die nominalen Variablen müssen ganzzahlig in 1-er Schritten kodiert sein. Sind sie das nicht, dann müssen sie in der Eingabe-Box "Kein-Wert-Angabe und Umkodierungen" entsprechend umkodiert werden. Beispiel: Die 3 Ausprägungen einer nominalen Variablen seien mit den Zahlen 2, 3.5, 7 kodiert worden. Die Variable muß in folgender Weise umkodiert werden:

V10 ( 2=1; 3.5=2; 7=3)

Die Codeziffern der nominalen Variablen müssen nicht notwendigerweise bei 1 beginnen. Folgende Kodierung ist korrekt: 3, 4, 5. Die Variable ist ganzzahlig und mit 1-er Schritten kodiert.

## Eingabe-Box 7: Deskriptions-Variable

Deskriptions-Variable Hilfe

quantitative Variable

---

ordinale Variable  
(werden wie quantitative behandelt)

---

nominale Variable Hilfe

Ordinale Variable werden wie quantitative behandelt. Es gilt das selbe was wir oben bei Eingabe-Box 6, bei Eingabefeld 2 schon ausgeführt haben.

Die Deskriptions-Variablen haben keinen Einfluß auf die Gewinnung der Cluster. Sie können weggelassen werden. Ihr Sinn ist folgender: Almo ermittelt zuerst aus den Klassifikations-Variablen die Cluster. In unserem Beispiel werden 3 Cluster gefunden. Dann errechnet Almo für die quantitativen Deskriptions-Variablen den Mittelwert je Cluster und für die nominalen Deskriptions-Variable die Anteilswerte je Cluster.

In unserem Beispiel gibt Almo folgendes aus:

Zellenmittelwerte der Deskriptionsvariablen:  
(Mittelwerte bei quantitativen Variablen, Anteilswerte bei nominalen)

Variable	C1	C2	C3	
v8				Geschlecht
1	0.49	0.73	0.43	männl
2	0.51	0.27	0.57	weibl
v9	3.16	3.09	3.56	Bildungsgrad
v10	22.81	24.52	23.19	Alter

Betrachten wir Cluster C3. Das sind die modisch-elegant gekleideten Jugendlichen, die eher Wein und Limonade trinken etc.

Der Anteil der Männer in diesem Cluster beträgt 0.43 (also 43 %), der der Frauen 0.57 (57 %). Die Frauen überwiegen also in C3 deutlich.

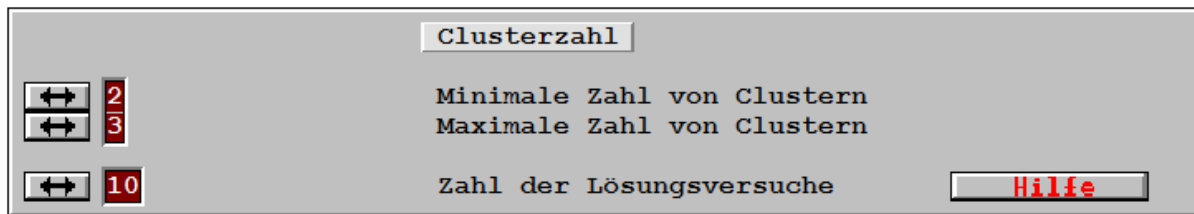
Der Bildungsgrad ist mit einem Mittelwert von 3.56 in diesem Cluster höher als in den Clustern C1 und C2.

Deskriptionsvariable dienen der inhaltlichen Validierung der Clusterlösung. Wir verwenden sie, um zu überprüfen, ob die gefundenen Cluster inhaltlich "stimmig" sind. Man könnte auch anders argumentieren: Deskriptionsvariable sind "ursächliche" Variable, mit deren Hilfe wir versuchen, die gefundenen Cluster zu erklären.

**Eingabe-Box 8:** Option: Ein- und Ausschließen von Untersuchungseinheiten  
Siehe "Arbeiten mit Almo-Datenanalyse-System", Abschnitt P0.7.

**Eingabe-Box 9:** Kein-Wert-Angabe und Umkodierungen: Siehe "Arbeiten mit Almo-Datenanalyse-System", Abschnitt P0.5.

**Eingabe-Box 10:** Clusterzahl



Clusterzahl	
2	Minimale Zahl von Clustern
3	Maximale Zahl von Clustern
10	Zahl der Lösungsversuche

Hilfe

*Eingabefeld 1 und 2*

Ein unangenehmes Problem der Clusteranalyse ist es, daß sie es dem Benutzer weitgehend überläßt, die Zahl der Cluster, die in den Daten gefunden werden sollen, selbst zu bestimmen. Zwar gibt es einige formale Verfahren, die dem Benutzer eine bestimmte Clusterzahl vorschlagen - wir werden darauf zurück kommen - die beste Methode ist es jedoch, sich mehrere Clusterlösungen anzuschauen und sich dann für eine zu entscheiden, die inhaltlich sinnvoll ist.

Wir empfehlen folgende Vorgehensweise:

Der Benutzer rechnet mehrere Analysen. Bei der ersten Analyse werden in obiger Eingabe-Box als minimale Clusterzahl 1 eingegeben und als maximale Clusterzahl etwa 8. Almo rechnet dann in einer Analyse eine Lösung mit 1 Cluster, mit 2 Cluster, mit 3 Cluster, ... bis 8 Cluster. Der Benutzer kann die 8 Clusterlösungen auf ihre Sinnhaftigkeit vergleichen und sich für eine entscheiden. Gleichzeitig macht Almo einen Vorschlag, welche Clusterzahl auf Grund eines formalen Kriteriums als entgültig betrachtet werden soll.

In einer weiteren Analyse wird man dann die Clusterzahl einschränken und beispielsweise sowohl als minimale wie auch als maximale Clusterzahl 3 angeben.

*Eingabefeld 3*

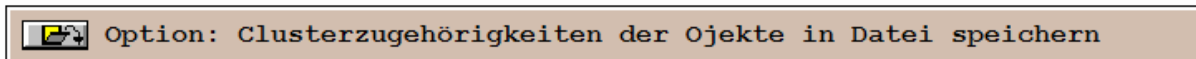
Geben Sie an wieviele Versuch gemacht werden sollen um (für jede Clusterzahl) eine Lösung zu finden

Beim K-Means-Verfahren kann wie bei allen iterativen Verfahren nicht ausgeschlossen werden, dass nur ein lokales Maximum erreicht wird. Daher ist es empfehlenswert, bei einer gegebenen Clusterzahl mehrere Startkonfigurationen zu untersuchen und schließlich jene auszuwählen, die ein Maximum bzw. Minimum berechnet. Mitunter ist es dabei erforderlich, mit mehr als 1000 Startkonfigurationen zu beginnen. In dem nachfolgenden Beispiel waren bei 4 und mehr Clustern 1000 Versuche je Clusterzahl zum Auffinden einer "besten" Lösung erforderlich, bei 12 Clustern sogar 2000.

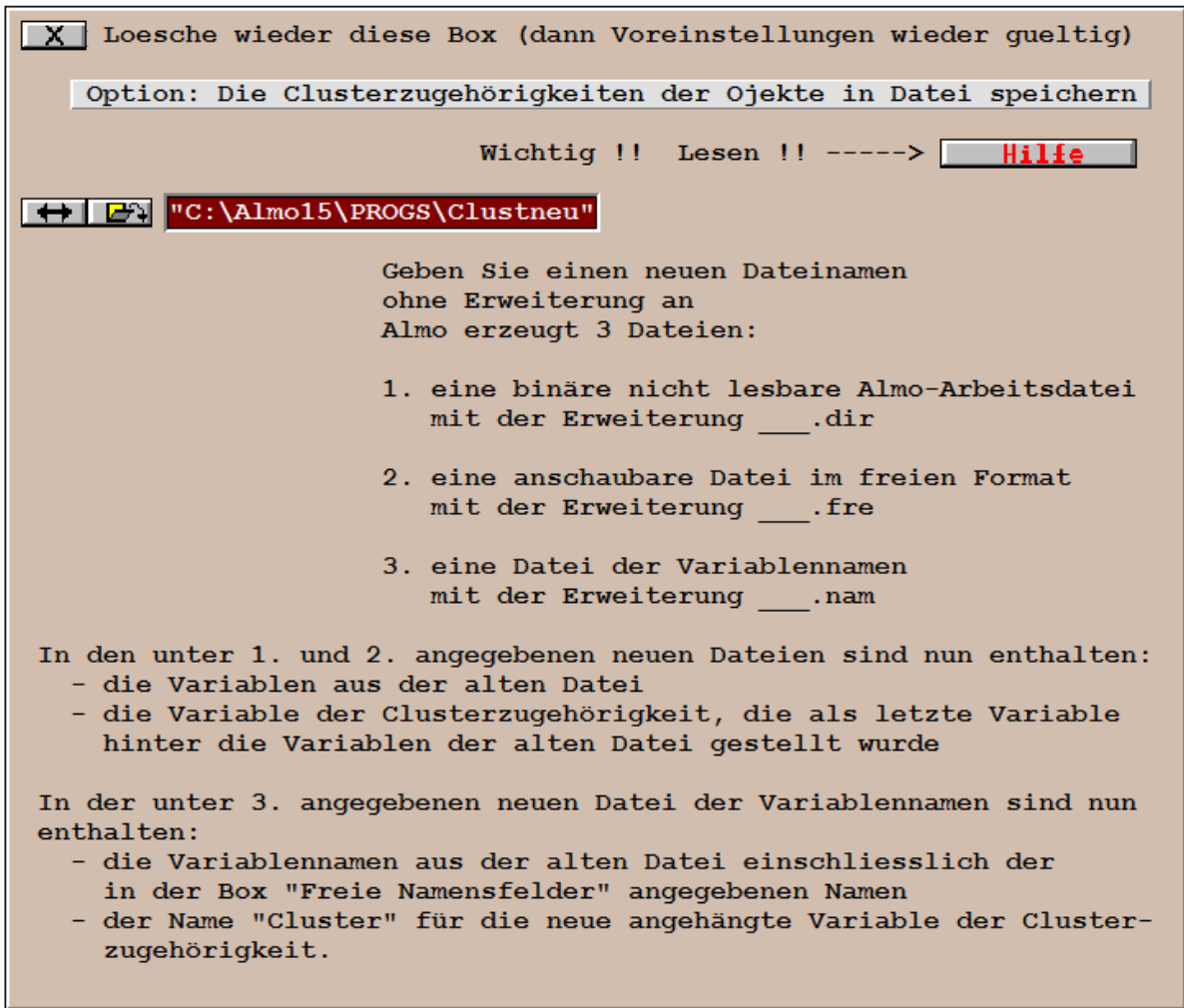
Cluster	Versuche je Clusterzahl				
	1	10	100	1000	2000
1	148.805	148.805	148.805	148.805	148.805
2	85.545	85.545	85.545	85.545	85.545
3	60.616	60.613	60.613	60.613	60.613
4	45.633	44.996	44.971	44.960	44.960
5	40.752	37.423	36.882	36.856	36.856
6	40.659	33.059	31.220	30.708	30.708
7	28.740	28.740	25.720	24.684	24.684
8	25.173	22.841	22.389	20.798	20.798
9	28.746	25.024	19.292	18.593	18.593
10	25.356	17.993	17.993	16.280	16.280
11	24.725	16.677	16.677	14.800	14.800
12	19.428	15.497	14.602	13.942	13.557

Zur Vermeidung eines langen Outputs, sollte auf jeden Fall die Option für eine stark verkürzte Ausgabe gewählt werden. Siehe Eingabe-Box 13.

**Eingabe-Box 11:** Option: Clusterzugehörigkeiten der Objekte in Datei speichern



Nach Klick auf den Knopf mit dem nach unten weisenden Pfeil, wird dann die eigentliche Optionsbox geöffnet.



Nachdem der Benutzer sich für eine Clusterlösung entschieden hat, ist es sinnvoll, die Clusterzugehörigkeit in eine Variable einzuschreiben und diese an die schon vorhandenen Variablen anzufügen. Damit die Originaldatei nicht gefährdet wird, verlangt Almo, daß eine neue Datei angelegt wird.

Geben Sie hier einen Namen für eine neue Datei an. Der Dateiname muß ohne Erweiterung geschrieben werden, weil Almo drei Dateien erzeugt, denen es selbst Erweiterungen anfügt.

Almo erzeugt folgende neue Datei Clustneu.fre (und Clustneu.dir)

```

V11: Clusterzugehörigkeit
6 3 2 2 5 5 2 2 2 18 1
13 4 3 2 1 5 3 1 5 30 2
10 4 2 6 5 6 1 2 4 21 1
5 4 3 2 1 7 1 1 3 30 2
9 5 1 2 3 5 2 1 3 21 2
. . . . .
. . . . .
. . . . .

```

Mit dieser Datei kann man nun ein Prog45mf "Ursachen für die Zielvariable" rechnen, wobei man V11 als Zielvariable angibt und die Deskriptionsvariablen

- V8 Geschlecht
- V9 Bildungsgrad

V10 Alter

als ursächliche Variable angibt. Mit dieser Analyse versucht man also die Determinanten der Clusterzugehörigkeit zu ermitteln. Wir zeigen dies in Abschnitt P45.20 und P45.21.

**Eingabe-Box 12:** Option: Grafik-Optionen

Siehe "Arbeiten mit Almo-Datenanalyse-System", Abschnitt P0.10.

**Eingabe-Box 13:** Ausgabe der Ergebnisse

Ausgabe der Ergebnisse

0

0= Ergebnisse stark verkürzt ausgeben  
1= Ergebnisse mittelstark verkürzt ausgeben  
2= Ergebnisse leicht verkürzt ausgeben  
3= Ergebnisse in voller Länge ausgeben

Bei Eingabe von "3" werden die Ergebnisse in voller Länge ausgegeben. Im Almo-Handbuch "P36, P37 Clusteranalyse" Abschnitt P37.2.4 und P37.2.7 werden die Ergebnisse ausführlich kommentiert. Für die Zwecke des Data Mining genügt es "0" einzusetzen, d.h. sich mit den "stark reduzierten Ergebnissen" zufrieden zu geben. Sie umfassen die wesentlichen Ergebnisse und sind ohnehin lange genug.

## P45.19.2 Ausgabe der Ergebnisse

Ausgabe der Ergebnisse

-----

Fuer Analyse ausgewaehlte Variable:

Klassifikationsvariablen:

V7	Kleidung	nominal	UG =	1	OG =	3
V1	Rauchen	quantitativ				
V2	Bier	quantitativ				
V3	Wein	quantitativ				
V4	Schnaps	quantitativ				
V5	Aufputschdrinks	quantitativ				
V6	nichtalkoh.drinks	quantitativ				

Deskriptionsvariablen:

V8	Geschlecht	nominal	UG =	1	OG =	2
V9	Bildungsgrad	quantitativ				
V10	Alter	quantitativ				

-----

Es wurden 589 Datensaeetze eingelesen,  
davon werden 589 Datensaeetze analysiert.

### \*\*\*\*\* Erläuterung:

Um den Output kurz zu halten, haben wir im Kurzprogramm Prog45mn in der Eingabe-Box 10 "Clusterzahl" als minimale Clusterzahl 2 und als maximale Clusterzahl 4 angegeben. Almo liefert nun im folgenden die Lösung für 2 Cluster, für 3 Cluster und für 4 Cluster.

-----

Ergebnisse fuer 2-Clusterloesung

-----

=====

Clustergroessen:

C1 = 282

C2 = 307

### \*\*\*\*\* Erläuterung:

Cluster 1 umfasst 282 Personen

Cluster 2 umfasst 307 Personen

Zellenmittelwerte der Klassifikationsvariablen:  
(Mittelwerte bei quantitativen Variablen, Anteilswerte bei nominalen)

```
-----
Variable      C1      C2
-----
V7
  1      0.36    0.38      Kleidung
  2      0.45    0.26      konventionell
  3      0.19    0.36      unkonventionell
                                     elegant

V1      13.35    5.26      Rauchen
V2       4.63    3.36      Bier
V3       2.04    3.05      Wein
V4       3.39    2.24      Schnaps
V5       4.33    2.98      Aufputschdrinks
V6       4.49    5.57      nichtalkoh.drink
-----
```

\*\*\*\*\* **Erläuterung:**

Dies ist nun die für uns wichtigste Tabelle. Wir werden diese Tabelle jedoch für die nachfolgende 3-Clusterlösung erläutern, da diese als die inhaltlich am besten interpretierbare erscheint.

```
-----
Ergebnisse fuer 3-Clusterloesung
-----
```

```
=====
Clustergroessen:
C1 =      182
C2 =      217
C3 =      190
-----
```

Zellenmittelwerte der Klassifikationsvariablen:  
(Mittelwerte bei quantitativen Variablen, Anteilswerte bei nominalen)

```
-----
Variable      C1      C2      C3
-----
V7
  1      0.07    0.77    0.20      Kleidung
  2      0.73    0.06    0.34      konventionell
  3      0.21    0.17    0.46      unkonventionell
                                     elegant

V1      11.01    11.34    4.82      Rauchen
V2       3.93     5.06    2.74      Bier
V3       1.97     2.07    3.71      Wein
V4       4.07     2.24    2.19      Schnaps
V5       5.08     2.79    3.19      Aufputschdrinks
V6       4.62     4.97    5.56      nichtalkoh.drink
-----
```

\*\*\*\*\* **Erläuterung:**

Almo teilt zunächst die Anteilswerte der nominalen Klassifikationsvariablen je Cluster mit.

Betrachten wir zuerst Cluster C1.

Von den 182 Personen, die Cluster 1 bilden, trägt ein Anteil von 0.07, also 7% konventionelle Kleidung. 73% tragen unkonventionelle ("ausgeflippte", schlampige) Kleidung und 21% tragen elegante Kleidung. Zusammen ergibt das 100%. Die Anteilswerte einer nominalen Variablen summieren sich für jedes Cluster zu 1.0. Cluster 1 enthält also vor allem die unkonventionell gekleideten Jugendlichen

Betrachten wir nun Cluster C2.

Von den 217 Personen, die Cluster 2 bilden, tragen 77% konventionelle Kleidung.

In Cluster 3 tragen die meisten Jugendlichen elegante (modische) Kleidung. Betrachten wir nun die Mittelwerte der quantitativen Klassifikationsvariablen je Cluster. Deutlich am wenigsten rauchen die Jugendlichen in Cluster 3. Das sind diejenigen, die wir bereits als die elegant gekleideten identifiziert haben. Die beiden anderen Cluster C1 und C2 unterscheiden sich nur wenig.

Die Biertrinker finden wir vor allem in Cluster 2. Das sind die Jugendlichen, die wir als die konventionell gekleidet identifiziert haben. Die 3 Cluster können wir inhaltlich etwa so beschreiben:

#### Cluster 1

Die Jugendlichen sind überwiegend unkonventionell ("ausgeflippt", schlampig) gekleidet. Sie rauchen und trinken (im Vergleich zu den anderen) häufiger Schnaps und Aufputzmittel.

#### Cluster 2

Die Jugendlichen sind überwiegend konventionell gekleidet. Sie rauchen und trinken (im Vergleich zu den anderen) häufiger Bier.

#### Cluster 3

Die Jugendlichen sind eher elegant (modisch) gekleidet. Sie rauchen wenig und trinken (im Vergleich zu den anderen) häufiger Wein und nicht-alkoholische Getränke.

Allgemein gilt: Zur Interpretation der Cluster sollte man sich vor allem die maximalen und minimalen Werte je Zeile der Tabelle anschauen. In der weiter unten folgenden Tabelle "Signifikanz der z-Werte" kann man dann nachschauen, ob der jeweilige Wert vom Mittelwert über alle Personen signifikant abweicht.

-----  
 Ergebnisse fuer 4-Clusterloesung  
 -----

=====  
 Clustergroessen:

C1 = 149  
 C2 = 130  
 C3 = 190  
 C4 = 120

Zellenmittelwerte der Klassifikationsvariablen:

(Mittelwerte bei quantitativen Variablen, Anteilswerte bei nominalen)

Variable	C1	C2	C3	C4	
V7					Kleidung
1	0.13	0.08	0.82	0.28	konventionell
2	0.62	0.59	0.06	0.23	unkonventionell
3	0.25	0.33	0.12	0.50	elegant
V1	14.11	2.74	10.74	7.33	Rauchen
V2	4.25	3.08	5.08	2.82	Bier
V3	2.11	2.00	2.05	4.56	Wein
V4	4.29	2.93	2.06	1.93	Schnaps
V5	5.12	4.06	2.66	2.83	Aufputzschdrinks
V6	4.30	6.00	4.99	5.07	nichtalkoh.drink

=====

Cluster- zahl	Streuungsquadratsummen		F-Wert	ETA**2	PRE
	innerhalb	zwischen			
2	3718.232	699.268	110.394	0.158	KW
3	3105.806	1311.694	123.745	0.297	0.165
4	2815.557	1601.943	110.947	0.363	0.093

Beachte: Zur Berechnung der Streuungszerlegung wurden die Variablen standardisiert um Vergleichbarkeit zu erhalten.

\*\*\*\*\* **Erläuterung:**

Der höchste F-Wert (von 123.745) entsteht bei der 3-Clusterlösung. Also verwendet dies als formales Entscheidungskriterium für die richtige Zahl der Cluster und untersucht nun im folgenden die 3-Clusterlösung im Detail. In unserem Beispiel ist die 3-Clusterlösung auch jene, die am besten inhaltlich interpretierbar ist. Es muß aber ausdrücklich darauf hingewiesen werden, daß das nicht immer so ist. Unsere Empfehlung ist, jene Clusterlösung zu wählen, die inhaltlich gut interpretierbar ist, aber auch einen "ordentlichen" (nicht notwendigerweise den maximalen) F-Wert besitzt. Siehe auch die ausführliche Diskussion dieser Problematik im Handbuch "Johann Bacher: P36, P37 Clusteranalyse" Abschnitt P37.2.7

=====  
Die 3-Clusterloesung wird weiter untersucht  
=====

Clustergroessen:

C1	182	( 30.900 %)
C2	217	( 36.842 %)
C3	190	( 32.258 %)

KW-Faelle (ungewichtet)= 0

=====  
Zellenmittelwerte der Klassifikationsvariablen  
(Mittelwerte bei quantitativen / ordinalen Variablen)  
(Anteilswerte bei nominalen Variablen)

Variable	C1	C2	C3	
V7				Kleidung
1	0.07	0.77	0.20	konventionell
2	0.73	0.06	0.34	unkonventionell
3	0.21	0.17	0.46	elegant
V1	11.01	11.34	4.82	Rauchen
V2	3.93	5.06	2.74	Bier
V3	1.97	2.07	3.71	Wein
V4	4.07	2.24	2.19	Schnaps
V5	5.08	2.79	3.19	Aufputschdrinks
V6	4.62	4.97	5.56	nichtalkoh.drink

Standardabweichungen:

Variable	C1	C2	C3	
V7				Kleidung
1	0.25	0.42	0.40	konventionell
2	0.45	0.24	0.47	unkonventionell
3	0.41	0.38	0.50	elegant
V1	7.07	7.35	5.99	Rauchen
V2	0.98	0.82	0.87	Bier
V3	1.04	0.99	1.44	Wein
V4	1.52	1.17	1.20	Schnaps
V5	1.07	1.13	1.26	Aufputschdrinks
V6	1.42	1.24	1.23	nichtalkoh.drink

Z-Werte:

Variable	C1	C2	C3	
V7				Kleidung
1	-16.40	14.00	-5.79	konventionell
2	11.16	-18.27	-0.52	unkonventionell
3	-2.25	-4.15	5.14	elegant
V1	3.56	4.42	-9.90	Rauchen
V2	-0.44	19.61	-19.30	Bier
V3	-7.72	-7.34	10.89	Wein
V4	11.28	-6.94	-6.79	Schnaps
V5	18.22	-10.94	-4.74	Aufputschdrinks
V6	-4.10	-1.00	5.71	nichtalkoh.drink

Signifikanz der z-Werte:

Variable	C1	C2	C3	
V7				Kleidung
1	100.00	100.00	100.00	konventionell
2	100.00	100.00	39.90	unkonventionell
3	97.43	100.00	100.00	elegant
V1	99.95	100.00	100.00	Rauchen
V2	33.85	100.00	100.00	Bier
V3	100.00	100.00	100.00	Wein
V4	100.00	100.00	100.00	Schnaps
V5	100.00	100.00	100.00	Aufputschdrinks
V6	100.00	68.37	100.00	nichtalkoh.drink

\*\*\*\*\* **Erläuterung:**

Die Tabelle gibt an, ob die Mittelwerte der Variablen je Cluster vom Gesamtmittelwert über alle Personen signifikant abweichen. Betrachten wir die Zeile V6 nichtalkoh.drinks. In der Tabelle der "Zellenmittelwerte der Klassifikationsvariablen" finden wir folgende Werte:

	C1	C2	C3
	4.62	4.97	5.56

Für Cluster C2 finden wir einen Mittelwert von 4.97. Dieser Mittelwert weicht nur mit einer Sicherheitswahrscheinlichkeit von 68.37% vom Gesamtmittelwert aus allen Personen ab. Umgekehrt betrachtet: Die Irrtumswahrscheinlichkeit beträgt 31.63 %. Üblicherweise wird eine Mindestsicherheit von 95% bzw. eine maximale

Irrtumswahrscheinlichkeit von 5% gefordert. Die Mittelwerte von C1 und C3 weichen mit 100%iger Sicherheit vom Gesamtmittelwert ab. Diese 100 % entstehen Also-intern durch Aufrunden. Der richtige Wert ist 99.9999....%

=====

Zellenmittelwerte der Deskriptionsvariablen:  
(Mittelwerte bei quantitativen Variablen, Anteilswerte bei nominalen)

Variable	C1	C2	C3	
V8				Geschlecht
1	0.49	0.73	0.43	männl
2	0.51	0.27	0.57	weibl
V9	3.16	3.09	3.56	Bildungsgrad
V10	22.81	24.52	23.19	Alter

\*\*\*\*\* **Erläuterung:**

Wie wir bereits bei der Erläuterung der Programm-Eingabe zu Eingabe-Box 7 "Deskriptionsvariable" ausgeführt haben, sind die Deskriptionsvariablen für die Clusteranalyse nicht notwendig. Sie haben keinen Einfluß auf die Clusterbildung. Also liefert nun für die gefundenen 3 Cluster die Mittelwerte der quantitativen Deskriptionsvariablen bzw. die Anteilswerte der nominalen Deskriptionsvariablen. Cluster 1 (unkonventionell gekleidet, Raucher, Schnaps und Aufputschgetränke) weist keinen Unterschied der Geschlechter auf. Das Bildungsniveau liegt in der Mitte, weicht aber nicht signifikant vom Gesamtmittelwert ab (siehe nachfolgende Tabelle "Signifikanz der z-Werte"). Das durchschnittliche Alter ist das niedrigste.

Im Cluster 2 (konventionell gekleidet, Raucher, Biertrinker) sind die Männer deutlich überrepräsentiert. Das Bildungsniveau ist das niedrigste. Das durchschnittliche Alter ist das höchste.

Im Cluster 3 (elegant-modisch gekleidet, Wenig-Raucher, Wein und Limonade) sind die Frauen überrepräsentiert. Das Bildungsniveau ist das höchste. Das durchschnittliche Alter weicht nicht signifikant vom Gesamtmittelwert ab.

Wenn wir die Deskriptionsvariable wie ursächliche Variable (die die Clusterzugehörigkeit determinieren) betrachten, dann können wir sagen: Das Geschlecht bestimmt die Zugehörigkeit zu Cluster 2 und 3. Männer gehen eher in Cluster 2, Frauen eher in Cluster 3. Eine hohe Bildung determiniert eher die Zugehörigkeit zu Cluster 3. Die Jüngsten tendieren eher zu Cluster 1 und die Älteren eher zu Cluster 2

Standardabweichungen:

Variable	C1	C2	C3	
V8				Geschlecht
1	0.50	0.44	0.49	männl
2	0.50	0.44	0.49	weibl
V9	1.35	1.26	1.27	Bildungsgrad
V10	3.85	3.86	4.09	Alter

Z-Werte:

Variable	C1	C2	C3	
V8				Geschlecht
1	-1.77	5.73	-3.72	männl
2	1.77	-5.73	3.72	weibl
V9	-1.01	-2.04	3.21	Bildungsgrad
V10	-2.63	3.64	-1.25	Alter

Signifikanz der z-Werte:

Variable	C1	C2	C3	
V8				Geschlecht
1	92.17	100.00	99.97	männl
2	92.17	100.00	99.97	weibl
V9	68.84	95.77	99.83	Bildungsgrad
V10	99.08	99.96	78.57	Alter

\*\*\*\*\* Erläuterung:

Die Tabelle gibt an, ob die Mittelwerte der Deskriptionsvariablen je Cluster vom Gesamtmittelwert über alle Personen signifikant abweichen.

Betrachten wir die Zeile V10 Alter. In der Tabelle der "Zellenmittelwerte der Deskriptionsvariablen" finden wir für die 3 Cluster folgende Mittelwert:

	C1	C2	C3
	22.81	24.52	23.19

Diese Mittelwerte weichen nur mit der oben angegebenen Sicherheitswahrscheinlichkeit von vom Gesamtmittelwert des Alters aus allen Personen ab. Die Mittelwerte von C1 und C2 weichen mit einer Sicherheit von 99.08 % bzw. 99.96 % vom Gesamtmittelwert ab. Wenn wir die Deskriptionsvariablen als Variable betrachten, die ursächlich für die Clusterzugehörigkeit sind, dann können wir folgen, daß das Lebensalter wesentlich die Zugehörigkeit zu Cluster 1 und 2 determiniert. Anders bei Cluster 3: Die Angehörigen dieses Clusters weichen in ihrem Alters-Mittelwert vom Gesamtmittelwert nicht signifikant ab. Es ist üblich, einen Wert von 95 % als Signifikanzgrenze zu betrachten. Das Alter ist keine Determinante für die Zugehörigkeit zu Cluster 3.

Gesamtstatistiken fuer Klassifikationsvariablen:

	F-Wert	Signifikanz (1-p)*100	ETA**2	Name
V7				Kleidung
1	208.403	100.000	0.416	konventionell
2	141.499	100.000	0.326	unkonventionell
3	26.837	100.000	0.084	elegant
V1	55.472	100.000	0.159	Rauchen
V2	343.218	100.000	0.539	Bier
V3	132.980	100.000	0.312	Wein
V4	126.882	100.000	0.302	Schnaps
V5	213.937	100.000	0.422	Aufputschdrinks
V6	25.253	100.000	0.079	nichtalkoh.drinks

\*\*\*\*\* **Erläuterung:**

Die Tabelle gibt an, ob eine Variable insgesamt signifikant zur Trennung der Cluster beiträgt.

In unserem Beispiel haben alle Klassifikationsvariable einen signifikanten Einfluß auf die Clusterbildung.

Betrachten wir den ETA\*\*2-Wert für Bier (=0.539). Er sagt aus, daß 53.9% der Streuung im Bierkonsum auf die Trennung in 3 Cluster zurückgeführt werden kann.

Gesamtstatistiken fuer Deskriptionsvariablen:

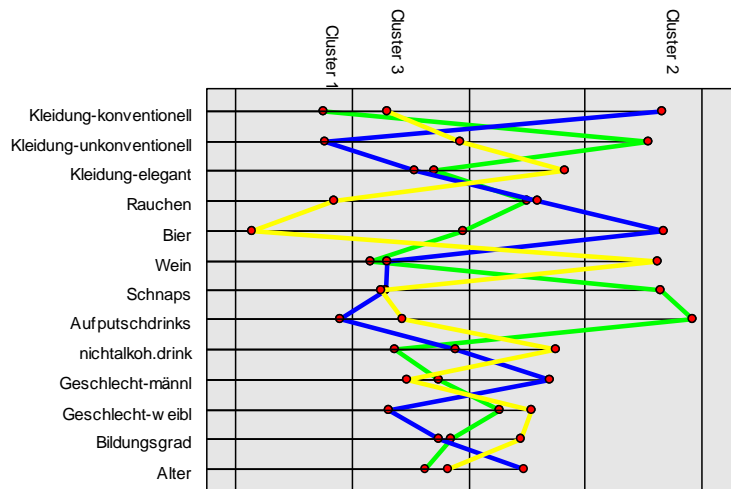
	F-Wert	Signifikanz (1-p)*100	ETA**2	Name
V8				Geschlecht
1	23.207	100.000	0.073	männl
2	23.207	100.000	0.073	weibl
V9	7.534	99.909	0.025	Bildungsgrad
V10	10.519	99.987	0.035	Alter

\*\*\*\*\* **Erläuterung:**

Die Tabelle gibt an, welcher Teil der Streuung in den Deskriptionsvariablen auf die Trennung in 3 Cluster zurückgeführt werden kann und ob dieser signifikant von .0 verschieden ist. Das Alter z.B. hat einen ETA\*\*2-Wert von 0.035. D.h. 3.5% der Streuung des Alters können auf die Trennung in 3 Cluster zurückgeführt werden.

Dieser Streuungsanteil ist mit 99.987%-iger Signifikanz von .0 verschieden.

Almo zeichnet noch abschließend ein Liniendiagramm der Clustermittelwerte für alle Variable. Dabei sind die Variablen standardisiert, d.h. mit Standardabweichungen als Maßeinheiten gezeichnet.



Die Linien in diesem Diagramm sind auf dem Bildschirm verschieden färbig. Hier im Druck sind sie als Grau-Abstufungen nicht gut voneinander zu trennen.

### P45.19.3 Technische Anmerkung zur Clusteranalyse im Almo-Data-Mining

Almo rechnet eine k-means Clusteranalyse. Die Parameter und Optionen sind so gewählt, daß mit jeder Datenkonstellation eine Clusteranalyse gerechnet werden kann. Die Eigenschaften von Prog45mn sind folgende (siehe dazu auch das Handbuch zu P36, P37; Johann Bacher: Clusteranalyse):

1. Es können beliebig viele quantitative und nominale Variable eingeführt werden. Ordinale Variable müssen entweder als quantitative oder als nominale deklariert werden.
2. Die Zahl der zu clusternden Untersuchungsobjekte ist nicht begrenzt.
3. Die nominalen Variablen werden intern in 0-1-kodierte Dummies aufgelöst.
4. Es wird ein Modell 3 gerechnet. Siehe Handbuch, Abschnitt P37.2.4. D.h. es wird ein Minimaldistanzverfahren mit gewichteten euklidischen Distanzen gerechnet. Als Gewichtungskriterium werden die Varianzen der Variablen verwendet.
5. Der Test auf die (richtige) Clusterzahl über das F-Max-Kriterium wird mit standardisierten Variablen gerechnet.
6. Die nominalen Variablen werden mit 0.5 gewichtet.

### P45.19.4 Weiterführende Hinweise

In Almo sind 2 verschiedene Verfahren der Clusteranalyse enthalten, das hier im Almo-Data-Mining verwendete k-means-Verfahren und die hierarchische Clusteranalyse. Beide Verfahren wurden von Johann Bacher programmiert. Die beiden Verfahren sind als Prog36 und Prog37 mit einer Vielzahl von Varianten in Almo enthalten. Im Almo-Handbuch „Johann Bacher: P36, P37 Clusteranalyse“ werden diese Verfahren ausführlich beschrieben.

Zur Literatur: Sowohl als Einführung, wie auch als umfassende Gesamtdarstellung sei empfohlen: Johann Bacher: Clusteranalyse, 1994.

# Kapitel 11: Ursachen für die Clusterzugehörigkeit

Häufig wird man wissen wollen, was die Ursachen dafür sind, daß die Untersuchungsobjekte (in unserem Beispiel: die Jugendlichen) verschiedene Typen bilden, in der Sprache der Clusteranalyse: verschiedenen Cluster angehören. Hier bieten sich 3 Lösungsmöglichkeiten an:

1. Die vermuteten ursächlichen Variablen werden als „Deskriptionsvariable“ eingeführt. Wir haben das bereits in unserem Beispiel gezeigt.
2. Wir rechnen mit Prog45mq ein „Allgemeines Lineares Modell“. Dabei ist die Clusterzugehörigkeit die nominale Zielvariable.
3. In Abschnitt P45.15.0 haben wir auf die Probleme hingewiesen, die entstehen, wenn das Allgemeine Lineare Modell auf nominale, insbesondere polytome Zielvariable angewendet wird. Als Alternative, die derartige Probleme vermeidet, haben wir die Logitanalyse empfohlen. Um die Ursachen der Clusterzugehörigkeit zu erkunden wird deswegen das Logitmodell Prog45mw angeboten.
4. Eine Möglichkeit, die Probleme des ALM mit nominalen Zielvariablen, annähernd zu lösen, ist es mit Prog45gw ein „gewichtetes“ ALM zu rechnen. Wir werden das aber im folgenden nicht vorführen. Der Benutzer, der sich mit Prog45gw in Abschnitt P45.15.2.2 beschäftigt hat, wird problemlos dieses Programm auf die Analyse der Clusterursachen anwenden können.

## ***P45.20 Schritt 14a: Ursachen für die Clusterzugehörigkeit: Analyse mit ALM***

Wir rechnen zuerst mit Prog45mq ein Allgemeines Lineares Modell. Prog45mq entspricht exakt Prog45mf.

**Prog45mf.Msk**

**Wirkungsstärke der ursächlichen Variablen  
hinsichtlich der Zielvariablen**

Die Zielvariable kann nominal, ordinal oder quantitativ sein  
Es wird ein Allgemeines Lineares Modell gerechnet

Was ist ein Kurzprogramm ? -->

Bedienung -->

1

Vereinbare Variable=  ;

2

Option: Weitere Vereinbarungen - nur wenn Almo dazu auffordert

3

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

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

4

**Name 11=Cluster:C1,C2,C3;**

**erzeuge zusätzliche Namensfelder**

5

"C:\Almo7\TESTDAT\Clustneu.dir"

6

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 Zielvariable**

---

**ordinale Zielvariable**

---

**nominale Zielvariable**


**Cluster**



14

 Option: Gewichtete Kleinste-Quadrate-Schätzung

15

 Option: Prognosewerte und Residuen

16

 Option: Wertemuster


17

 Option: "Aussehen" der auszugebenden Tabelle bzw. Matrix

18



 Grafik-Optionen

19



 Option: Die errechneten Koeffizienten in eine Datei speichern

20

**Ausgabe der Ergebnisse**

  **1**

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

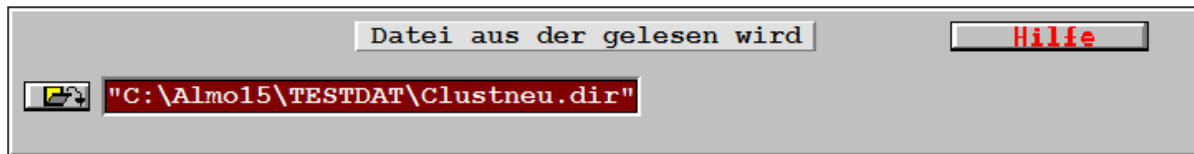
  **0**

1= Basisstatistiken ausgeben  
2= Basisstatistiken und "diverse Werte" ausgeben  
0= nicht

## P45.20.1 Erläuterungen zu den Eingabe-Boxen

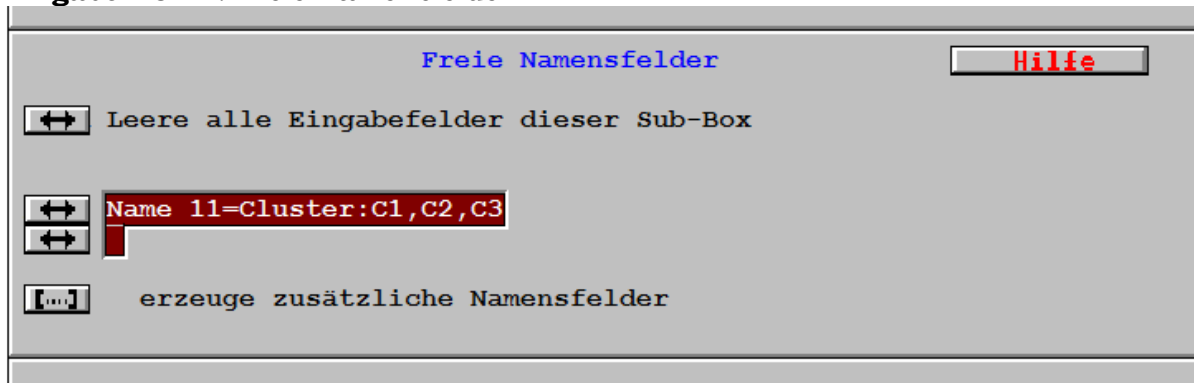
Wir wollen nur die 4. und die 5. Eingabe-Box erläutern, da alle anderen analog den Eingabe-Boxen von Prog45mf (Abschnitt P45.15.1.1) ausgefüllt sind.  
Zuerst Eingabe-Box 5.

**Eingabe-Box 5:** Datei aus der gelesen wird



Mit dem Clusteranalyse-Programm Prog45mn haben wir mit Eingabe-Box 10 „Clusterzugehörigkeit der Objekte in Datei speichern“ die neue Datei „Clustneu.dir“ erzeugt. In der Variablen V11 ist dabei die Clusterzugehörigkeit gespeichert worden.

**Eingabe-Box 4:** Freie Namensfelder



In der Datei der Variablennamen „Clustana.nam“ ist für V11, die Clusterzugehörigkeit, noch kein Name enthalten. Einen solchen schreiben wir jetzt in Eingabe-Box 4.

**Eingabe-Box 6** und **Eingabe-Box 7:** Zielvariable und ursächliche Variable

Zielvariable

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 Zielvariable



ordinale Zielvariable

Hilfe



nominale Zielvariable

Hilfe



Ursächliche Variable

ursächliche nominale Variable

**Geschlecht**

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

paarweise Vergleiche (Kontraste) für die  
ursächlichen nominalen Variablen rechnen

---

ursächliche quantitative Variable

**Bildungsgrad,Alter**

---

ursächliche ordinale Variable

Die nominale Zielvariable ist die in der Eingabe-Box "Freie Namensfelder" definierte Variable "Cluster". Als ursächliche Variable verwenden wir

Geschlecht (nominal)  
Bildungsgrad und Alter (quantitativ)

## P45.20.2 Ausgabe der Ergebnisse

Almo liefert folgende Ergebnisse (gekürzt):

Fuer Analyse ausgewaehlte Variable

V8 Geschlecht: männl weibl  
 V9 Bildungsgrad  
 V10 Alter  
 V11 Cluster: C1 C2 C3

V8 wird auch bezeichnet mit A  
 die Auspraegungen (bzw.Dummies) mit  
 A1 =männl  
 A2 =weibl

Prozentwerte der unabhaengigen nominalen Variablen  
 je Auspraegung der abhaengigen nominalen Variablen  
 (zeilenweise auf 100 % addiert)

		Cluster C1 V11-1	Cluster C2 V11-2	Cluster C3 V11-3
Geschlec männl	V8-1	27.27	48.18	24.55
Geschlec weibl	V8-2	35.52	22.39	42.08

### \*\*\*\*\* Erläuterung:

Die Prozentwerte summieren sich zeilenweise zu 100%.

Von den Männern befinden sich die meisten, nämlich 48.18 % in Cluster 2 (=die konventionell gekleideten Biertrinker). Von den Frauen befinden sich die meisten, nämlich 42.08% in Cluster 3 (=die eleganten Weintrinker).

Mittelwerte der unabhaengigen quantitativen Variablen  
 je Auspraegung der abhaengigen nominalen Variablen

		Cluster C1 V11-1	Cluster C2 V11-2	Cluster C3 V11-3
Bildungs	V9	3.16	3.09	3.56
Alter	V10	22.81	24.52	23.19

Haeufigkeiten je Auspraegung der nominalen Variablen

-----

V8 Geschlecht  
 V8-1 männl 330  
 V8-2 weibl 259

V11 Cluster  
 V11-1 C1 182  
 V11-2 C2 217  
 V11-3 C3 190

Haeufigkeitstabelle:

		Cluster C1 V11-1	Cluster C2 V11-2	Cluster C3 V11-3
Geschlec männl	A1	90	159	81
Geschlec weibl	A2	92	58	109

Koeffizienten fuer quantitat./ordinale Variable aus univariater Analyse

hinsichtlich der abhaeng. Var. V11-1 Cluster: C1

Variable	Regr. koeff.	part. Korrel.	Signifikanz p	(1-p)100
V9 Bildungsgrad	-0.0284	-0.078	0.060	93.98
V10 Alter	-0.0129	-0.112	0.007	99.31

**\*\*\*\*\* Erläuterung:**

Das Alter ist eine signifikante Determinante für die Zugehörigkeit zu Cluster 1. Nicht jedoch der Bildungsgrad. Die Signifikanz dieser Variablen liegt (wenn auch knapp) unter der üblichen Signifikanzschwelle von 95%.

hinsichtlich der abhaeng. Var. V11-2 Cluster: C2

V9 Bildungsgrad	-0.0124	-0.034	0.414	58.62
V10 Alter	0.0178	0.151	0.000	99.98

**\*\*\*\*\* Erläuterung:**

Nur das Alter ist eine signifikante Determinante für die Zugehörigkeit zu Cluster 2.

hinsichtlich der abhaeng. Var. V11-3 Cluster: C3

V9 Bildungsgrad	0.0408	0.111	0.007	99.29
V10 Alter	-0.0048	-0.042	0.311	68.87

**\*\*\*\*\* Erläuterung:**

Nur der Bildungsgrad ist eine signifikante Determinante für die Zugehörigkeit zu Cluster 3.

"multivariate" partielle Korrelation zwischen der abhaengigen nominalen Variablen und den einzelnen unabhengigen quantitat./ordinalen Variablen

Variable	part. Korrel	Signifikanz p	(1-p)100
V9 Bildungsgrad	0.1141	0.021	97.87
V10 Alter	0.1564	0.001	99.89

**\*\*\*\*\* Erläuterung:**

Hier wird pauschal festgestellt, daß sowohl Bildungsgrad als auch Alter signifikant die Trennung der Jugendlichen in 3 Cluster bestimmen.

Koeffizienten der Dummies  
hinsichtlich der abh. Var. V11-1 Cluster: C1

Effekte von A Geschlecht

	Effekte partielle		Signifikanz	
	Korrelat.	p	(1-p)100	
A1 männl	-0.0389	-0.0909	0.0277	97.23%
A2 weibl	0.0496	0.0909	0.0277	97.23%

**\*\*\*\*\* Erläuterung:**

Das Geschlecht ist eine signifikante Determinante für die Zugehörigkeit zu Cluster 1.

Koeffizienten der Dummies  
hinsichtlich der abh. Var. V11-2 Cluster: C2

Effekte von A Geschlecht

	Effekte partielle		Signifikanz	
	Korrelat.	p	(1-p)100	
A1 männl	0.0999	0.2263	0.0000	100.00%
A2 weibl	-0.1273	-0.2263	0.0000	100.00%

**\*\*\*\*\* Erläuterung:**

Das Geschlecht ist eine signifikante Determinante für die Zugehörigkeit zu Cluster 2.

Koeffizienten der Dummies  
hinsichtlich der abh. Var. V11-3 Cluster: C3

Effekte von A Geschlecht

	Effekte partielle		Signifikanz	
	Korrelat.	p	(1-p)100	
A1 männl	-0.0610	-0.1414	0.0007	99.93%
A2 weibl	0.0777	0.1414	0.0007	99.93%

**\*\*\*\*\* Erläuterung:**

Das Geschlecht ist eine signifikante Determinante für die Zugehörigkeit zu Cluster 2.

Zusammenfassung

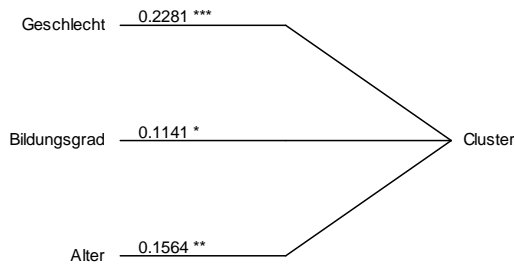
Streuungsquelle	Korrel Koeff.	F-Wert	df	Signifikanz	
				p	(1-p)100
V8 Geschlecht	0.2281	16.0196	2	0.0000	99.9986
V9 Bildungsgrad	0.1141	3.8496	2	0.0213	97.8681
V10 Alter	0.1564	7.3211	2	0.0011	99.8881

**\*\*\*\*\* Erläuterung:**

Hier wird nochmals pauschal festgestellt, dass alle 3 ursächlichen Variablen signifikant die Trennung der Jugendlichen in 3 Cluster bestimmen. Am stärksten mit einer Korrelation von 0.2281 wirkt das Geschlecht, am schwächsten (aber noch signifikant) mit einer Korrelation von 0.1141 der Bildungsgrad.

Almo zeichnet noch obige Tabelle als Flussdiagramm. Auf den Pfeilen stehen die Korrelationskoeffizienten.

Partielle  
Korrelationskoeffizienten



Zusammenfassung: Effekte und Regressionskoeffizienten  
und ihre Signifikanzen  
hinsichtlich der abhaengigen Variablen  
Cluster: C1

	Effekte Regress.koeff	Signifikanz (1-p)*100
-----		
A1 männl	-0.038937	97.233276
A2 weibl	0.049611	97.233276
Bildungsgrad	-0.028362	93.984016
Alter	-0.012933	99.313048

Zusammenfassung: Effekte und Regressionskoeffizienten  
und ihre Signifikanzen  
hinsichtlich der abhaengigen Variablen  
Cluster: C2

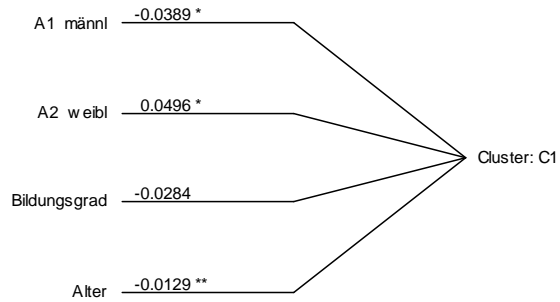
	Effekte Regress.koeff	Signifikanz (1-p)*100
-----		
A1 männl	0.099937	99.995000
A2 weibl	-0.127333	99.995000
Bildungsgrad	-0.012432	58.624282
Alter	0.017761	99.977091

Zusammenfassung: Effekte und Regressionskoeffizienten  
und ihre Signifikanzen  
hinsichtlich der abhaengigen Variablen  
Cluster: C3

	Effekte Regress.koeff	Signifikanz (1-p)*100
-----		
A1 männl	-0.061000	99.927416
A2 weibl	0.077722	99.927416
Bildungsgrad	0.040794	99.290631
Alter	-0.004828	68.872545

Almo zeichnet für diese 3 Tabellen Flußdiagramme. Auf den Pfeilen stehen die Effekte der Dummies der nominalen Variablen bzw. die Regressionskoeffizienten der quantitativen Variablen. Um Platz zu sparen zeigen wir hier nur das Flußdiagramm zur 1. Tabelle.

Effekte und Regressionskoeffizienten  
A Geschlecht: A1=männl A2=weibl



***P45.21 Schritt 14b: Ursachen der Clusterzugehörigkeit: Analyse mit Logitmodell***

**Eingabe mit Programm Prog45mu**

Das Programm ist identisch mit Prog45m9.

Prog45m9.Msk

Wirkungsstärke der ursächlichen Variablen  
hinsichtlich einer nominalen Zielvariablen

Es wird eine Logit-Analyse gerechnet  
(alternativ kann auch eine Probit-Analyse gerechnet werden)

Almo-Struktur -->   
Bedienung -->

1    
Vereinbare Variable=  ;

2  Option: Weitere Vereinbarungen - nur wenn Almo dazu auffordert

3    
  "C:\Almo7\TESTDAT\Clustana.nam"  
        zeige = Namensdatei in Output zeigen  
leer = nicht

4

5    
 "C:\Almo7\TESTDAT\Clustneu.dir"

6

7   
ursächliche nominale Variable  
    
ursächliche quantitative Variable

8  Option: Alternative: Probitanalyse

9  Option: Auflösung der unabhäng. nominalen Variab. in Dummies

10  Option: Ein- und Ausschliessen von Untersuchungseinheiten

11  Option: Umkodierungen und Kein-Wert-Angaben

12  Option: Prognosewerte ermitteln

13  Option: Wertemuster

14  Option: Die errechneten Koeffizienten in eine Datei speichern

15  Grafik-Optionen

16 **Ausgabe der Ergebnisse**

1  
 0= Ergebnisse in voller Länge ausgeben  
 1= Ergebnisse verkürzt ausgeben

0  
 1= Basisstatistiken ausgeben  
 0= nicht

### P45.21.1 Erläuterungen zu den Eingabe-Boxen

Die **Eingabe-Boxen**

- 4:** (Freie Namensfelder)
- 5:** (Datei aus der gelesen wird)
- 6:** (Zielvariable)
- 7:** (Ursächliche Variable)

sind mit kleinen Modifikationen dieselben, wie im vorausgehend beschriebenen Prog45mq. Das Programm ist außerdem weitgehend identisch mit dem in Abschnitt P45.16.1 und P45.16.2 bereits ausführlich erläuterten Logit-Programm. Wir ersparen es uns deswegen, die einzelnen Eingabe-Boxen zu erklären.

### P45.21.2 Ausgabe der Ergebnisse (gekürzt)

Ergebnisse fuer 2. Auspraegung "C2" der abhaengigen Variablen V11 Cluster  
 (als Referenz wird die 1. Auspraegung "C1" verwendet)

unabhaengige Variable			Regress. koeff.ß	Risiko epx(ß)	relatives Risiko	Signifikanz (1-p)*100	partielle Korrelation
A1	Geschlec:	männl	0.49256	1.63650	63.64951	100.00	0.11526
A2	Geschlec:	weibl	-0.49256	0.61106	-38.89380	100.00	-0.11526
V9	Bildungsgrad		0.05017	1.05145	5.14545	45.27	0.03562
V10	Alter		0.09678	1.10162	10.16191	99.97	0.09351

Ergebnisse fuer 3. Auspraegung "C3" der abhaengigen Variablen V11 Cluster  
 (als Referenz wird die 1. Auspraegung "C1" verwendet)

unabhaengige Variable			Regress. koeff.ß	Risiko epx(ß)	relatives Risiko	Signifikanz (1-p)*100	partielle Korrelation
A1	Geschlec:	männl	-0.06588	0.93624	-6.37600	44.99	-0.03568
A2	Geschlec:	weibl	0.06588	1.06810	6.81022	44.99	0.03568
V9	Bildungsgrad		0.22440	1.25158	25.15765	99.19	0.06234
V10	Alter		0.02710	1.02747	2.74675	68.99	0.02739

**\*\*\*\*\* Erläuterung:**

Die Zielvariable "Cluster" besitzt 3 Ausprägungen. Wir haben also eine polytome Logitanalyse gerechnet. Almo liefert Ergebniss für Cluster C2 und C3. Cluster C1 wird als Referenz verwendet. Siehe auch die ausführliche Erläuterung zum Risiko bei polytomer Zielvariablen in Abschnitt P45.16.2.1. Betrachten wir die Ergebnisse

für C2 und beschränken wir uns zunächst auf die ursächliche Variable des Geschlechts. Das relative Risiko für männliche Jugendliche beträgt 63.7 % und das der weiblichen Jugendlichen -38.9 %. Was bedeutet dies ?

Die Männer haben im Vergleich zum Durchschnitt aller Personen eine um 63.7 % erhöhte Wahrscheinlichkeit dem Cluster C2 als dem Referenz-Cluster C1 anzugehören.

Die Frauen haben im Vergleich zum Durchschnitt aller Personen eine um 38.9 % verringerte Wahrscheinlichkeit dem Cluster C2 als dem Referenz-Cluster C1 anzugehören.

Bei der Interpretation der Ergebnisse ist also zu berücksichtigen, dass 2 Referenzgruppen bestehen,

1. eine auf Seiten der ursächlichen nominalen Variablen. Dies ist der Durchschnitt aus allen Personen.
2. Und eine auf Seiten der Zielvariablen. Dies ist das Cluster C1.

Betrachten wir jetzt die ursächliche quantitativen Variablen. Der Bildungsgrad erweist sich mit  $(1-p)100 = 45.27\%$  als nicht signifikant. Die Wirkung des Alters ist folgendermaßen zu interpretieren: Nimmt das Alter um 1 Jahr zu, dann erhöht sich die Wahrscheinlichkeit, eher dem Cluster C2 als dem Cluster C1 anzugehören um 10.2 %. Wir haben hier also nur eine Referenzgruppe - die auf Seiten der Zielvariablen.

Trefferhaeufigkeiten bei Individualdaten fuer abhaengige Variable V11 Cluster

		tatsaechlich			prognostiziert absolut		
		1 C1	2 C2	3 C3	1 C1	2 C2	3 C3
C1	1	182	0	0	51	78	53
C2	2	0	217	0	29	149	39
C3	3	0	0	190	28	70	92

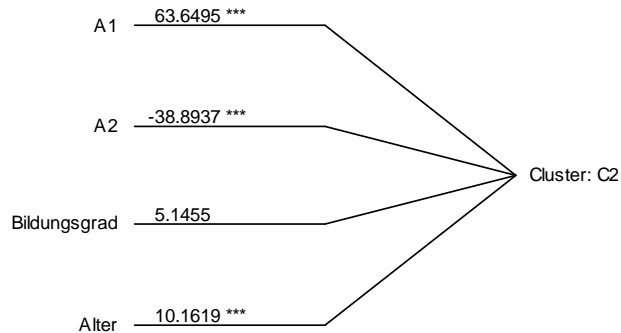
		prognostiziert relativ			erwartet Zufall		
		1 C1	2 C2	3 C3	1 C1	2 C2	3 C3
C1	1	59.6	62.1	60.3	56.2	67.1	58.7
C2	2	62.3	92.7	62.0	67.1	79.9	70.0
C3	3	60.1	62.2	67.7	58.7	70.0	61.3

**\*\*\*\*\* Erläuterung:**

Aus der Tabelle "tatsaechlich" entnehmen wir, dass 182 Personen zum Cluster C1 gehören. Aus der Tabelle "prognostiziert, absolut" erkennen wir (im Diagonalglied), dass wir nur 51 Personen richtig als zu C1 gehörend mit unserem Logitmodell prognostizieren konnten. Aus der Tabelle "erwartet, Zufall" müssen wir erkennen, dass der Zufall besser wäre als unser Logitmodell: Bei zufälliger Zuweisung der Personen zu den 3 Clustern würden 56.2 Personen richtig zugewiesen werden. Bei Cluster C2 und C3 ist unser Logitmodell allerdings deutlich besser als der Zufall.

Almo zeichnet nun noch ein Flussdiagramm der relativen Risikokoeffizienten. Wir zeigen nur das für C2.

relative Risikokoeffizienten  
für unabhängige Variable  
A Geschlecht: A1=männl A2=weibl



Auf den Strichen stehen die relativen Risikokoeffizienten.

3 Sterne = mit 99.9% signifikant

2 Sterne = mit 99.0% signifikant

1 Stern = mit 95.0% signifikant

0 Signifikanz unter 95

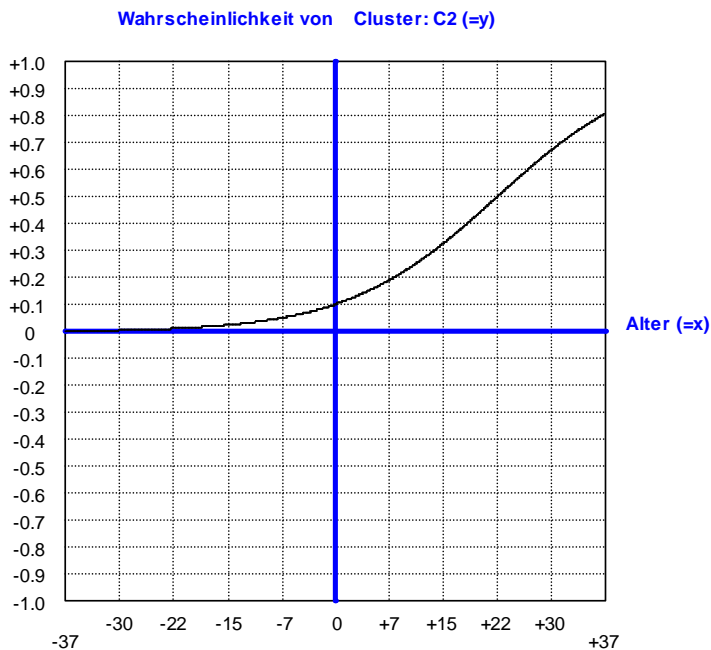
Almo zeichnet noch die logistische Funktion für die quantitativen ursächlichen Variablen. Wir zeigen hier die

logistische Funktion für die  
Zielvariable V11 Cluster: 2. Ausprägung: C2

hinsichtlich der  
ursächlichen Variablen V10 Alter

Die Referenzgruppe ist Cluster C1

Logistische Funktion  
 $Y = 1/(1+e^{*-(-2.1641+0.096781*X)})$



Die logistische Kurve verläuft sehr flach. D.h. der Zusammenhang zwischen Alter und der Wahrscheinlichkeit für C2 ist schwach.

Betrachten wir eine 22-jährige „Durchschnitts-Person“. Für sie müssen wir folgendermaßen interpretieren: Mit einer Wahrscheinlichkeit von ca. 0.5 (=50%) wird diese Person eher dem Cluster C2 als C1 (der Referenzgruppe) angehören. Wir haben diese Person als 22-jährige „Durchschnitts-Person“ bezeichnet. Damit meinen wir, daß diese Person in allen ursächlichen Variablen den Mittelwert (bzw. den Anteilswert bei den nominalen Variablen) besitzt. Nur in der Variablen „Alter“ besitzt sie nicht den Mittelwert, sondern den Wert 22.

## **P45.22 Clusterzugehörigkeit als unabhängige oder als abhängige Variablen**

In unserem Beispiel war die Clusterzugehörigkeit die abhängige (Ziel-)Variable. Natürlich wäre es auch möglich, die Clusterzugehörigkeit als ursächliche Variable in einer Analyse zu verwenden. In unserem „Rückzahlungs-Beispiel“ wäre es z.B. möglich, die ursächlichen Variablen Wohnort, Geschlecht, Beruf, Hausbesitz etc. als Klassifikationsvariable zu betrachten, eine Clusteranalyse zu rechnen und dadurch die Clusterzugehörigkeit der Personen zu bestimmen. Diese kann dann als ursächliche nominale Variable verwendet werden, mit der Absicht die Zielvariable der Rückzahlung bzw. Nicht-Rückzahlung des Kredits zu bestimmen. Eine derartige Analyse ist möglich, aber sie wird nicht sehr erfolgreich sein. Die normale Analyse mit einzelnen ursächlichen Variablen hat in aller Regel einen höheren Erklärungswert (eine höhere multiple Korrelation) und sie erlaubt auch eine differenziertere inhaltliche Erklärung. Nichts spricht aber dagegen, die Clusterzugehörigkeit in einer zweiten, zusätzlichen Analyse als ursächliche Variable einzuführen.

### **Schlagwortverzeichnis**

- 2-dimensionale Tabelle 44
- 2I-Test 69
- 3D-Balkendiagramm 60
- Allgemeines Lineares Modell 84
- ALM 84
- Analyse mit Logitmodell 268
- Anteilswert 116, 182
- Auspartiellierung 104
- Balkendiagramm 21, 60
- Binomialtest 71
- Chi-Quadrat-Beitrag 56, 69
- Clusteranalyse 237
- Clusterzahl 243
- Clusterzugehörigkeit 244
- Codeziffer 13
- Cramers V 20, 63
- d\_Kreuzprodukt 95
- Deskriptionsvariable 242, 252, 256
- Dezimalwerte 13
- dichotome Variable 7
- dichotome Zielvariable 84, 175
- Diskriminanzanalyse 86
- Dummy-Kodierung 179, 208
- Dummy-Variable 12, 18, 149, 153, 179
- Durchschnittsperson 116, 198
- Effekte 108, 110, 127, 128, 139, 141, 167, 209
- Effekt-Koeffizient 190
- erklärte Streuung 111
- Erwartungswert 56
- Eta-Korrelation 20
- exp( $\beta$ ) 187, 209
- Fehlerstreuung 111
- fitting\_constants 96
- Flußdiagramm 100, 105, 184
- F-Max-Kriterium 250, 255
- Gesamtstreuung 111
- gewichtete Kleinste-Quadrate 85, 97, 127, 149
- Gewinn-zu-Verlust-Verhältnis 189
- gleiche Zellenhäufigkeiten 97
- Grafik 22, 60, 100, 184
- Grafik-Editor 24
- Grafiktyp 60
- Groß-Gamma 17
- Gruppierungsvariable 79, 101, 117, 185, 201
- Heteroskedastizität 85
- Interaktion 92
- Interaktionstabelle 50, 65
- kanonische Korrelation 20
- Kein-Wert-Behandlung 28
- KFA 70
- Klassifikationsvariable 241, 248, 274
- k-means Clusteranalyse 255
- Kodierung 190
- Konfigurationsfrequenzanalyse 56, 70
- Konstante 92, 108, 139, 174
- Kontingenzkoeffizient 63
- Kontingenztafel 52

Kontraste 93  
 Kontrollvariable 50  
 Korrelationskoeffizient 111, 135  
 Korrelationsmatrix 28  
 korrelieren 7  
 Kovarianz 29  
 Kovarianz-Matrix 34  
 Kreuzprodukt 33, 95  
 lineare Funktion 116, 118  
 Liniendiagramm 61  
 logistische Funktion 174, 184, 186, 195  
*Logitanalyse* 172  
 Median 54  
 Mehrfach-Tabelle 48, 63  
 Minimaldistanzverfahren 255  
 MIT-Anweisung 120  
 Mittelwert 54  
 Mittelwerts-Tabelle 54, 72  
 multidimensionale Kontingenztafel 52, 68  
 multiple Korrelation 104, 128, 144  
 multivariate Analyse 135, 143  
 nominale Variable 7  
 odds 189  
 Ogive 178  
 ordinale Variable 7  
 paarweise Vergleiche 93, 109  
 paarweises Ausschneiden 27  
 Partialtafel 50, 65  
 partielle Korrelation 15, 192  
 partielle multiple Korrelation 104  
 Phi-Korrelation 20  
 Pillais Spur 20, 135  
 polytome Variable 7  
 polytome Zielvariable 84, 131, 149, 207, 210  
 PRE-Koeffizient 17, 63  
 Probit-Analyse 178  
 Prognoseerfolg 126  
 Prognosewert 121, 123, 192, 214  
 Prozentuieren 62  
 punktbiseriale Korrelation 20  
 Quadratsumme 95  
 Quadratsummenmatrix 31  
 qualitative Variable 7  
 quantitative Variable 8  
 Quartilsabstand 54  
 Quasi-Korrelationsmatrix 30  
 Rangwert-Variable 8  
 Regression 78  
 Regressionsebene 83  
 Regressionsgerade 79  
 Regressionskoeffizient 35, 79, 107, 128, 139, 167, 174, 186  
 Reproduzierbarkeit 85, 123, 130  
 Residuen 123, 129, 170  
 Risiko 187, 209  
 sequentiell 96  
 Signifikanz 15, 29, 144  
 Spaltenvariable 44  
 SPSS 36  
 SS Typ 96  
 Standardabweichung 29, 35, 54  
 Streudiagramm 38, 75  
 Streuungsmatrizen 95  
 Tafel 38  
 Trefferhäufigkeit 126, 193, 215  
 ursächliche Variable 91  
 Validierung 242  
 Varianzheterogenität 127  
 vollständiges Ausschneiden 28  
 Wahrscheinlichkeitsanalyse 173  
 Wahrscheinlichkeitsfunktion 127  
 weighted squares of means 95, 116  
 Wertemuster 97, 121, 206  
 Zeilenvariable 44  
 Zellenmittelwert 122, 251  
 Zielvariable 38  
 zwei-dimensionale Tafel 59

## Literatur

- Aldrich & Nelson:** Linear Probability, Logit and Probit Models, Sage Publications 1984  
**Allison P. D.:** Multiple imputation for missing data, in: Sociological Methods & Research, Vol. 28, Feb. 2000, S. 301 – 309  
**Arminger Gerhard:** Faktorenanalyse, Teubner Verlag, 24, Stuttgart, 1979  
**Bacher Johann:** Clusteranalyse, Oldenbourg Verlag, München, Wien, 1994

- Bacher Johann:** P36, P37, Clusteranalyse, Almo-Handbuch
- Bortz, Lienert, Boehnke:** Verteilungsfreie Methoden in der Biostatistik, Springer, Berlin, 1990
- Bortz, Jürgen:** Statistik für Sozialwissenschaftler, Springer Verlag, Berlin, 1993
- Cleve J., U. Lämmel:** Data Mining, Oldenbourg Verlag, 2014
- Denz Hermann:** Analyse latenter Strukturen, UTB 1198, Franke Verlag, München, 1982
- Denz Hermann:** Das Groß-Gamma-Modell, in Holm: Befragung 6, Francke Verlag, München, UTB 436, 1979
- Denz Hermann:** Regressionsanalyse mit ordinalen Variablen, in Holm: Befragung 5. Francke Verlag, München, UTB 435, 1977
- Fischer Gerhard:** Einführung in die Theorie psychologischer Tests, Huber Verlag, Bern, 1974
- Fleischer Karlheinz, Rässler Susanne:** Erfolgreiche Fusion von Datensätzen, Ein Mythos?, Universität Erlangen-Nürnberg
- Gerich Joachim:** Nichtparametrische Skalierung nach Mokken, Trauner Verlag, Linz, 2001
- Goldberger A. S.:** Econometrie Theory, Wiley, New York, 1964
- Haller/Holm:** Österreich im Wandel, Verlag für Geschichte und Politik, Oldenbourg Verlag, Wien, München 1996
- Holm Kurt:** Befragung Bd. 6, 1979, UTB 436, Abschnitt 13.4, Francke, München
- Holm Kurt:** Die Befragung 3 (Faktorenanalyse), 1976, UTB 433, Francke, München
- Krauth:** Einführung in die Konfigurationsfrequenzanalyse, Beltz Verlag, Weinheim, 1993
- Little R. J. A. & Rubin D. B.:** The analysis of social science data with missing values, in J. Fox /J. S. Long (Hrsg.): Modern methods of data analysis, Sage, Newsbury Park, 1990
- Peterson Helge:** Data Mining, Oldenbourg Verlag, 2005
- Rässler Susanne, Fleischer Karlheinz:** Aspects Concerning Data Fusion Techniques, Discussion Paper 16/1997, Universität Erlangen-Nürnberg, Lehrstuhl für Statistik
- Rost Jürgen:** Testtheorie, Testkonstruktion, Huber Verlag, Bern 1996
- Rubin D. B.:** Multiple imputation for nonresponse in surveys, Wiley, New York, 1987
- Runkler, Thomas A.:** Data Mining, Vieweg+Teubner Verlag, 2011
- Sixtl Friedrich:** Skalierungsverfahren, in Holm (Hrsg.): Befragung 4, UTB 434, Franke Verlag, 1976
- Überla K.:** Faktorenanalyse, Springer, Berlin, Heidelberg, New York, 1968
- Urban Dieter:** Logit Analyse, Gustav Fischer Verlag, 1993
- Urban Dieter:** Regressionstheorie und Regressionstechnik, Teubner Verlag, 36, Stuttgart, 1982