



Kurt Holm

Statistische Datenanalyse Data-Mining

Teil 1

Ein Standard-Auswertungssystem

2014

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

Das vorliegende Dokument enthält die 1. Hälfte des Almo-Handbuchs "P45 Almo-Data-Mining". Der Text wurde etwas überarbeitet und aktualisiert. Die 2. Hälfte und das Kapitel "Arbeiten mit dem Almo-Data-Mining-System" werden als Teil 2 und Teil 3 als separate Almo-Dokumente Nr. 25 und 26 veröffentlicht. Siehe nachfolgendes Inhaltsverzeichnis.

Zum Begriff "Data-Mining"

Inzwischen sind wir mit dem Begriff "Data-Mining" als Überschrift über dieses Almo-Dokument 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 jedoch akzeptieren, dass der Begriff in Almo eingengt wird auf die Auswertung von Daten, die die Form der Datenmatrix besitzen, wie wir sie gleich in Abschnitt P45.0 beschreiben. Wie diese Daten-Matrix zustande gekommen ist, wird im Almo-Data-Mining ausgeklammert - mit folgenden zwei Ausnahmen: Die Datenmatrix darf fehlende Werte besitzen und sie kann aus der Fusion mehrerer heterogener Dateien entstehen - wobei allerdings strenge Bedingungen erfüllt sein müssen. Wir werden in Kapitel 3 und 4 darauf eingehen, wie Ersatzwerte für fehlende Werte in die Datenmatrix "imputiert" werden können und wie mehrere verschiedene Dateien zu einer gemeinsamen Datenmatrix fusioniert werden können.

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 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 (0,7 MB)
32. Datei-Operationen mit Almo (1,1 MB)
33. Wählerstromanalyse und Wahlhochrechnung.PDF (1,6 MB)
34. Soziometrie. Auswertung soziometrischer Daten (0,5 MB)

Übersicht

TEIL I

KAPITEL 1: EINE ALMO-ARBEITSDATEI ERSTELLEN	12
<i>Schritt 1a: „Tabulator-getrennte“ Daten aus Excel nach Almo übertragen.....</i>	<i>19</i>
<i>Schritt 1b: Daten im Format FREI oder FIX in eine Almo-Arbeitsdatei schreiben</i>	<i>27</i>
KAPITEL 2: DATEN KENNENLERNEN	37
<i>Schritt 2: Daten anschauen.....</i>	<i>37</i>
<i>Schritt 3: Kennwerte der Variablen anschauen.....</i>	<i>37</i>
<i>Schritt 4: Variable auszählen.....</i>	<i>44</i>
KAPITEL 3: DATEN BEREINIGEN	57
<i>Schritt 5a: Mittelwert für fehlende Werte einsetzen</i>	<i>57</i>
<i>Schritt 5b: Prognosewerte für fehlende Werte durch das Allgemeine Lineare Modell ermitteln</i>	<i>64</i>
<i>Schritt 5c: Prognosewerte für fehlende Werte durch Logitanalyse ermitteln.....</i>	<i>83</i>
<i>Schritt 5d: Multiple Imputation.....</i>	<i>91</i>
KAPITEL 4: DATEIEN VEREINEN	109
<i>Schritt 6a: Datenfusion mit dem Allgemeinen Linearen Modell.....</i>	<i>109</i>
<i>Schritt 6b: Datenfusion mit der Logitanalyse.....</i>	<i>124</i>
<i>Schritt 6c: Fusionierte Dateien vereinen</i>	<i>135</i>
KAPITEL 5: MEHRERE VARIABLE ZU EINER MESSUNG KOMBINIEREN.....	149
<i>Schritt 7a: Aus mehreren Variablenwerten einen Gesamtpunkt看 bilden.....</i>	<i>149</i>
<i>Schritt 7b: Mit Faktorenanalyse einen gewichteten Gesamtpunkt看 (Faktorwert) bilden</i>	<i>151</i>
<i>Schritt 7c: Rasch-Skalierungsverfahren.....</i>	<i>168</i>

TEIL II

(enthalten in Almo-Dokument Nr. 25)

Kapitel 6: Zusammenhänge „blind“ suchen

Schritt 8: Variable miteinander korrelieren

Kapitel 7: Einzelne Zusammenhänge genauer untersuchen

Schritt 9a: Variable zwei- und beliebig-dimensional tabellieren

Schritt 9b: Streudiagramm für 2 oder 3 Variable

Kapitel 8: Mehrfach-Zusammenhänge untersuchen

Schritt 11a: Ursachen für die Zielvariable: Allgemeines Lineares Modell (ALM)

 Ursachen für die Zielvariable: Zielvariable ist dichotom

 Ursachen für die Zielvariable: Zielvariable ist nominal-polytom

Schritt 11b: Gewichtete Kleinste-Quadrate-Schätzung für nominal- polytome Zielvariable

 Ursachen für die Zielvariable: Zielvariable ist quantitativ

Schritt 11c: Alternative wenn Zielvariable nominal (dichotom oder polytom): Die Logit-Analyse
Logitanalyse mit dichotomer Zielvariablen
Logit-Analyse mit polytomer Zielvariablen

Kapitel 9: Prognose leisten

Schritt 12a: Werte für Zielvariable mit ALM prognostizieren
Schritt 12b: Werte für Zielvariable mit Logitanalyse prognostizieren

Kapitel 10: Zusammengehörigkeiten zwischen Objekten suchen

Schritt 13: Cluster von Objekten bilden: Clusteranalyse

Kapitel 11: Ursachen für die Clusterzugehörigkeit

Schritt 14a: Ursachen für die Clusterzugehörigkeit: Analyse mit ALM
Schritt 14b: Ursachen der Clusterzugehörigkeit: Analyse mit Logitmodell

Teil III

(enthalten in Almo-Dokument Nr. 26)

Arbeiten mit dem Almo-Datenanalyse-System

Inhaltsverzeichnis

Teil I

P45 ALMO-DATA-MINING	8
P45.0 Einführung	8
P45.0.1 Anwendungsbereiche für den Data Mining Prozeß	9
P45.0.2 Die Schritte des Data-Mining-Prozesses	10
P45.0.3 Wie man mit Almo arbeitet	11
P45.0.4 Unsere Testdaten	11
KAPITEL 1: EINE ALMO-ARBEITSDATEI ERSTELLEN	17
P45.1 Schritt 1a: Daten aus Excel und SPSS nach Almo übertragen	19
P45.1.1 Erläuterungen zu den Eingabe-Boxen	23
P45.2 Schritt 1b: Daten im Format FREI oder FIX in eine Almo-Arbeitsdatei schreiben	27
P45.2.1 Erläuterungen zu den Eingabe-Boxen	31
KAPITEL 2: DATEN KENNENLERNEN	37
P45.3 Schritt 2: Daten anschauen	37
P45.3.1 Erläuterungen zu den Eingabe-Boxen	39
P45.3.2 Ausgabe	40
P45.4 Schritt 3: Kennwerte der Variablen anschauen	41
P45.4.1 Erläuterungen zu den Eingabe-Boxen	45
P45.5 Schritt 4: Variable auszählen	47
P45.5.1 Erläuterungen zu den Eingabe-Boxen	50
P45.5.2 Ausgabe	51
KAPITEL 3: DATEN BEREINIGEN	57
P45.6 Schritt 5a: Mittelwert für fehlende Werte einsetzen	58
P45.6.1 Erläuterung zu den Eingabe-Boxen	61
P45.7 Prognosewerte für fehlende Werte einsetzen	67
P45.7.1 Schritt 5b: Prognosewerte für fehlende Werte durch das Allgemeine Lineare Modell ermitteln	67
P45.7.1.1 Wie Prog45mm rechnet	68
P45.7.1.2 Eingabe in das Maskenprogramm Prog45mm	69
P45.7.1.3 Erläuterungen zu den Eingabe-Boxen	73
P45.7.1.4 Ausgabe aus Prog45mm	81
P45.7.2 Schritt 5c: Prognosewerte für fehlende Werte durch Logitanalyse ermitteln	86
P45.7.2.1 Eingabe in Prog45mz	86
P45.7.2.2 Erläuterungen zu den Eingabe-Boxen	90
P45.7.2.3 Ausgabe aus Prog45mz	91
P45.7.3 Schritt 5d: Multiple Imputation	95
P45.7.4 Ein Experiment	104
KAPITEL 4: DATEIEN VEREINEN	109
P45.8 Datenfusion	109
P45.8.1 Schritt 6a: Datenfusion mit dem Allgemeinen Linearen Modell	113
P45.8.1.1 Eingabe in Programm Prog45mw zur einseitigen Datenfusion	114
P45.8.1.2 Erläuterungen zu den Eingabe-Boxen	118
P45.8.1.3 Ausgabe aus Prog45mw	126
P45.8.2 Schritt 6b: Datenfusion mit der Logitanalyse	131
P45.8.2.1 Eingabe in Prog45my	132
P45.8.2.2 Erläuterungen zu den Eingabe-Boxen	135
P45.8.2.3 Ausgabe aus Prog45my	136
P45.8.3 Schritt 6c: Fusionierte Dateien vereinen	140
P45.8.3.1 Eingabe in Prog45mx	141
P45.8.3.2 Erläuterung zu den Eingabe-Boxen	144
P45.8.3.3 Ausgabe	148
P45.8.3.4 Weiterführende Hinweise	148

KAPITEL 5: MEHRERE VARIABLE ZU EINER MESSUNG KOMBINIEREN.....	149
P45.9 Schritt 7a: Aus mehreren Variablenwerten einen Gesamtpunktwert bilden	149
P45.9.1 Erläuterung zu den Eingabe-Boxen.....	152
P45.9.2 Ausgabe der Ergebnisse.....	156
P45.10 Schritt 7b: Mit Faktorenanalyse einen gewichteten Gesamtpunktwert (Faktorwert) bilden	156
P45.10.1 Erläuterungen zu den Eingabe-Boxen	158
P45.10.2 Ausgabe aus Prog45ms.....	160
P45.10.3 Beispiel für eine Faktorenanalyse aus der Umfrageforschung.....	169
P45.10.4 Weiterführende Hinweise.....	172
P45.11 Schritt 7c: Rasch-Skalierungsverfahren	173
P45.11.1 Eingabe mit Prog45mr.....	176
P45.11.2 Erläuterungen zu den Eingabe-Boxen	179
P45.11.3 Ausgabe aus Programm P45mr.....	184
P45.11.4 Einschreiben der Skalenwerte in den Datensatz.....	191
P45.11.5 Weitere in Almo vorhandene Skalierungsverfahren	193
P45.11.6 Faktorenanalyse oder Rasch-Skalierung.....	193
Schlagwortverzeichnis.....	193

P45 Almo-Data-Mining

P45.0 Einführung

Data Mining heißt: „Zusammenhänge und Zusammengehörigkeiten in einer Datentabelle entdecken“.

Das Almo-Data-Mining-System besteht aus einer Folge von statistischen Programmen, mit denen eine **Standardauswertung** einer Datentabelle durchgeführt werden kann.

Die Datenmatrix

Die Datenmatrix (oder Datentabelle) besitzt eine rechteckige Form. Ihre Zeilen nennen wir „Objekte“ und ihre Spalten „Variable“. Selbstverständlich kann die Datentabelle auch transponiert sein, so dass die Variablen die Zeilen und die Objekte die Spalten bilden (siehe dazu Handbuch Teil 2, Abschnitt 42.3).

Betrachten wir ein Beispiel

	Alter	Geschlecht	Kauf
Konsument1	34	w	ja
2	64	m	nein
3	17	w	ja
.	.	.	.
.	.	.	.
.	.	.	.

Die Datenmatrix, die sich im Computer befindet, hat dann folgende Gestalt

1	34	1	1
2	64	2	2
3	17	1	1
.	.	.	.
.	.	.	.
.	.	.	.

Dabei wurde Geschlecht w=1 und m=2 kodiert und Kauf ja=1 und nein =2.

Die „Objekte“ dieser Datentabelle sind Konsumenten. Die „Variablen“ sind das Alter, das Geschlecht und der Kauf (von z.B. Musik-CDs). Die Variablen sind Merkmale, die den Objekten anhaften.

Einzelne Zellen der Datenmatrix können leer sein, weil die entsprechenden Daten fehlen. Durch speziell adaptierte Verfahren und *Imputations*-Algorithmen können die dabei auftretenden Probleme gelöst werden. Auch ist es möglich, dass die Datenmatrix durch *Fusion* mehrerer verschiedener Dateien mit nur teilweise identischen Variablen und Objekten entsteht.

„Zusammenhänge zwischen Variablen“ entdecken wir im Almo-Data-Mining-System durch Korrelieren, durch Tabellieren, durch das „Allgemeine Lineare Modell“ und die Logitanalyse.

„Zusammengehörigkeiten zwischen Objekten“ entdecken wir durch die Clusteranalyse.

„Zusammenhänge entdecken“ können wir auf 2 Weisen. Wir suchen

(1) „blind“ oder

(2) „gezielt“ nach Zusammenhängen.

Wir suchen „blind“, wenn wir eigentlich nicht wissen, nach was wir suchen. Wir haben Daten aber keine Fragen. Im Almo-Datamining-System suchen wir blind, in dem wir jede Variable mit jeder anderen korrelieren. Dabei treten einige Probleme auf, die wir später behandeln werden.

Wenn wir die ermittelten Korrelationen betrachten, dann stoßen wir auf Variable, die mit anderen interessante Zusammenhänge aufweisen. Wir werden angeregt, Fragen zu stellen und gelangen so in das Stadium des „gezielten Suchens“.

Wir suchen „gezielt“, wenn uns eine oder mehrere Variable besonders interessieren (wir nennen sie „Zielvariable“) und wir deren Determinanten erkunden wollen. Das „gezielte“ Suchen erfolgt im Almo-Datamining-System hauptsächlich durch das Allgemeine Lineare Modell aber auch durch Tabellieren und die Logitanalyse.

Betrachten wir unser kleines Beispiel:

Wir stellen etwa fest, daß zwischen dem Alter und dem Kauf von Musik-CDs ein Variablen-Zusammenhang besteht: Je jünger umso eher werden Musik-CDs gekauft.

Wir stellen weiterhin fest, daß bestimmte Objekte *zusammengehören* (ein Cluster bilden). So können wir z.B. feststellen, daß die Konsumenten 1, 3, 9, 13, ... etc. zusammengehören. Sie sind jung, weiblich und kaufen Musik-CDs. Ein weiteres Cluster könnte etwa von den Konsumenten 2, 8, 12, ... etc. gebildet werden. Sie sind alt, männlich und kaufen keine Musik-CDs.

Von einem Data-Mining-Programm wird gefordert, daß es dem Benutzer nicht zu viele Statistikkenntnisse abverlangt, daß es einfach zu bedienen ist und daß es viele seiner Ergebnisse grafisch darstellt.

Das Almo-Datamining-Programm versucht diese Forderungen zu erfüllen. Der Benutzer kann die Ergebnisausgabe steuern. Er kann einstellen, ob er mehr Grafik oder mehr Zahlenmaterial von Almo geliefert haben möchte. Auch die Komplexität der Ergebnisse kann über mehrere Stufen eingestellt werden.

P45.0.1 Anwendungsbereiche für den Data Mining Prozeß

Wir definierten das Almo-Data-Mining-System als ein Standard-Auswertungssystem für Datentabellen. Als solches kann es überall da angewendet werden, wo Daten, die oben beschriebene Form der rechteckigen Datenmatrix einnehmen oder in eine solche gebracht werden können. Das können Kundendaten sein, Daten aus der Umfrageforschung, Daten aus wissenschaftlichen Untersuchungen etc.

P45.0.2 Die Schritte des Data-Mining-Prozesses

Die Auswertung einer Datentabelle verläuft in folgenden Schritten:

Kapitel 1: Eine Almo-Arbeitsdatei erstellen

Prog45md: Schritt 1a: Daten aus Excel nach Almo übertragen
und daraus eine Almo-Arbeitsdatei erstellen
Prog45mh: Schritt 1b: Daten, die im Format FREI oder FIX vorliegen
in eine Almo-Arbeitsdatei schreiben

Kapitel 2: Daten kennenlernen

Prog45mg: Schritt 2: Daten anschauen
Prog45ml: Schritt 3: Kennwerte der Variablen anschauen
Prog45m4: Schritt 4: Variable auszählen

Kapitel 3: Daten bereinigen

Prog45mo: Schritt 5a: Mittelwerte für fehlende Werte einsetzen
Prog45mm: Schritt 5b: Prognosewerte für fehlende Werte einsetzen (mit ALM)
Prog45mz: Schritt 5c: Prognosewerte für fehlende Werte einsetzen (mit
Logitanalyse)
Prog00ml: Schritt 5d: Rechenhilfe für multiple Imputation

Kapitel 4: Dateien vereinen

Prog45mw: Schritt 6a: Dateien fusionieren (mit ALM)
Prog45my: Schritt 6b: Dateien fusionieren (mit Logitanalyse)
Prog45mx: Schritt 6c: Fusionierte Dateien vereinen

Kapitel 5: Mehrere Variable zu einer Messung kombinieren

Prog45mv: Schritt 7a: Aus mehreren Variablenwerten einen Gesamtpunktwert
bilden
Prog45ms: Schritt 7b: Mit Faktorenanalyse einen gewichteten Gesamtpunktwert
bilden
Prog45mr: Schritt 7c: Mit Rasch-Skalierungsverfahren aus mehreren Variablen
einen Messwert bilden

Teil II

(enthalten in Almo-Dokument Nr. 25)

Kapitel 6: Zusammenhänge "blind" suchen

Prog45m6: Schritt 8: Variable miteinander korrelieren

Kapitel 7: Einzelne Zusammenhänge genauer untersuchen

Prog45mb: Schritt 9: Variable tabellieren
Prog45m7: Schritt 10: Streudiagramm für 2 oder 3 Variable

Kapitel 8: Mehrfach-Zusammenhänge untersuchen

Prog45mf: Schritt 11a: Ursachen für die Zielvariable: Allgemeines Lineares
Modell
Prog45gw: Schritt 11b: Alternative wenn Zielvariable polytom: Gewichtetes ALM
Prog45m9: Schritt 11c: Alternative wenn Zielvariable nominal (dichotom,
polytom):
Logit-Analyse

Kapitel 9: Prognose leisten

Prog45mp: Schritt 12a: Werte für Zielvariable mit ALM prognostizieren
Prog45mt: Schritt 12b: Werte für Zielvariable mit Logitanalyse prognostizieren

Kapitel 10: Zusammengehörigkeiten zwischen Objekten suchen

Prog45mn: Schritt 13: Cluster von Objekten bilden: Clusteranalyse

Kapitel 11: Ursachen für die Clusterzugehörigkeit

Prog45mq: Schritt 14a: Ursachen für die Clusterzugehörigkeit: ALM

Prog45mu: Schritt 14b: Ursachen für die Clusterzugehörigkeit: Logit-Analyse

Im folgenden werden wir einen Schritt nach dem anderen ausführlich behandeln.

P45.0.3 Wie man mit Almo arbeitet

Wenn Sie mit Almo noch nicht gearbeitet haben, dann lesen Sie zunächst das Almo-Dokument Nr. 0 "Arbeiten mit Almo" und den 3. Teil des Data-Mining-Dokuments "Arbeiten mit dem Almo-Datenanalyse-System".

Wie der Almo-Texteditor zu bedienen ist, wird dort und besonders ausführlich im Almo-Handbuch, Teil 1, Bedienungsanleitung ausführlich beschrieben. Den Texteditor wird man hauptsächlich, dazu verwenden, um im Almo-Output zu editieren, z.B. um Teile des Outputs "herauszuschneiden" und in ein eigenes Dokument einzufügen. Die schnellen Tastenkombinationen, mit denen die vielfältigen Editorfunktionen durchgeführt werden können, sind ebenfalls im Dokument Nr. 0 und im Almo-Handbuch, Teil 1 dargestellt. Auch die Bedienung des Almo-Grafikeditors ist dort ausgeführt. Mit ihm können Sie die von Almo im Output automatisch erstellten Grafiken in vielfältiger Weise verändern.

P45.0.4 Unsere Testdaten

Wir werden überwiegend, aber nicht ausschließlich eine simulierte Datentabelle verwenden – mit 1.000 Personen und 10 Variablen. Die Daten beschreiben folgenden Sachverhalt: Personen kaufen bestimmte Produkte auf Kredit. Die Frage lautet nun: Zahlen sie ihren Kredit zurück? Von den 1000 Personen kennen wir folgende Variable.

V1 = Wohnort:Stadt, Land;
V2 = Geschlecht:männlich, weiblich;
V3 = Beruf:Selbständig, Unselbständig;
V4 = Einkommen;
V5 = Kinderzahl;
V6 = Hausbesitz:kein Haus, hat Haus;
V7 = Rückzahlungsrate;
V8 = Kreditlaufzeit;
V9 = gekauftes Produkt:Kleidung, Möbel, Technik;
V10 = Rückzahlung:nein, ja;

Die Daten sind in 3 Formaten unter folgenden Namen in Almo enthalten

- (1) ./Testdat/DatMin.txt
- (2) ./Testdat/DatMin.fre
- (3) ./Testdat/DatMin.dir

Datei 1 befindet sich im Format "Tabulator-getrennt" ("tab-delimited"). In diesem Format können z.B. Daten aus Excel oder SPSS ausgegeben werden oder umgekehrt geladen und angeschaut werden.

Datei 2 befindet sich im „freien“ Format. Die Zahlenwerte sind durch Leerzeichen voneinander getrennt. Die Datei kann in Almo geladen und angeschaut werden.

Datei 3 befindet sich im „direkten“ Format. Dies ist die Almo-Arbeitsdatei. Sie befindet sich in einem internen, binären (platzsparenden und schnellen) Format und kann nicht angeschaut werden.

P45.0.4 Format der Daten

Almo unterscheidet 3 Formate, in denen Daten von Almo gelesen und geschrieben (gespeichert) werden können. Es sind dies die Formate

FREI
FIX
DIREKT.

P45.0.4.1 Format FREI (Werte sind durch Trennzeichen getrennt)

Beispiel:

```

1 2 23.5 24, 0, 4 . . . .
2 5 35 33, 1, 7 . . . .
1 4 42 47, 0, 3 . . . .
. . . . . . . . . .
. . . . . . . . . .
. . . . . . . . . .

```

Dieses Format wird oft auch als "free-field format" oder "Ascii data" bezeichnet. Die Werte eines Datensatzes sind durch

ein oder mehrere Blanks
und/oder ein Komma

voneinander getrennt. In Almo dürfen diese Trennzeichen gemischt in einer Datenmatrix vorkommen - was eher ungewöhnlich ist und was besser vermieden wird. Die Regeln, nach denen Variablen- und Ausprägungsnamen zu schreiben sind werden detailliert im Almo-Dokument Nr.0 „Arbeiten mit Almo.PDF“ ausgeführt.

Anmerkung:

Almo kann auch Daten mit dem **Tabulatorzeichen als Trennzeichen** lesen und schreiben.

Aber: Mit der Format-Angabe „FREI“ in den Almo-Programm-Masken kann es solche Daten aber nur dann korrekt lesen, wenn die Daten von Almo selbst geschrieben wurden. "Fremde" Daten im Tabulator-getrennten Format (z.B. aus Excel) können (müssen aber nicht) Probleme machen, wenn sie von Almo mit der Format-Angabe „FREI“ eingelesen werden sollen.

Im nachfolgenden Abschnitt "P45.0.4.4 Fremdformate" werden wir zeigen wie diese "fremden" Tabulator-getrennten Daten in Almo korrekt zu lesen sind.

Der Punkt dient als Dezimalpunkt. Das Komma hinter '24' und hinter '0' ist also ein Trennzeichen. Der Punkt in '23.5' ist ein Dezimalpunkt. (im nachfolgend beschriebenen Format „erweitert-FREI“ ist auch das Komma als Dezimalzeichen möglich)

Die Werte der aufeinander folgenden Datensätze müssen nicht, wie in unserem Beispiel, exakt untereinander stehen.

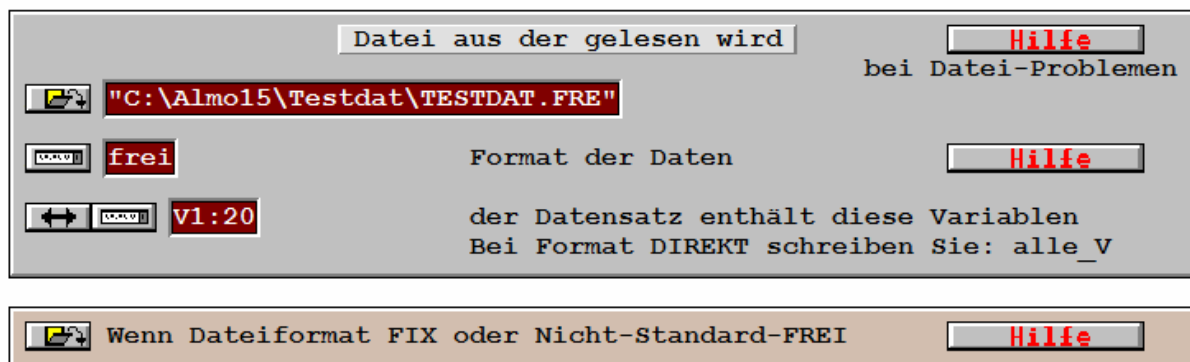
In Almo wird innerhalb des freien Formats noch zwischen

Standard-FREI und
Nicht-Standard-FREI (bzw. erweitert-Frei)

unterschieden. Das Format "Nicht-Standard-FREI" bezeichnen wir auch als "erweitert-FREI".

WICHTIG: Befinden sich die Daten des Benutzers im Format "Nicht-Standard-FREI" dann muss der Benutzer zusätzlich in der Programm-Maske die nachfolgende Optionsbox

Wenn Dateiformat FIX oder Nicht-Standard-FREI
öffnen und entsprechende Einträge vornehmen.



Standard-FREI liegt vor, wenn

- die einzulesenden Datensätze nur aus numerischen Variablen bestehen, also keine Zeichenvariablen (=Worte) vorhanden sind
- wenn bei Dezimalvariablen als Dezimalzeichen der Punkt (und nicht das Komma) verwendet wurde.
- wenn der jeweils nächste Datensatz in einer neuen Zeile beginnt - wobei ein langer Datensatz durchaus über mehrere Zeilen gehen darf. Ein neuer Datensatz muss aber prinzipiell in einer neuen Zeile beginnen. Beispiel:

```

1  2  23.5  24  0  4  77  88  2  5
27 33  4
2  5  35    33  1  7  77  93  4  2
13 21  3
      .
      .

```

Beim 1. Datensatz stehen in der 2. Zeile noch die drei Werte 27 33 4
beim 2. die Werte 13 21 3. Der 2. Datensatz beginnt in einer neuen Zeile

Nicht-Standard-FREI (bzw. erweitert-FREI)

Die Daten befinden sich im erweiterten FREI-Format, also nicht im "Standard-FREI - Format", wenn

- Zeichenvariable (=Worte) vorhanden sind.

Beispiel:

```
1 2 23.5 Beamter 24, 0, 4
2 5 35 Bauer 33, 1, 7
1 4 42 Arbeiter 47, 0, 3
```

Das Wort "Beamter" ist eine Ausprägung der Zeichenvariablen "Beruf".

- b. wenn 2 oder mehr Datensätze hintereinander in einer Zeile stehen; d.h. wenn ein neuer Datensatz nicht in einer neuen Zeile beginnt
- c. wenn bei Dezimalvariablen als Dezimalzeichen nicht der Punkt sondern das Komma verwendet wurde
- d. wenn bei Zeichenvariablen das Trennzeichen hinter der Zeichenvariablen nicht ein Blank ist sondern ein Komma oder ein Tabulatorzeichen

Wenn nur eine dieser 4 Bedingungen gegeben ist, dann befinden sich die Daten im "erweitert-FREI-Format" bzw. "Nicht-Standard-FREI-Format".

Wenn Ihre Daten sich im Format "Standard-FREI" aber auch im Format "erweitert-FREI" befinden, dann geben Sie in der Box "Datei aus der gelesen wird" im 2. Eingabefeld als Format "frei" an.

Befinden sich die Daten im Format "Nicht-Standard-FREI" (erweitert-FREI), dann muss der Benutzer zusätzlich in der Programm-Maske die nachfolgende Optionsbox

Wenn Dateiformat FIX oder Nicht-Standard-FREI
öffnen und entsprechende Einträge vornehmen.

P45.0.4.2 FORMAT FIX (Werte sind spaltengebunden)

Die Werte stehen ohne Trennzeichen direkt hintereinander.

Beispiel:

```
1223.5      Beamter2404
25 35      Bauer3317
14 42Angestellter4703
.. .      .      ...
.. .      .      ...
.. .      .      ...
```

Hier kommt noch die Anweisung hinzu:

```
Feld 1,1,4,12,2,1,1
```

Die Feldbreite der einzulesenden Variablen muss angegeben werden. Diese schreiben Sie in der nachfolgenden Optionsbox. In unserem Beispiel hat die erste Variable eine Feldbreite von 1, die 2. Variable von 1, die 3. Variablen (23.5) von 4 etc. Die 4. Variable ist eine Zeichenvariable. Ihre Ausprägung lautet " Beamter". Sie besitzt eine Feldbreite von 12 wobei 5 Blanks links vor dem Wort "Beamter" stehen. Die Feldbreite muss 12 sein, damit auch der Beruf "Angestellter" in das Feld hineinpasst. Die Daten werden üblicherweise "rechtsbündig" geschrieben. Sie dürften aber auch "linksbündig" oder "mitten hinein" geschrieben werden. Der 2 Datensatz könnte also auch so geschrieben sein:

```
rechtsbündig:      25 35      Bauer3317
```

<i>linksbündig:</i>	2535	Bauer	3317
<i>"mitten hinein":</i>	25 35	Bauer	3317

P45.0.4.3 FORMAT DIREKT

Dies ist ein Almo-spezifisches binäres Format. Es ist sehr schnell, erlaubt den DIREKT-Zugriff auf ausgewählte Datensätze (sogar Variable). Es ist z.B. möglich die Variable V48 im Datensatz 500 zu lesen, ohne dass die Datensätze vom 1. bis zum 500. sukzessive eingelsen werden müssen. Weiterhin ermöglicht dieses Format das platzsparende Speichern der Daten auf Platte (oder sonstigem Datenträger) mit 1 bis 8 Byte je Variablenwert (je nach Größe des Wertes).

Eine von Ihnen im Format FREI oder FIX geschriebene Datei kann mit den Maskenprogrammen Prog45mh, Prog00m8, Prog00m9 in das Format DIREKT gewandelt werden. Diese Maskenprogramme finden Sie durch Klick auf den Knopf "Verfahren" und dann Klick auf den Eintrag "Datei-Operationen". Dort finden Sie auch Programme für den umgekehrten Transfer von DIREKT zu FREI oder FIX.

EMPFEHLUNG: Wenn Sie mit denselben Daten mehrere Analysen durchführen wollen, dann ersparen Sie sich Arbeit, wenn Sie die Daten zuerst in eine Datei im Format DIREKT einschreiben.

Klicken Sie auf den Knopf "Verfahren" und dann auf den Eintrag "Datei-Operationen". Eines der dort angegebenen Maskenprogramme wird sicherlich in der Lage sein, das Format Ihrer Daten korrekt zu erfassen und dann Ihre Daten in ein Almo-lesbares Format zu bringen. Am besten geeignet ist wahrscheinlich Prog45mh.

P45.0.4.4 Fremdformate

Wenn Ihre Daten in einem Fremdformat vorliegen, dann können sie relativ problemlos in die 3 Almo-Formate konvertiert werden, sofern sie sich in einem der beiden folgenden Formate befinden

- (1) dem Format Tabs getrennt (*.txt)
 bzw. tab-delimited (*.txt)
 bzw. Tabulator-getrennt (*.dat)
- (2) dem Format *.sav von SPSS

Das 1. Format ist das "Tabulator-getrennte" Format. Sehr häufig wird es durch die Datei-Endung *.txt gekennzeichnet, in manchen Programmen (vor allem in Statistik-Programmen) auch durch die Endung *.dat

Das 2. Format ist ein spezifisches binäres SPSS-Format. Eine Datei im sav-Format enthält neben den Daten auch die Variablennamen ("variable names") und Ausprägungsnamen ("value labels").

Befinden sich die "fremden" Daten *nicht* in einem dieser beiden Formate, dann ist es häufig möglich, die Daten in einem Programm zu öffnen, das über vielfältige Konvertierungsoptionen verfügt. So könnte man beispielsweise die "fremden" Daten mit Excel öffnen und dann als "Tabs getrennt (*.txt)" speichern und danach mit dem im folgenden vorgestellten Almo-Programm Prog45md in das Almo-Format konvertieren.

P45.0.4.5 Konvertieren von Fremdformaten in Almo-Formate

Daten im Tabulator-getrennten Format können durch die 3 Almo-Maskenprogramme

Prog00m4 Prog00m6 Prog45md

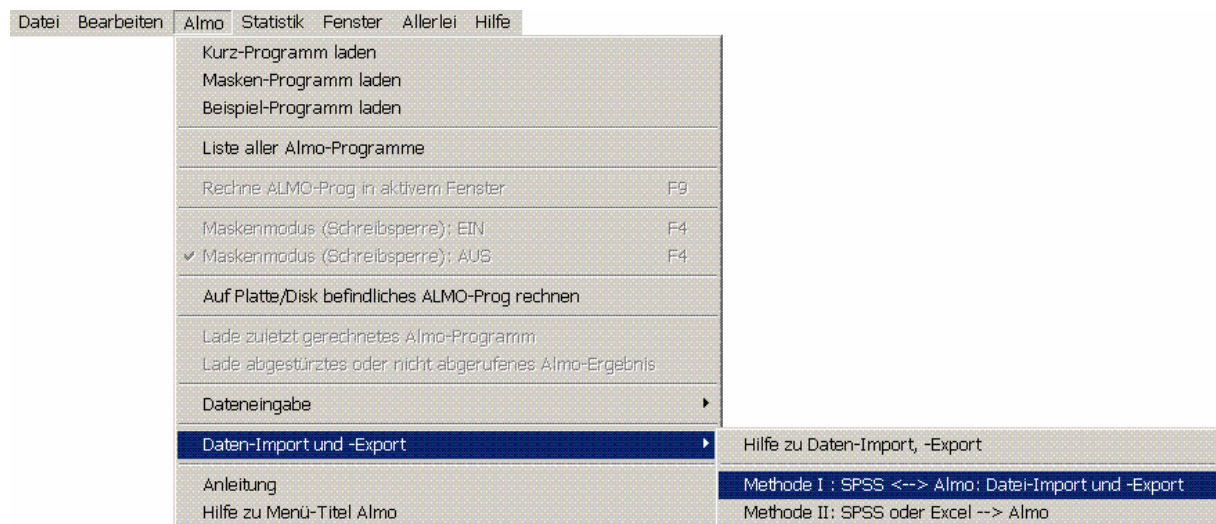
In die Almo-Formate übertragen werden. Die ersten beiden werden gefunden durch Klick auf den Knopf „Verfahren“, dann „Datei-Operationen“. Das 3. Programm wird über das Menü „Datamining“ gefunden. Es wird im folgenden dargestellt und ausführlich erläutert. Die 3 Programme sind im Prinzip identisch und austauschbar. Sie übertragen die Daten und auch die Variablennamen nach Almo - nicht jedoch die Ausprägungsnamen.

.Daten im sav-Format von SPSS können in Almo sehr komfortabel in die Almo-Formate konvertiert werden. Klicken Sie im Menü ALMO auf den Eintrag

Daten-Import und -Export

dann auf

Methode I : SPSS <--> Almo: Datei-Import und -Export



Almo startet dann das Konvertierungsprogramm von Joachim Gerich. Zuvor sollten Sie jedoch die Hilfe und danach das Manual zu diesem Programm lesen. Klicken Sie dazu auf den 1. Eintrag

Hilfe zu Daten-Import, -Export

Der Transfer der Datenmatrix von SPSS nach Almo ist unproblematisch. Probleme können jedoch beim Transfer der "variable names" und der "value labels" von SPSS in die Variablennamen und Ausprägungsnamen von Almo entstehen. Die Regeln, nach denen "variable names" (Variablennamen) und "value labels" (Ausprägungsnamen) in SPSS und Almo geschrieben werden, stimmen nicht vollständig überein. Die Regeln, nach denen in Almo Variablen- und Ausprägungsnamen zu schreiben sind werden im Almo-Dokument Nr.0 Arbeiten mit Almo.PDF im Abschnitt P03.x ausführlich behandelt. Will der Benutzer Komplikationen vermeiden, dann sollte er die Variablen- und Ausprägungsnamen, die das Konvertierungsprogramm von Gerich ausgibt nicht verwenden. Er sollte die Namen selber schreiben.

Kapitel 1: Eine Almo-Arbeitsdatei erstellen

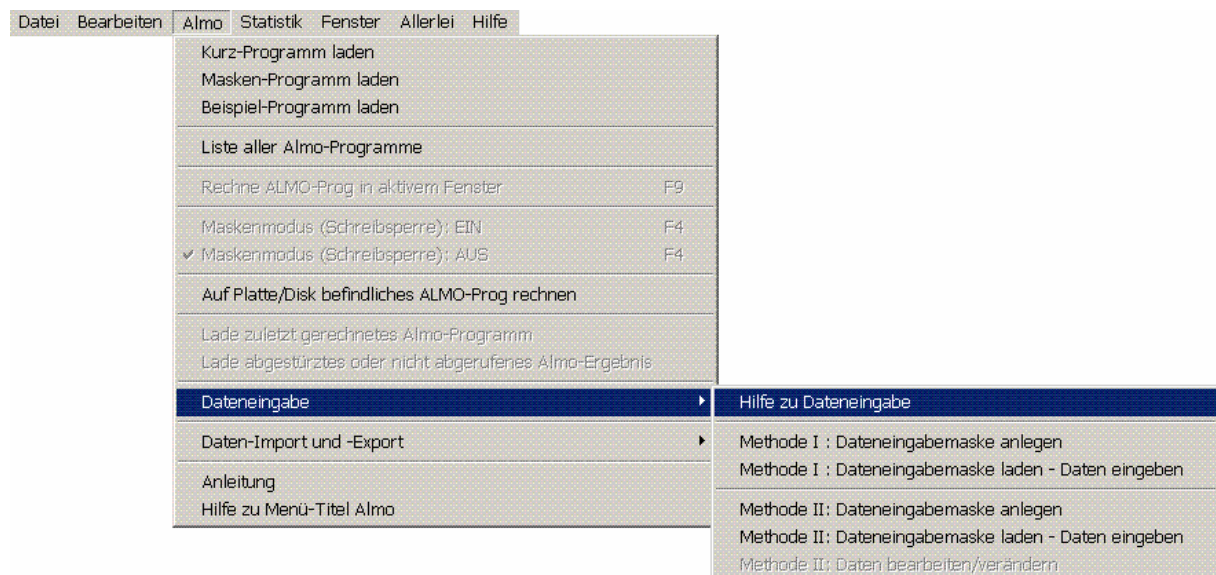
Einige Programme des Almo-Data-Mining-System benötigt Daten, die in dem speziellen Almo-Format DIREKT vorliegen. Aus den vorhandenen Daten müssen wir zuerst eine Almo-Arbeitsdatei in diesem Format erstellen.

Wurden die Daten noch nicht in den Computer geschrieben, d.h. liegt noch keine Daten-Datei vor, dann bietet Almo 3 Möglichkeiten an, eine Daten-Datei zu erzeugen:

P45.0 Daten schreiben

1. Methode I verwendet das spezielle Eingabe-Prog "WinMask"
2. Methode II verwendet ein anderes spezielles Eingabe-Prog
3. Methode III Daten einfach in ein neues leeres Fenster schreiben

Methode I und II verwenden spezielle Eingabe-Programme. Beide Programme werden in Almo gestartet über das Menü "Almo/Dateneingabe". Lesen Sie zuerst die „Hilfe“.



Methode I : Das Programm „WinMask“

Dies ist ein sehr komfortables spezielles Daten-Eingabe-Programm. Es wurde von Joachim Gerich für die Eingabe von Daten in Almo oder SPSS entwickelt. Es trägt den Namen "WinMask". Die Rechte zu diesem Programm liegen bei Joachim Gerich. Die Bedienungsanleitung zu diesem Programm befindet sich im Ordner ALMO_EMK unter dem Namen "WMmanual.pdf". Eine kurze Anleitung finden Sie in diesem Ordner unter dem Namen "WManleit.pdf". Lesen Sie diese Datei mit dem Acrobat-Reader.

"WinMask" besteht aus 2 Programmen, aus (1) 'WManlege.exe', mit dem eine Eingabemaske für die eigentliche Dateneingabe erzeugt wird und (2) 'WMeingab.exe', das die Eingabemaske zur Dateneingabe präsentiert. Beide Programme können aus Almo gestartet werden. Wenn sie nach dem Start nicht auf dem Bildschirm zu sehen sind, dann befinden sie sich in der Startleiste am unteren Bildschirmrand. Beide Programme befinden sich im Almo-Ordner "Almo_Emk" und können auch von dort gestartet werden.

WinMask hat allerdings eine Einschränkung: Es sind nur numerische Variablenwerte eingebbar. Zeichenvariable (z.B. Worte) können nicht geschrieben werden.

Methode II: Almo-Maske verwenden

Etwas weniger komfortabel aber ebenfalls sehr leistungsfähig ist ein von Johann Bacher für Almo entwickeltes Daten-Eingabe-Programm. Mit diesem Programm können auch Zeichenvariable (=Worte) eingegeben werden, was gegenwärtig mit "WinMask" noch nicht möglich ist. Die Bedienungsanleitung zu diesem Programm befindet sich im Ordner ALMO_EMK unter dem Namen "Dateingb.pdf" oder "Dateingb.doc". Lesen Sie diese Datei mit dem Acrobat-Reader bzw. MS Word

Methode III: Daten in ein leeres Fenster schreiben

Diese Vorgehensweise ist empfehlenswert, wenn nicht sehr viele Daten zu schreiben sind. Sie benötigen kein spezielles Daten-Eingabe-Programm. Gehen Sie folgendermaßen vor:

1. Erzeugen Sie ein neues Almo-Fenster durch Klick auf das Menü

Datei / Neue Datei anlegen.

Wählen Sie einen Almo-Sub-Ordner, z.B. "Progs" und geben Sie der Datei einen Namen z.B. "meineDat.fre".

Dann Mausklick auf den Knopf "Öffnen". Almo erzeugt ein neues leeres Fenster. Schreiben Sie in dieses Fenster Ihre Daten - z.B. so

```
1 1 1 1 4.2 4.1 2.2 4.1 1.2 1.5
1 1 1 1 5.1 3.1 1.2 6.1 1.2 9.3
1 1 1 2 4.2 2.2 3.2 1.1 1.2 7.2
1 1 1 2 2.1 1.2 4.1 2.2 1.1 3.0
1 1 2 1 4.2 3.2 1.2 8.2 1.1 5.5
1 1 2 1 4.1 5.2 1.2 9.2 1.1 4.9
1 1 2 2 2.2 2.2 3.2 3.2 1.2 2.7
1 1 2 2 4.2 4.1 4.2 1.1 1.1 2.8
1 1 3 1 3.1 3.1 2.2 7.2 1.1 4.4
. . . . .
. . . . .
. . . . .
```

Als Trennzeichen zwischen den Zahlen wird (mindestens) 1 Blank und/oder ein Beistrich verwendet. Das Dezimalzeichen ist ein Punkt. Die Zahlen müssen nicht so schön wie hier untereinander stehen.

Wenn Sie Ihre Dasten so schreiben, dann befinden diese sich im Format FREI und können dann mit dem nachfolgend dargestellten Programm Prog45mh problemlos in eine Almo-Arbeitsdatei (im Format DIREKT) übertragen werden. Siehe dazu den folgenden Abschnitt P45.2.

2. Erzeugen Sie ein 2. neues Almo-Fenster für die Variablennamen.

Geben Sie es auch in den Almo-Sub-Ordner "Progs" und nennen Sie es z.B. "meineDat.nam". Schreiben Sie in das Fenster die Variablennamen, z.B. so

```
Name1=Geschlecht:männlich,weiblich;
N2=Wohnort:Stadt,Land;
N3=Beruf:(1)Arbeiter,(2)Angestellter,(3)Selbständiger;
N4=Schulbildung:niedrig,hoch;

N6=Alter;
```

N7=Einkommen;
N8=Kinderzahl;
N9=Item1;
N10=Item2;

Almo unterscheidet zwischen Variablennamen (z.B. Geschlecht) und Ausprägungsnamen (z.B. männlich, weiblich). Variablennamen können in Almo 80 Zeichen lang sein, wobei zur Identifizierung der Variablen nur die ersten 32 verwendet werden. Variablen- und Ausprägungsnamen können auch entfallen. So wurde z.B. für Variable 5 kein Name vergeben. Als Analysevariable in den Almo-Programmen schreiben Sie dann einfach "V5".

P45.1 Schritt 1a: Tabulator-getrennte Daten (z.B. aus Excel) nach Almo übertragen

Die Daten des Benutzers werden in der Regel nicht innerhalb des Almo-Systems entstanden sein.

Häufig werden jedoch Daten in einem Format vorliegen, das Almo nicht kennt.

In sehr vielen Datenbanken, Tabellenkalkulationen, Data-Warehouse-Programmen und Statistikprogrammen ist es allerdings möglich, die Daten nach Excel zu übertragen. Von Excel können diese Daten dann mit dem speziellen Almo-Programm Prog45md in eine Almo-Arbeitsdatei transferiert werden.

Wir beschreiben uerst die Übertragung von Excel nach Almo.

Der Benutzer muß zuerst die Daten in Excel laden und dann dort über das Menü

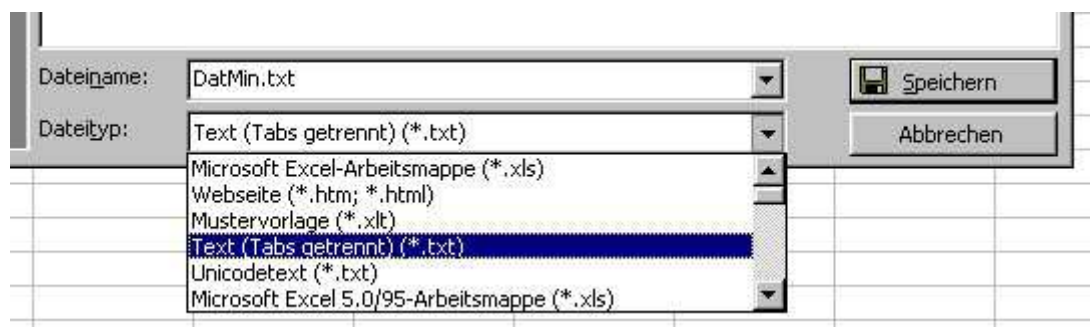
„Datei/Speichern unter ...“

ausgeben, wobei als Dateityp

„Text (Tabs getrennt) (*.txt)“

zu wählen ist. Beachten Sie: Die Daten dürfen nicht mit dem Dateityp (*.xls) oder (*.xlsx) ausgegeben werden.

In der Excel-Dialogbox „Speichern“ sieht das folgendermaßen aus (wir zeigen nur einen Ausschnitt):



Excel schreibt die Daten dann in folgender Form.

1. Die Werte (Zahlen oder Worte) werden durch das Tabulatorzeichen getrennt.
2. Ein fehlender Wert, d.h. eine leere Zelle im Excel-Datenblatt, fehlt schlicht und einfach. Man bemerkt das daran, daß z.B. zwei Tabulatorzeichen aufeinander folgen oder eine Zeile mit einem Tabulatorzeichen endet (dann fehlt dahinter ein Variablenwert).

Das Tabulatorzeichen selbst ist nicht sichtbar. Wird die *.txt-Datei in einen Text-Editor geladen, dann wird das Tabulatorzeichen üblicherweise durch ein oder mehrere Blanks angezeigt. Ein fehlender Wert ist dann eine Lücke in der Daten-Matrix. Wird die Datei in ein Almo-Fenster geladen, dann wird das Tabulatorzeichen negiert. Ein fehlender Variablenwert ist einfach nicht da. Sein linker und sein rechter Nachbar-Wert stehen unmittelbar nebeneinander.

Nachdem die Daten aus Excel in der beschriebenen Weise ausgegeben wurden, rechnen Sie mit dem nachfolgend beschriebenen Prog45md oder mit dem im Standard-Almo enthaltenen äquivalenten Programm Prog00m4.

Prog45md.Msk
Datensätze aus einer Excel-Datei lesen
und in eine Almo-Arbeitsdatei schreiben

Der Benutzer muß zuerst in Excel die Daten über das Menü
"Datei / Speichern unter ..."

ausgegeben - wobei als Dateityp anzugeben ist:
"Text (Tabs getrennt) (*.txt)"

Auch die Variablennamen in der 1. Zeile des Excel-Daten-
blatts können von Almo übernommen werden

Dieses Almo-Prog Prog45md erzeugt 3 Almo-Dateien:

1. Eine Almo-Arbeitsdatei im binären Format DIREKT
Sie kann nicht angeschaut werden
2. Eine Datei im Format FREI
Sie ist inhaltlich identisch mit der Almo-Arbeitsdatei
Sie kann jedoch angeschaut werden
3. Eine Datei der Variablennamen

Handbuch P45 Data-Mining, Abschnitt P45.1

Almo bietet eine 2. Möglichkeit, Excel-Daten nach Almo zu
übertragen. Klicken Sie im Menü "Almo" auf den Eintrag
"Daten-Import und -Export / Methode II".

Programm-Bedienung ---> Hilfe

Speicher fuer x Variable

Vereinbare Variable=

Excel-Datei aus der gelesen wird

"C:\Almo15\Testdat\DatMin.txt"

ein Datensatz enthält sovieler Variable

Zeichenvariable
Diese Variable aus der Excel-Datei sind Zeichenvariable (also Worte)

Dezimalzeichen
Punkt= Dezimalwerte sind mit dem Punkt als Dezimalzeichen geschrieben
Komma= Dezimalwerte sind mit dem Komma als Dezimalzeichen geschrieben

0= Keine Variablennamen vorhanden schon in der 1. Zeile der Excel-Datei stehen Daten

1= in der 1. Zeile der Excel-Datei stehen Variablennamen, in der 2. und den folgenden Zeilen stehen die Daten

BEACHTEN: Stehen in den oberen Zeilen sonstige Texte, so müssen diese in der Excel-Datei gelöscht werden

Kein-Wert-Angabe und Umkodierungen von Zahlenvariablen

Die Variablen werden mit ihren unkodierten Werten bzw. mit dem Kein-Wert-Code in die neue Datei übernommen

Kein-Wert-Angabe
Umkodierungen

erzeuge zusätzliche Felder für Umkodierungen / Kein_Wert-Angaben

Umkodierungen von Zeichenvariablen in Zahlenvariablen

Zeichenvariablen, die nicht unkodiert werden, gehen wieder als Zeichenvariablen in die neue Datei ein

erzeuge zusätzliche Felder für Umkodierungen

Almo-Dateien, die erzeugt werden sollen



"C:\Almo15\PROGS\DatMin"

Geben Sie den Dateinamen ohne Erweiterung an. Almo erzeugt 3 Dateien:

1. eine nicht lesbare Almo-Arbeitsdatei mit der Erweiterung __.dir
2. eine anschauliche Datei im freien Format mit der Erweiterung __.fre
3. eine Datei der Variablennamen mit der Erweiterung __.nam sofern in der 1. Zeile der Excel-Datei Namen vorhanden waren



v1:10

der Datensatz soll diese Variablen enthalten



120

Zeilenlänge

Hilfe

P45.1.1 Erläuterungen zu den Eingabe-Boxen

Eingabe-Box 1: Speicher für Variable

Geben Sie hier mindestens so viele Variable an, wie ein Datensatz besitzt. Siehe dazu Almo-Dokument Nr.0 "Arbeiten mit Almo", Abschnitt P0.1.

Eingabe-Box 2: Excel-Datei aus der gelesen wird

Excel-Datei aus der gelesen wird

ein Datensatz enthält so viele Variable

Zeichenvariable
Diese Variable aus der Excel-Datei sind
Zeichenvariable (also Worte)

Dezimalzeichen
Punkt= Dezimalwerte sind mit dem Punkt
als Dezimalzeichen geschrieben
Komma= Dezimalwerte sind mit dem Komma
als Dezimalzeichen geschrieben

0= Keine Variablennamen vorhanden
schon in der 1. Zeile der
Excel-Datei stehen Daten
1= in der 1. Zeile der Excel-Datei stehen
Variablennamen, in der 2. und den
folgenden Zeilen stehen die Daten

BEACHTEN: Stehen in den oberen Zeilen sonstige
Texte, so müssen diese in der
Excel-Datei gelöscht werden

Eingabefeld 1: Geben Sie den Pfad- und Dateinamen der Datei an, in die Sie aus Excel die Daten im „Tabs getrenntem“ Format geschrieben haben.

Eingabefeld 2: Geben Sie die Länge eines Datensatzes an, d.h. die Zahl der Variablen, aus denen ein Datensatz besteht, bzw. die Zahl der Spalten im Excel-Datenblatt.

Eingabefeld 3: Geben Sie die Nummern der Variablen an, die Zeichenvariable (also Worte) sind. Beispiel: Im Excel-Datenblatt sei die 5., 7. und 10. Spalte eine Zeichenvariable. Sie enthalten Buchstaben oder sonstige Zeichen, die nicht Zahlen sind. Geben Sie dann im 3. Eingabefeld ein: V5, 7, 10.

Eingabefeld 4: Es muss angegeben werden, welches Dezimalzeichen in den Daten verwendet wird. Sind alle Daten ganzzahlig, dann schreibt man "Punkt".

Eingabefeld 5: Es geht darum ob in der Exceldatei Variablennamen (Spaltennamen) vorhanden sind und wie sie von Almo übernommen werden können.

Man wird zu diesem Zwecke die aus Excel ausgegebene *.txt-Datei in Almo laden. Man sieht dann folgendes.

1. Möglichkeit:

```
1 02 02 03238405 01 05211 02001 02
2 01 02 03477001 01 04236 01402 02
2 02 02 025653 01 01 05545 01401 01
2 01 02 041643 0001 04748 02003 02
2 01 02 018641 01 01 01568 01003 02
2 02 01 027226 02 01 03772 01102 01
```

Die Kreise symbolisieren das eigentlich unsichtbare Tabulatorzeichen. Gleich in der 1. Zeile stehen Daten. Variablenamen sind nicht vorhanden. In das Eingabefeld wird dann 0 eingegeben (durch Klick auf dem Knopf mit nach unten weisendem Pfeil).

2. Möglichkeit:

```
Wohnort 0Geschlecht 0Beruf 0Einkommen 0Kinderzahl 0Hausbesitz
1 02 02 03238405 01 05211 02001 02
2 01 02 03477001 01 04236 01402 02
2 02 02 025653 01 01 05545 01401 01
2 01 02 041643 0001 04748 02003 02
2 01 02 018641 01 01 01568 01003 02
2 02 01 027226 02 01 03772 01102 01
```

In der 1. Zeile stehen die Spaltennamen aus dem Excel-Datenblatt, in der 2. und den folgenden Zeilen die Daten. In diesem Fall schreiben Sie in das Eingabefeld eine 1.

3. Möglichkeit:

```
Rückzahlungsdaten ■
000
Wohnort 0Geschlecht 0Beruf 0Einkommen 0Kinderzahl 0Hausbesitz
1 02 02 03238405 01 05211 02001 02
2 01 02 03477001 01 04236 01402 02
2 02 02 025653 01 01 05545 01401 01
2 01 02 041643 0001 04748 02003 02
2 01 02 018641 01 01 01568 01003 02
2 02 01 027226 02 01 03772 01102 01
```

In den ersten Zeilen stehen verschiedene Informationen. Sie müssen gelöscht werden (mit Strg+y, wieder zurückholen mit Alt+r). Dann wird die Datei neu gespeichert, entweder in der Form der 1. oder 2. Möglichkeit.

Eingabe-Box 3: Kein-Wert-Angabe und Umkodierung von Zahlenvariablen

Kein-Wert-Angabe und Umkodierungen von Zahlenvariablen

Die Variablen werden mit ihren unkodierten Werten bzw. mit dem Kein-Wert-Code in die neue Datei übernommen

Kein-Wert-Angabe

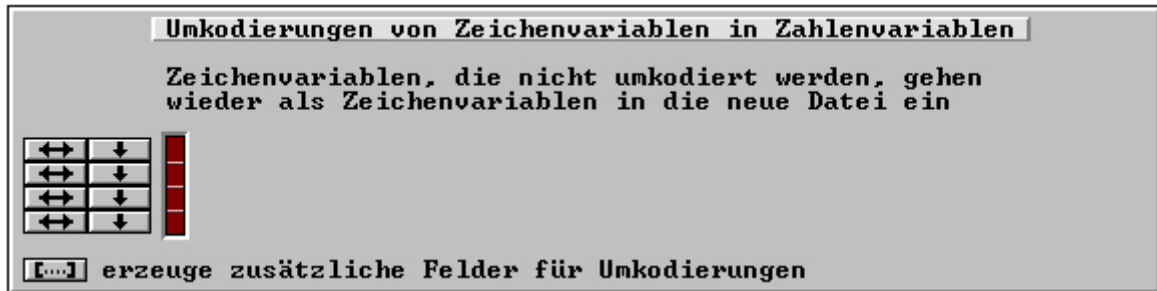
Umkodierungen

erzeuge zusätzliche Felder für Umkodierungen / Kein_Wert-Angaben

Almo unterscheidet zwischen Zahlen- und Zeichenvariablen. Zahlenvariable sind Variable, die Zahlenwerte besitzen. Zeichenvariable bestehen aus Worten, allgemein aus Zeichen, die nicht Zahlen sind. In dieser Box geht es nur um die Umkodierung von Zahlenvariablen.

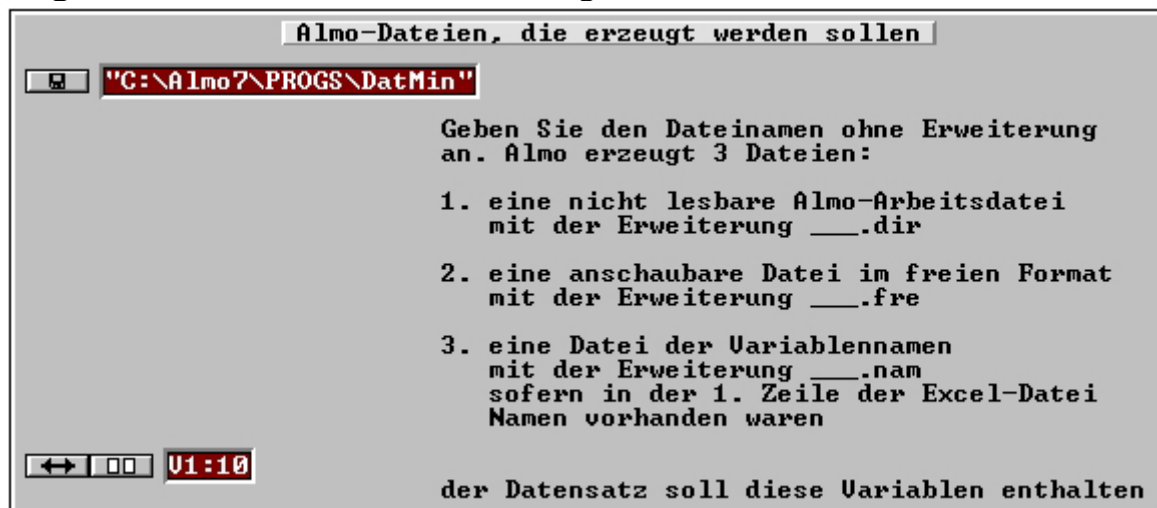
Siehe unsere ausführliche Darstellung im "Arbeiten mit Almo-Datenanalyse-System", Abschnitt P0.5.

Eingabe-Box 4: Umkodierungen von Zeichenvariablen in Zahlenvariablen



Hier geht es nur um die Umkodierung von Zeichenvariablen. Siehe P0.5
Jene Zeichenvariable, die in Eingabe-Box 4 nicht in eine Zahlenvariable umkodiert wurden, werden in die neue Datei wieder als Zeichenvariable übernommen.

Eingabe-Box 5: Almo-Dateien, die erzeugt werden sollen



Eingabefeld 1: Geben Sie hier einen Dateinamen ohne Erweiterung an
In unserem Beispiel ist dies:

"C:\Almo6\PROGS\DatMin"

Beachte: Der Dateiname darf keine Erweiterung besitzen.
Almo erzeugt nun 3 Dateien:

1. eine nicht lesbare Almo-Arbeitsdatei im Format Direkt mit der Erweiterung __.dir

In unserem Beispiel würde diese Datei dann heißen

"C:\Almo6\PROGS\DatMin.dir"

2. eine anschaulbare Datei im freien Format mit der Erweiterung __.fre

In unserem Beispiel würde diese Datei dann heißen

"C:\Almo6\PROGS\DatMin.fre"

3. eine Datei der Variablennamen mit der Erweiterung __.nam sofern in der 1. Zeile der Excel-Datei Namen vorhanden waren.

In unserem Beispiel würde diese Datei dann heißen

"C:\Almo6\PROGS\DatMin.nam"

Eingabefeld 2: Geben Sie hier an, welche Variablen aus der Excel-Datei übernommen werden sollen. In der Regel wird man alle übernehmen. In unserem Beispiel haben wir geschrieben:

V1:10

Der Doppelpunkt heißt „bis“.

Die Variablen V1 bis V10 - also alle - sollen aus der Excel-Datei übernommen werden.

Die in unserem Beispiel von Almo erzeugte Datei "C:\Almo6\PROGS\DatMin.dir" ist die Almo-Arbeitsdatei, die wir für einige der folgenden Data-Mining-Programme benötigen. Sie ist in einem binären Format geschrieben. In Almo sprechen wir vom Format DIREKT, weil dieses Format auch Direktzugriffe auf einzelne Datensätze sogar Variable erlaubt. Diese Datei kann sehr schnell eingelesen werden, kann aber nicht unmittelbar angeschaut werden.

Die in unserem Beispiel von Almo erzeugte Datei "C:\Almo6\PROGS\DatMin.fre" ist im freien Format geschrieben. Sie wird in den folgenden Data-Mining-Programmen nur selten gebraucht. Sie wird hauptsächlich erzeugt, damit sich der Benutzer die Daten anschauen kann. Zu diesem Zweck wird diese Datei in Almo geladen. Dateien im freien Format haben in Almo folgende Eigenschaften:

1. Die Werte (Zahlen oder Worte) sind durch ein Leerzeichen getrennt.
2. Enthält ein Wort selbst ein Leerzeichen, dann wird das Wort zwischen Apostrophe gestellt.
3. Enthält ein Wort einen Apostroph (z.B. d'Artagnan) dann wird der Apostroph in das Zeichen ~ gewandelt (da Almo den Apostroph als Wortbegrenzer verwendet. Siehe Punkt 2).
4. Fehlende Werte (= leere Zellen im Excel-Datenblatt) werden für Zahlenvariable durch „kw“ (= keinWert) und für Zeichenvariable durch ein zwischen Apostroph stehendes Leerzeichen vertreten.

Almo hat in unserem Beispiel weiterhin eine Datei der Variablennamen erzeugt. Ihr Name lautet "C:\Almo\Progs\DatMin.nam". Diese enthält jedoch keine Ausprägungsnamen. Für unser Beispiel entsteht folgende Namensdatei:

```
Name1=Wohnort;  
Name2=Geschlecht;  
.  
.  
.
```

Da in einer Excel-Datei in der Regel nur Variablennamen, aber keine Ausprägungsnamen gespeichert werden, müssen wir obige Namensdatei nachbearbeiten. Wir ergänzen:

```
Name1=Wohnort:Stadt, Land;  
Name2=Geschlecht:m,w;  
.  
.  
.
```

Wir geben also bei den nominalen Variablen noch die Ausprägungsnamen an.

Zur Art und Weise, wie in Almo Variablen- und Ausprägungsnamen geschrieben werden, siehe "Arbeiten mit Almo-Datenanalyse-System", Abschnitt P0.3.

P45.2 Schritt 1b: Daten im Format FREI oder FIX in eine Almo-Arbeitsdatei schreiben

Häufig wird es möglich sein, Daten aus anderen Programmen im freien oder fixen Format auszugeben. Diese Daten können, wenn sie nicht sehr ungewöhnliche Eigenschaften besitzen von Almo gelesen werden. Mit Prog45mh kann der Benutzer die Eigenschaften seiner Daten charakterisieren. Almo verarbeitet diese Angaben und erzeugt eine Almo-Arbeitsdatei im Format DIREKT.

Prog45mh.Msk
 Datensätze aus Ursprungs-Datei lesen
 und in eine Almo-Arbeitsdatei im Format DIREKT schreiben

Die Ursprungs-Datei kann Zahlenvariable und Zeichenvariable enthalten.
 Die zu lesenden Daten in dieser Ursprungs-Datei können sich im Format FREI oder FIX befinden.

Die zu schreibenden Daten werden im Format Direkt geschrieben. Dabei gibt es die Möglichkeit, die Zahlenvariablen je nach ihrer Größe platzsparend mit 1 Byte oder 2 oder 4 oder 8 Byte in die neue Datei zu schreiben.

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

1
 Vereinbare Variable= ;

2 Option: Weitere Vereinbarungen - nur wenn Almo dazu auffordert

3
 "C:\Almo7\Testdat\DatMin.fre"
 frei Format der Daten
 01:10 der Datensatz enthält diese Variablen
 Bei Format DIREKT schreiben Sie: alle_U
 Zeichenvariable
 Diese Variable aus dem Datensatz sind Zeichenvariable (also Worte)
 Punkt Dezimalzeichen
 Punkt= Dezimalwerte sind mit dem Punkt als Dezimalzeichen geschrieben
 Komma= Dezimalwerte sind mit dem Komma als Dezimalzeichen geschrieben

4
 wenn das Format der Daten FREI ist
 irrelevant wenn das Format FIX ist

 Editfeld leer = in der Datei ist jeder Datensatz vom nachfolgenden durch einen Zeilenumbruch getrennt
 ohne_Zeilenumbruch = mehrere Datensätze stehen hintereinander in einer Zeile
 Blank Trennzeichen bei Zeichenvariablen
 Hinter den Zeichenvariablen steht als Trennzeichen ein Blank oder Komma oder Tabulator

5

<Alte> Datei, aus der gelesen wird - Teil III
 wenn das Format der Daten FIX ist
 irrelevant wenn das Format FREI ist

schreiben Sie hier die Feldanweisung

Zeilenlänge
 Zahl der Spalten eines Datensatzes
 nur wenn Datensatz über mehrere Zeilen geht

Leerfeld bei Zahlenvariablen
 KeinWert = die aus einem Leerfeld bestehende
 Variable erhält den KeinWert-Code
 0 = erhält den Wert 0

Blank bei Zahlenvariablen
 0 = ein Blank mitten in oder hinter
 einer Zahl wird als 0 betrachtet
 negiert = wird als nicht existent betrachtet

6

Kein-Wert-Angabe und Umkodierungen von Zahlenvariablen

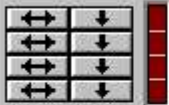
Die Variablen werden mit ihren umkodierten Werten bzw.
 mit dem Kein-Wert-Code in die neue Datei übernommen

Kein-Wert-Angabe
 Umkodierungen

erzeuge zusätzliche Felder für Umkodierungen / Kein_Wert-Angaben

Umkodierungen von Zeichenvariablen in Zahlenvariablen

Zeichenvariablen, die nicht umkodiert werden, gehen wieder als Zeichenvariablen in die neue Datei ein



erzeuge zusätzliche Felder für Umkodierungen

Zeichenvariablen in der neuen Datei

Das sind jene, die oben nicht in eine Zahlenvariable umkodiert wurden

[Hilfe](#)



oben: Zeichenvariable
darunter: deren maximale Länge, mit der sie in die neue Datei geschrieben werden sollen

Neue Datei (im Format DIREKT), die erzeugt werden soll

"C:\Almo7\PROGS\DatMin"

Geben Sie den Dateinamen ohne Erweiterung an
Der Pfad- und Dateiname muss anders lauten als der der alten Datei !!

Almo erzeugt 2 Dateien:

1. eine nicht lesbare Almo-Arbeitsdatei mit der Erweiterung `__.dir`
2. eine anschauliche Datei im freien Format mit der Erweiterung `__.fre`

U1:10

der Datensatz soll diese Variablen enthalten

Option: Inkludiere Programmteil zum platzsparenden Speichern

[Programmende](#)

P45.2.1 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 2: (Alte) Datei aus der gelesen wird



Eingabefeld 1: Geben Sie zuerst den Namen der Datei an, in der sich Ihre Daten befinden. Das machen Sie dadurch, daß Sie diesen Namen in das Editfeld schreiben oder einfach dadurch, daß Sie auf den Knopf mit dem Disketten-Symbol klicken. Es erscheint dann eine Dialogbox, in der Sie gefragt werden, ob Sie die Datei anschauen wollen oder ob Sie in das Editfeld einen (neuen) Namen einsetzen wollen. Klicken Sie auf "Namen einsetzen". Es erscheint dann die gewohnte Datei-Auswahlbox, in der Sie die Datei suchen, in der sich Ihre Daten befinden.

*Eingabefeld 2: **Format der Daten:***

Almo unterscheidet 3 Formate, in denen die Daten vorliegen können. Siehe Handbuch, Teil 2, Abschnitt 9.2 und 14.

1. FORMAT FREI

Beispiel: 1 2 23.5 Beamter 23, 0, 4

Die Werte eines Datensatzes sind durch Blank und/oder Komma und/oder das Tabulatorzeichen voneinander getrennt. Der Punkt dient als Dezimalpunkt.

2. FORMAT FIX

Die Werte stehen ohne Trennzeichen direkt hintereinander.

Beispiel:1223.5Beamter2304.....

3. FORMAT DIREKT

Dies ist ein Almo-spezifisches binäress Format. Es ist sehr schnell erlaubt den DIREKT-Zugriff auf ausgewählte Datensätze (sogar Variable) und ermöglicht das platzsparende Speichern der Daten auf Platte mit 1 bis 8 Byte (je nach Größe des Variablenwertes).

Klicken Sie auf den Knopf vor dem Editfeld und wählen Sie eines der 3 Formate aus.

Eingabefeld 3: Der Datensatz enthält diese Variable:

Geben Sie z.B. V1:20 an, wenn Ihr Datensatz 20 Variable umfasst. Der Doppelpunkt heißt „bis“.

Befinden sich Ihre Daten im Format DIREKT, dann schreiben Sie "alle_V" oder klicken Sie auf den Knopf vor dem Editfeld und treffen Sie Ihre Auswahl.

Wenn Ihre Datei im Format DIREKT vorliegt, dann negiert Also alle Eingaben in der nachfolgenden Eingabe-Box 3. Sie brauchen sich also um die Eingaben in der 2. Eingabe-Box nicht zu kümmern.

Eingabefeld 4: Zeichenvariable:

Folgende Datei liege vor

```
1 maennlich 25
2 weiblich 22
.      .      .
.      .      .
```

V1 ist eine fortlaufende Nummer. V2 ist das Geschlecht. V3 ist das Alter. V2 ist eine Zeichenvariable. Sie besteht nicht aus einem Zahlenwert, sondern aus Buchstaben (und sonstigen Zeichen). Also hat nun keine Problem, Zeichenvariable einzulesen. Der Benutzer muß lediglich angeben, welche Variable Zeichenvariable sind.

Schreiben Sie die Zeichenvariablen in das Eingabefeld oder klicken Sie zu diesem Zweck auf den Knopf mit den zwei kleinen Fenstersymbolen. Also präsentiert Ihnen dann die Variablen-Auswahl-Box. In ihr klicken Sie auf die Variablen, die Zeichenvariable sind.

Sollen Zeichenvariable als Analysevariable in irgend einer Weise ausgewertet werden (z.B. in eine Varianzanalyse als unabhängige, nominale Variable eingehen), dann müssen sie in Zahlenvariable umkodiert werden. Siehe dazu "Arbeiten mit Also-Datenanalyse-System", Abschnitt P0.5. Zu den Zeichenvariablen siehe Handbuch, Teil 2, Abschnitt 45.

Eingabefeld 5: Dezimalzeichen:

Die folgenden Ausführungen gelten nur für das Format FREI oder FIX. Daten, die Dezimalwerte besitzen, werden normalerweise so geschrieben sein:

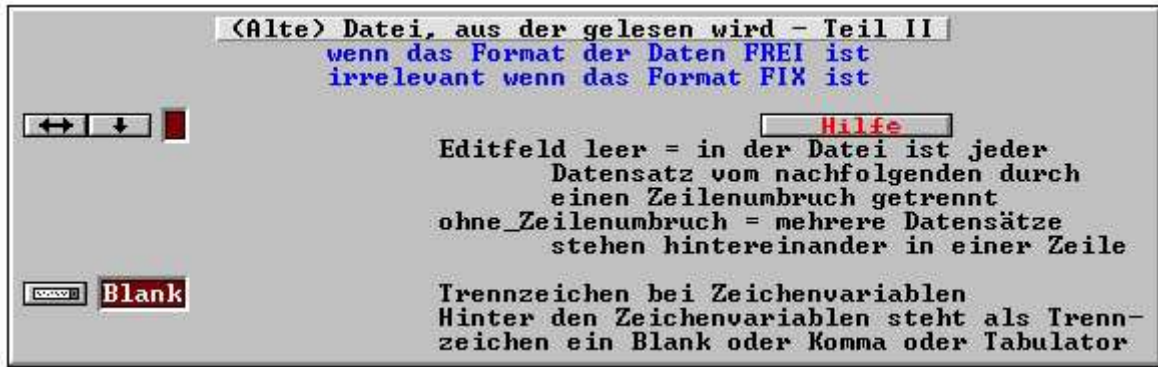
```
1.2    2.34    4.2    27.123
2.1    4.44    1.4    31.231
```

Als Dezimalzeichen wird der (Dezimal-) Punkt verwendet. Also ermöglicht es jedoch auch, Daten zu lesen (und zu schreiben), die als Dezimalzeichen das Komma verwenden. Beispiel:

```
1,2    2,34    4,2    27,123
2,1    4,44    1,4    31,231
```

Klicken Sie auf den Knopf mit dem nach unten weisenden Pfeil. Es erscheint eine Auswahlliste mit den beiden Einträgen "Punkt" und "Komma". Klicken Sie auf den zutreffenden Eintrag.

Eingabe-Box 3: (Alte) Datei aus der gelesen wird - Teil II



Eingabefeld 1: Datensätze mit Zeilenumbruch oder hintereinander:

Die folgenden Ausführungen gelten nur für das Format FREI.

Normalerweise werden Daten in folgender Weise geschrieben sein:

```

1 3 40 25
2 8 50 36
7 9 32 14
0 0 17 12

```

Jeder Datensatz beginnt in einer neuen Zeile. Anders formuliert: Die Datensätze sind durch einen Zeilenumbruch voneinander getrennt. Dabei kann ein sehr langer Datensatz durchaus über mehrere Zeilen gehen. Entscheidend ist, daß der nachfolgende Datensatz wieder in einer neuen Zeile beginnt.

Nun könnten die Datensätze auch in folgender Weise geschrieben sein:

```

1 3 40 25      2 8 50 36      7 9 32 14
0 0 17 12      7 5 23 22      1 0 87 12

```

Hier sind drei Datensätze ohne Zeilenumbruch hintereinander geschrieben. Erst nach drei Datensätzen erfolgt in diesem Beispiel ein Zeilenumbruch. Dabei wäre es in ALMO sogar erlaubt, mitten in einem Datensatz einen Zeilenumbruch zu machen (allerdings nicht mitten in einer Zahl).

Dies muß ALMO mitgeteilt werden. Klicken Sie auf den Knopf mit dem Pfeil nach unten. In das Editfeld wird dann eingetragen: "ohne_Zeilenumbruch"

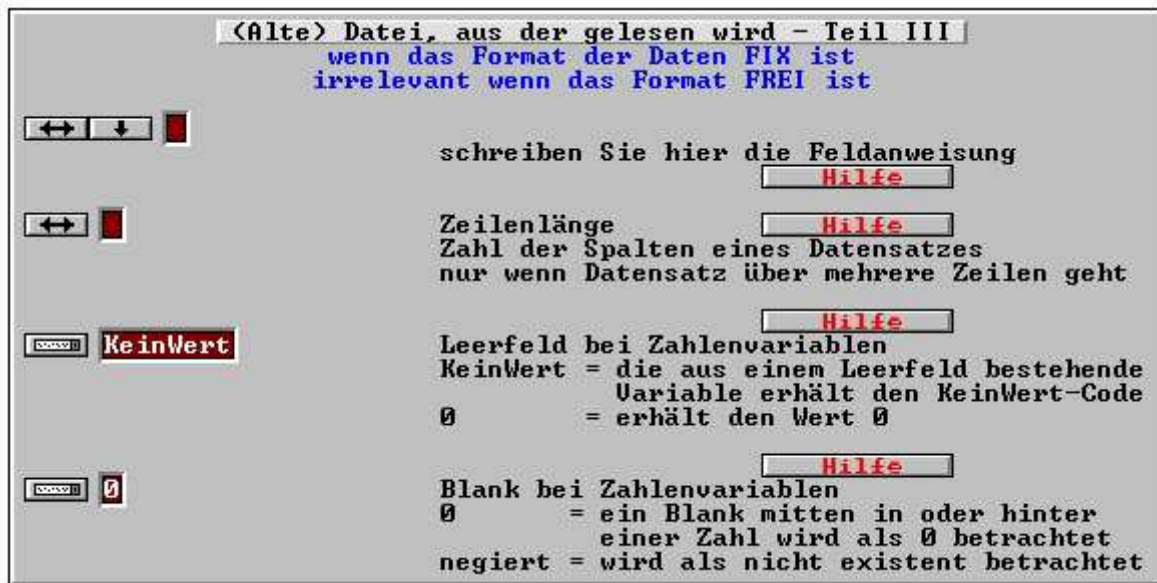
Eingabefeld 2: Trennzeichen:

Nur bedeutsam bei Format FREI

Als Trennzeichen zwischen den Variablenwerten akzeptiert ALMO:

- Blank
- Komma
- Tabulatorzeichen

Eingabe-Box 4: (Alte) Datei aus der gelesen wird - Teil III



Eingabefeld 1: Feldanweisung bei Format FIX:

Beim Format FIX stehen die Werte ohne Trennzeichen direkt hintereinander. Die Werte könnten z.B. folgende sein

```
1 2 4 0 23.55 Beamter 23 0 4
```

Im fixen Format würden sie dann folgendermaßen geschrieben sein

```
124023.55Beamter2304
```

Die Feldanweisung würde in diesem Fall lauten

```
Feld 1,1,1,1,5,7,2,1,1  
oder kürzer: Feld 4*1,5,7,2,1,1
```

Sie können also anstelle 1,1,1,1 kurz schreiben 4*1. Vergessen Sie nicht das Wort "Feld".

Eingabefeld 2: Zeilenlänge

Geht ein Datensatz nur über eine Zeile, dann ist die Angabe einer Zeilenlänge irrelevant - egal wie lang der Datensatz ist, ob z.B. 80 Spalten lang oder 170 Spalten.

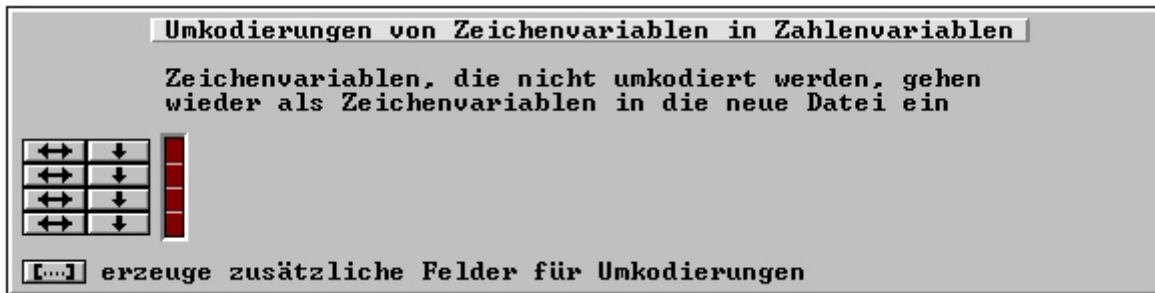
Geht ein Datensatz über 2 oder mehrere Zeilen, dann muss die Zeilenlänge angegeben werden. Die Angabe der Zeilenlänge und die Feldanweisung kann dann kompliziert werden. Siehe dazu Handbuch, Teil2, Abschnitt 9.2.1.5

Eingabefeld 3: Leerfeld

Von einem "Leerfeld" sprechen wir, wenn das Feld, das eine Variable einnimmt, aus Blanks besteht

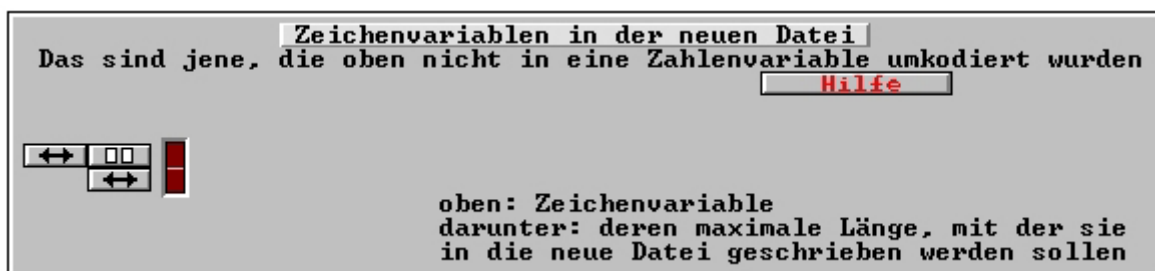
Beispiel:

Eingabe-Box 6: Umkodierungen von Zeichenvariablen in Zahlenvariablen



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

Eingabe-Box 7: Zeichenvariablen in der neuen Datei



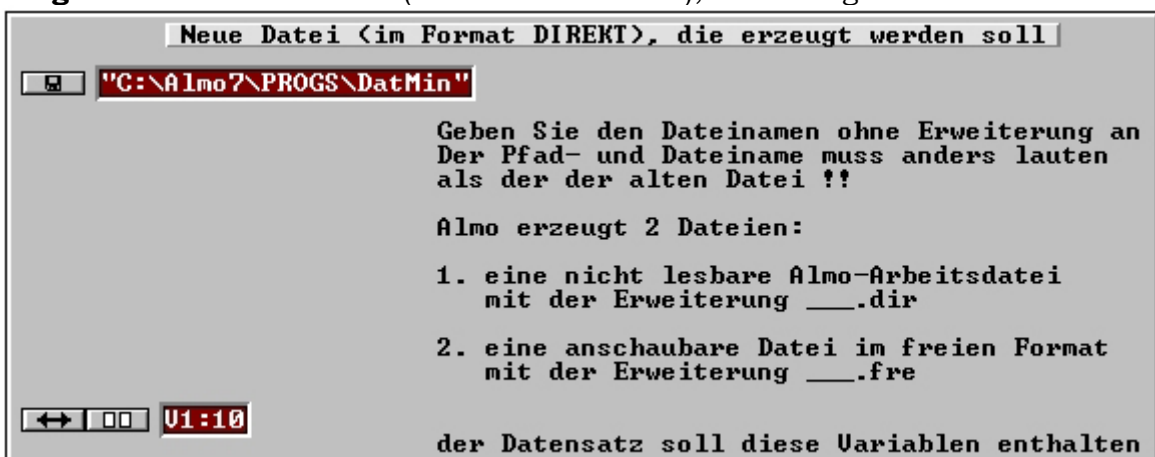
Jene Zeichenvariable, die in Eingabe-Box 6 nicht in eine Zahlenvariable umkodiert wurden, müssen hier angegeben werden. Sie werden in die neue Datei wieder als Zeichenvariable übernommen.

Dabei muss angegeben werden, wie lang, d.h. mit wie vielen Zeichen diese Variable maximal geschrieben werden sollen. Das könnte beispielsweise so ausschauen:



V4 soll mit maximal 8 Zeichen in die neue Datei geschrieben werden. Ist V4 länger, dann wird nach 8 Zeichen abgeschnitten. Ist V4 kürzer, dann wird mit Blanks aufgefüllt.

Eingabe-Box 8: Neue Datei (im Format DIREKT), die erzeugt werden soll



Diese Eingabe-Box entspricht der Eingabe-Box 6 aus Prog45md in Abschnitt P45.1 - mit der Ausnahme, dass keine Datei der Variablennamen erzeugt wird.

Eingabe-Box 8: Option: Inkludiere Programmteil zum platzsparenden Speichern



Wenn Sie diese Optionsbox öffnen, dann erscheint die Box, die wir schon bei Prog45md in Abschnitt P45.1 als Eingabe-Box 7 ausführlich erläutert haben.

Kapitel 2: Daten kennenlernen

P45.3 Schritt 2: Daten anschauen

Es ist sinnvoll, sich die Datenmatrix anzuschauen. Wenn der Benutzer in irgendeinem Almo-Fenster einen Dateinamen sieht, dann kann er – durch Doppelklick auf den Dateinamen (der zwischen Apostrophen stehen muß und die volle Pfadangabe enthalten muß) – die Datei in ein separates Fenster laden.

Die so geladene Datei ist allerdings nicht besonders schön angeordnet. Mit Prog45mg kann man sich nun die Datei in einer übersichtlicheren Form ausgeben lassen.

Prog45mg.Msk
Daten in übersichtlicher Form ausgeben

Was ist ein Kurzprogramm ? -->
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 zeige = Namensdatei in Output zeigen
leer = nicht

erzeuge zusätzliche Namensfelder

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

 diese Variablen sollen ausgegeben werden
Sie können Zahlen- oder Zeichenvariable sein

Option: Ein- und Ausschliessen von Untersuchungseinheiten

Option: Umkodierungen und Kein-Wert-Angaben

Option: Form der Ausgabe durch Benutzer beeinflussen

P45.3.1 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: die auszugebenden Variablen

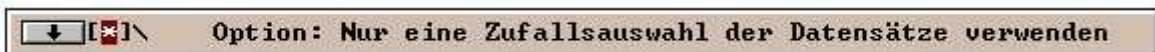


Tragen Sie hier die Variablen ein, die Sie ausgeben wollen. Sie können Variablennamen oder -nummern schreiben. Der Doppelpunkt zwischen „V1:10“ bedeutet „bis“.

Eingabe-Box 7: Ein- und Ausschliessen von Untersuchungseinheiten

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

Eingabe-Box 6: Nur eine Zufallsauswahl der Datensätze verwenden



Wird diese Optionsbox geöffnet, dann sieht man folgendes



Umfasst die Datei sehr viele Datensätze, dann kann man sich mit der Ausgabe einer Zufallsstichprobe von z. B. 10% begnügen.

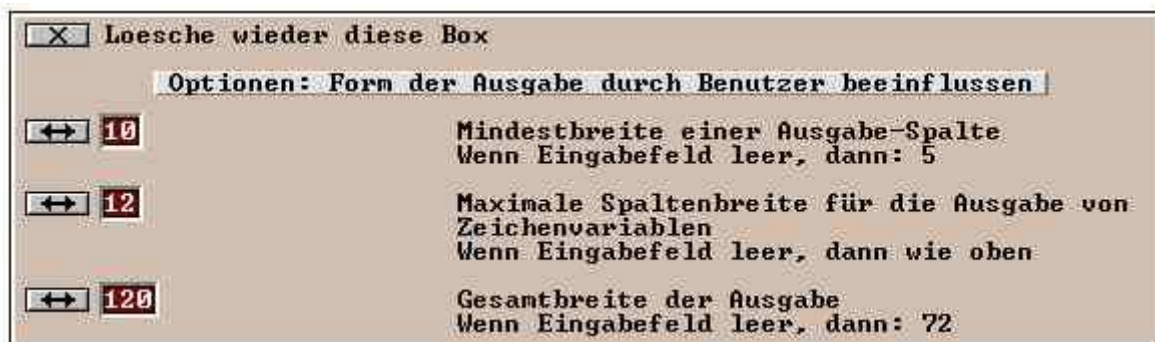
Eingabe-Box 7: Kein-Wert-Angabe und Umkodierungen

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

Eingabe-Box 8: Option: Form der Ausgabe



Wird die Optionsbox geöffnet, dann sieht man folgendes



Eingabefeld 1: Geben Sie an wie breit eine Ausgabespalte sein soll. In unserem Beispiel ist sie 10 Zeichen breit.

Eingabefeld 2: Für Zeichenvariable kann es notwendig sein, die Spaltenbreite zu verbreitern. In unserem Beispiel ist sie 12 Zeichen breit

Eingabefeld 3: Gesamtbreite der Ausgabe. Für das Ausdrucken ist ungefähr 96 die optimale Breite.

P45.3.2 Ausgabe

Almo liefert für unsere Testdaten folgende Ausgabe (gekürzt):

Zahl der eingelesenen Datensätze: 1000

Zahl der in die Analyse einbezogenen Datensätze: 1000

Nr.	V1 Wohnort	V2 Geschlecht	V3 Beruf	V4 Einkommen	V5 Kinderzahl	V6 Hausbesitz	V7 Rueckrate	V8 Laufzeit	V9 Produkt	V10 Rueckzahl
1	1	2	2	32384	5	1	5211	20	1	2
2	2	1	2	34770	1	1	4236	14	2	2
3	2	2	2	25653	1	1	5545	14	1	1
4	2	1	2	41643	0	1	4748	20	3	2
5	2	1	2	18641	1	1	1568	10	3	2
6	2	2	1	27226	2	1	3772	11	2	1
7	2	1	2	0	2	2	4901	21	3	2
8	1	1	2	28387	2	1	5411	13	3	2
9	2	2	2	49655	1	1	2409	15	1	2
10	2	2	2	50000	0	1	3124	21	2	2
.
.
.
.
.

P45.4 Schritt 3: Kennwerte der Variablen anschauen

Mit Prog45ml werden verschiedene Kennwerte der Analyse-Variablen ermittelt. Betrachten wir zuerst das Ergebnis aus diesem Programm.

Variable	Mittelwert	Stand.abwg.	Untergrenze	Diverse Werte	Kein-Wert Faele	
					abs	in %

Nominale Variable						

1 Wohnort	-	-	-	2.0000	0	0.00
2 Geschlecht	-	-	-	2.0000	0	0.00
3 Beruf	-	-	-	2.0000	0	0.00
6 Hausbesitz	-	-	-	2.0000	0	0.00
9 Produkt	-	-	-	3.0000	0	0.00
10 Rueckzahl	-	-	-	2.0000	0	0.00
Quantitative Variable						

4 Einkommen	24730.7920	13799.9172	-	887.0000	0	0.00
5 Kinderzahl	1.9700	1.4673	-	6.0000	0	0.00
7 Rueckrate	3459.7820	1362.3540	-	913.0000	0	0.00
8 Laufzeit	14.7710	5.5040	-	29.0000	0	0.00

Die Begriffe „nominale Variable“, „quantitative Variable“ und „ordinale Variable“ werden wir in P45.12 erläutern.

Mittelwerte und Standardabweichungen

Für die quantitativen Variablen wird uns der Mittelwert und die Standardabweichung mitgeteilt. Wir sehen z.B., daß die durchschnittliche Laufzeit eines Kredits 14.771 Monate beträgt und die Streuung um diesen Mittelwert herum 5.504 Monate ausmacht. Der Mittelwert ist sicherlich eine wichtige Information.

Die Standardabweichung zeigt uns, ob es Sinn macht, die Variable in die Analyse miteinzuschließen. Nehmen wir an, daß die Standardabweichung nur 0.5 Monate betragen würde. Das würde bedeuten, daß fast alle Personen eine Kredit-Laufzeit von 14.771 Monaten haben.

Es hätte keinen Sinn, diese Variable als erklärende Variable in eine Analyse einzuführen, in der erklärt werden soll, warum manche Personen ihre Kredite rechtzeitig zurückzahlen und andere nicht. Die Personen sind „zu ähnlich“, als daß diese Variable einen Erklärungswert besitzen könnte.

Allgemein gilt: Ist die Standardabweichung einer quantitativen Variable zu klein, dann sollte sie weder als erklärende noch als zu erklärende Variable in die Analyse aufgenommen werden.

Was „zu klein“ heißt, können wir nicht generell definieren. Der Benutzer muß das selbst entscheiden.

Kein-Wert-Fälle

In unseren Beispieldaten sind Kein-Wert-Fälle (fehlende Werte) nicht vorhanden. Allgemein gilt: Auch wenn Kein-Wert-Fälle vorhanden sind, ist eine Analyse möglich. Ist die Zahl der Kein-Wert-Fälle zu hoch (vielleicht über 50%), dann sollte die Variable nicht in die Analyse eingeschlossen werden.

Diverse Werte

Unsere 1000 Untersuchungspersonen haben 887 verschiedene Einkommen. Allgemein gilt: Hat eine Variable mehr als ca. 15-20 diverse Werte, dann sollte sie

für das nachfolgende Auszählungsprogramm Prog45m4 zu ca. 15-20 Ausprägungen zusammengefaßt werden. Wir werden das noch zeigen.

Wenn ordinale Variable genügend viele diverse Werte besitzen, dann kann man sie wie quantitative behandeln. Man kann dabei großzügig sein und schon ab ca. 6 diversen Werten diese Entscheidung treffen. Ordinale Variable machen bei der Analyse gewisse Schwierigkeiten. Wir werden darauf zurückkommen. Andererseits sind erfahrungsgemäß die Unterschiede in den Ergebnissen nicht sehr verschieden, wenn man dieselben Variablen einmal als ordinale und einmal als quantitative behandelt.

Eingabe in Prog45m1

Prog45ml.MSK

Einige wichtige Kennwerte der Variablen ermitteln

Das sind:

1. Werte-Unter- und Obergrenzen der Variablen
2. Zahl der diversen Werte, die die Variablen besitzen
3. Zahl der Kein-Wert-Fälle (fehlende Werte)
4. Mittelwerte der Variablen (auch Median, häufigster Wert)
5. Streuungen der 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

5

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

6

quantitative Variable

nominale Variable

ordinale Variable

7
8
9
10
11
12

<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> 1	Werte-Unter- und Obergrenzen oder Zahl der "diversen Werte" ermitteln 0 = für alle Variable werden die Unter- und Obergrenzen ermittelt 1 = für die nominalen und ordinalen Variablen werden die Unter- und Obergrenzen ermittelt für die quantitativen Variablen wird die Zahl der diversen Werte ermittelt 2 = für alle Variablen wird die Zahl der diversen Werte ermittelt
<input type="checkbox"/> ↓	Option: Ein- und Ausschliessen von Untersuchungseinheiten
<input type="checkbox"/> ↓	Option: Umkodierungen und Kein-Wert-Angaben
<input type="checkbox"/> ↓	Option: Untersuchungseinheiten gewichten
<input type="checkbox"/> ↓	Option: Ausgabe durch Benutzer beeinflussen
	Programmende

P45.4.1 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 5: Zu analysierende Variable

zu analysierende Variable Hilfe

quantitative Variable

↔ □□ Einkommen, Kinderzahl, Rueckrate, Laufzeit

nominale Variable

↔ □□ Wohnort, Geschlecht, Beruf, Hausbesitz, Produkt, Rueckzahl

ordinale Variable

↔ □□

Almo unterscheidet die 3 Meßniveaus: quantitativ, ordinal und nominal und rechnet dafür teilweise unterschiedliche Koeffizienten. Sie können aber auch ein und dieselbe Variable als quantitativ und als nominal und als ordinal angeben. Sie erhalten dann für diese Variable die Koeffizienten, die Almo für diese 3 Meßniveaus ermittelt.

Sie können die Analyse-Variablen „von Hand“ in die Eingabefelder schreiben oder Sie klicken auf den Knopf mit den 2 kleinen symbolischen Fenstern. Almo öffnet dann die Dialogbox „Variable für Analyse auswählen“. In dieser können Sie die Variable, die in die Eingabefelder eingeschrieben werden sollen per Mausklick selektieren. Siehe die ausführliche Beschreibung dieser Dialogbox in "Arbeiten mit Almo-Datenanalyse-System", Abschnitt P0.11.

Eingabe-Box 6: Werte-Unter- und Obergrenzen

Diverse Werte

↑ ↓ !

Werte-Unter- und Obergrenzen
oder
Zahl der "diversen Werte" ermitteln

0 = für alle Variable
werden die Unter- und Obergrenzen ermittelt

1 = für die nominalen und ordinalen Variablen
werden die Unter- und Obergrenzen ermittelt
für die quantitativen Variablen wird
die Zahl der diversen Werte ermittelt

2 = für alle Variablen wird
die Zahl der diversen Werte ermittelt

Wenn Sie 0 eingeben, dann berechnet Almo

- Unter- und Obergrenze (= Minimum und Maximum) aller Variablen (der nominalen, ordinalen, quantitativen)
- Mittelwerte für die quantitativen Variablen und Mediane für die ordinalen Variablen
- Standardabweichung für die quantitativen Variablen, halber Quartilsabstand für die ordinalen Variablen

Wenn Sie 1 eingeben, dann berechnet Almo

- die Unter- und Obergrenzen (= Minimum und Maximum) der nominalen und ordinalen Variablen – nicht jedoch für die quantitativen Variablen
- für die quantitativen Variablen wird die Zahl der diversen Werte ermittelt, die diese Variable annehmen
- für die quantitativen Variablen wird der Mittelwert und für die ordinalen der Median errechnet
- für die quantitativen Variablen wird die Standardabweichung ermittelt.

Wenn Sie 2 eingeben, dann berechnet Almo

- die Zahl der diversen Werte für alle Variable (nominale, ordinale, quantitative)
- Mittelwert für quantitative, Median für ordinale und Erwartungswert für nominale Variable. Zum Begriff des Erwartungswertes siehe Abschnitt P45.6.1, Eingabe-Box 7.
- Standardabweichung für quantitative und halber Quartilsabstand für ordinale Variable.

Eingabe-Box 7: Kein-Wert-Angabe und Umkodierungen

Siehe P0.5.

Eingabe-Box 8: Ein- und Ausschliessen von Untersuchungseinheiten

Siehe P0.7.

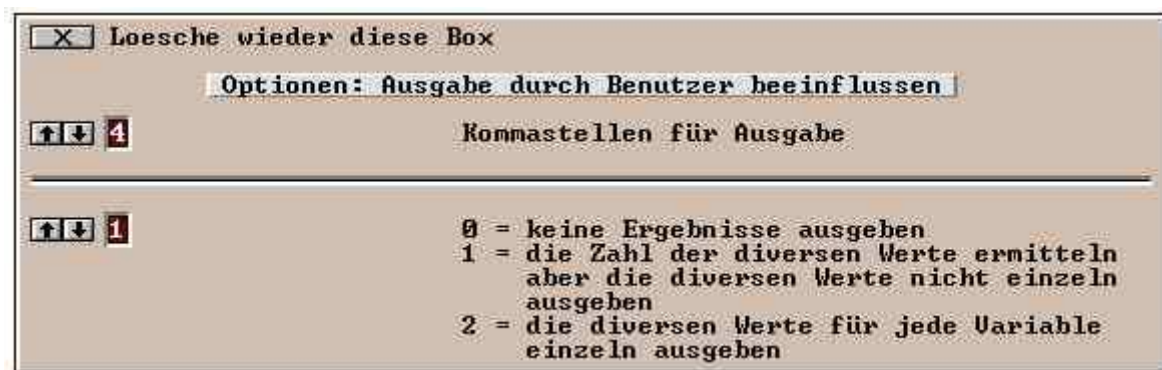
Eingabe-Box 9: Untersuchungseinheiten gewichten

Siehe P0.8.

Eingabe-Box 10: Option: Ausgabe durch Benutzer beeinflussen



Wird die Optionsbox geöffnet, dann sieht man folgendes



Eingabefeld 1: Wenn Sie hier z.B. 4 eingeben, dann werden die Ergebnisse mit 4 Kommastellen ausgegeben.

Eingabefeld 2: Art der Ergebnisausgabe

2 = alle Ergebnisse ausgeben. Falls für die quantitativen Variablen die Zahl der diversen Werte ermittelt wurde, dann gibt Almo alle aufgetretenen diversen Werte aus

1 = wie 2, aber die aufgetretenen diversen Werte nicht ausgeben

0 = Ergebnisse nicht ausgeben

P45.5 Schritt 4: Variable auszählen

Das nachfolgende Programm führt eine eindimensionale Auszählung von Variablen durch. Es entstehen Tabellen, wie etwa folgende.

Variable 9 gekauftes Produkt		
Wert	Faelle	%
1 Kleidung	204	20.40
2 Möbel	387	38.70
3 Technik	409	40.90
Summe	1000	100%

Prog45m4.Msk
Variable auszählen

Das Programm erstellt Häufigkeitsverteilungen folgender Art

Kauf von	Faelle	%
1 Kleidung	204	20.40
2 Möbel	387	38.70
3 Technik	409	40.90

Dabei besteht die Möglichkeit die Häufigkeitsverteilungen für eine "Gruppierungsvariable" wiederholt zu erzeugen. Beispiel: Die Gruppierungsvariable sei "Geschlecht" (männlich, weiblich) Dann können die Häufigkeitsverteilungen separat ausgezählt werden für Männer und Frauen - und dies in einem Programm-durchlauf

Grafik: Balkendiagramme

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
 Wohnort, Beruf, Einkommen, Hausbesitz, Produkt, Rueckzahl
- 7 Option: Gruppierungsvariable
- 8 Option: Ein- und Ausschliessen von Untersuchungseinheiten

9

Loesche wieder diese Box

Umkodierungen und Kein-Wert-Angaben

Umkodierungen
Kein_Wert-Angabe

Einkommen <0 Schritt 10000 his 60000=1>

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

10

Option: Untersuchungseinheiten gewichten

11

Grafik-Optionen

12

weitere Grafik-Optionen

13

Almo-Programmtext in Ergebnisliste zeigen oder nicht ?

zeige = Almo-Programmtext zeigen
Editfeld leer = nicht zeigen

P45.5.1 Erläuterungen zu den Eingabe-Boxen

Eingabe-Box 1 bis Eingabe-Box 5:

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

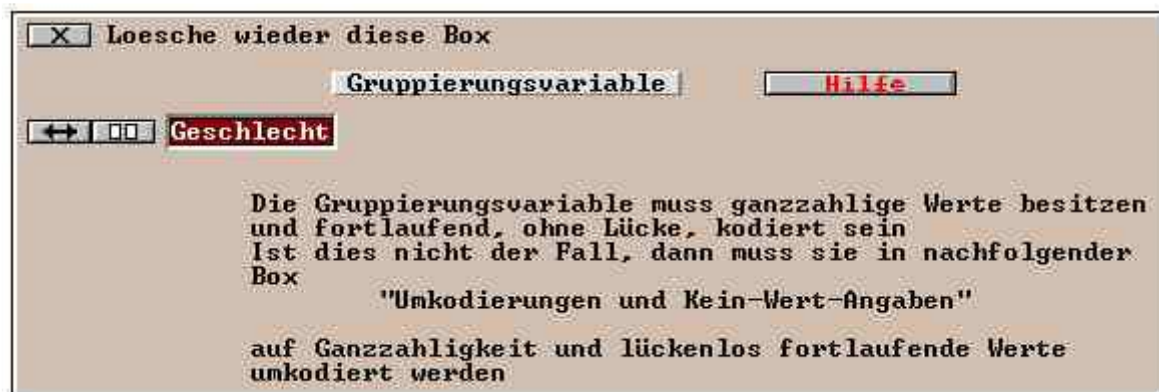
Eingabe-Box 6: Die auszuzählenden Variablen

Siehe P0.11.

Eingabe-Box 7: Gruppierungsvariable



Wird die Optionsbox geöffnet, dann sieht man folgendes



Es besteht die Möglichkeit die Häufigkeitsverteilungen für eine "Gruppierungsvariable" wiederholt zu erzeugen.

Beispiel: Die Gruppierungsvariable sei "Geschlecht" mit den Ausprägungen männlich und weiblich. Dann können die Häufigkeitsverteilungen für die auszuzählenden Variablen separat ausgezählt werden für Männer und Frauen - und dies in einem Programmdurchlauf.

Die Gruppierungsvariable muß ganzzahlige Werte besitzen und fortlaufend, ohne Lücke, kodiert sein. Ist dies nicht der Fall, dann muss sie im 3. Eingabefeld auf Ganzzahligkeit und lückenlos fortlaufende Werte umkodiert werden. Zum Umkodieren siehe P0.5.

Eingabe-Box 8: Ein- und Ausschliessen von Untersuchungseinheiten

Siehe P0.7.

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

Siehe P0.5.

Eingabe-Box 10: Kontrollieren, ob Umkodierung so erfolgt wie gewünscht

Siehe P0.6.

Eingabe-Box 11: Untersuchungseinheiten gewichten

Siehe P0.8.

Eingabe-Box 12: Grafik-Optionen

Siehe P0.10.

Eingabe-Box 13: Almo-Programmtext in Ergebnisliste zeigen oder nicht ?



Almo erzeugt aus dem Maskenprogramm einen in der Almo-Programmiersprache geschriebenen Programmtext. Diesen können Sie sich in die Ergebnisdatei ausgeben lassen - was empfehlenswert ist. Gelegentlich wird aber von Almo sehr viel Programmtext erzeugt, so daß die Ergebnisdatei sehr unübersichtlich wird. In diesem Fall verzichten Sie auf die Anzeige des Programmtexts.

P45.5.2 Ausgabe

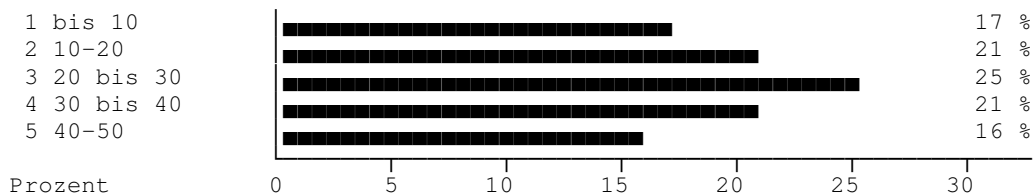
Wir zeigen nur einen Ausschnitt aus der Ergebnisliste

Eindimensionale Haeufigkeitsverteilungen

Zahl der eingelesenen Datensätze: 1000

Variable 4 Einkommen

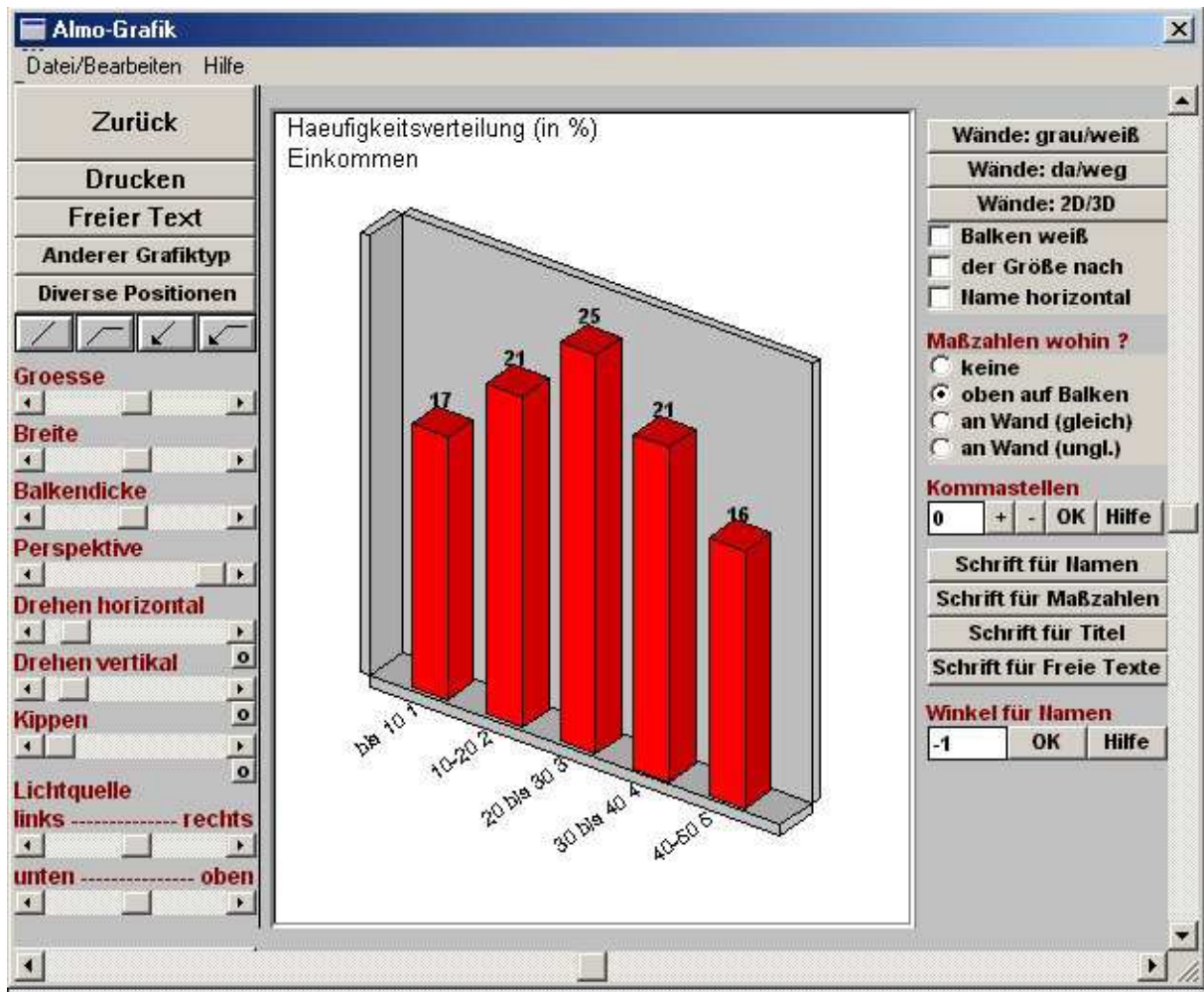
Wert	Faelle	%	% kumuliert
1 bis 10	166	16.60	16.60
2 10-20	209	20.90	37.50
3 20 bis 30	252	25.20	62.70
4 30 bis 40	211	21.10	83.80
5 40-50	162	16.20	100.00
Summe	1000	100%	



***** Erläuterung:

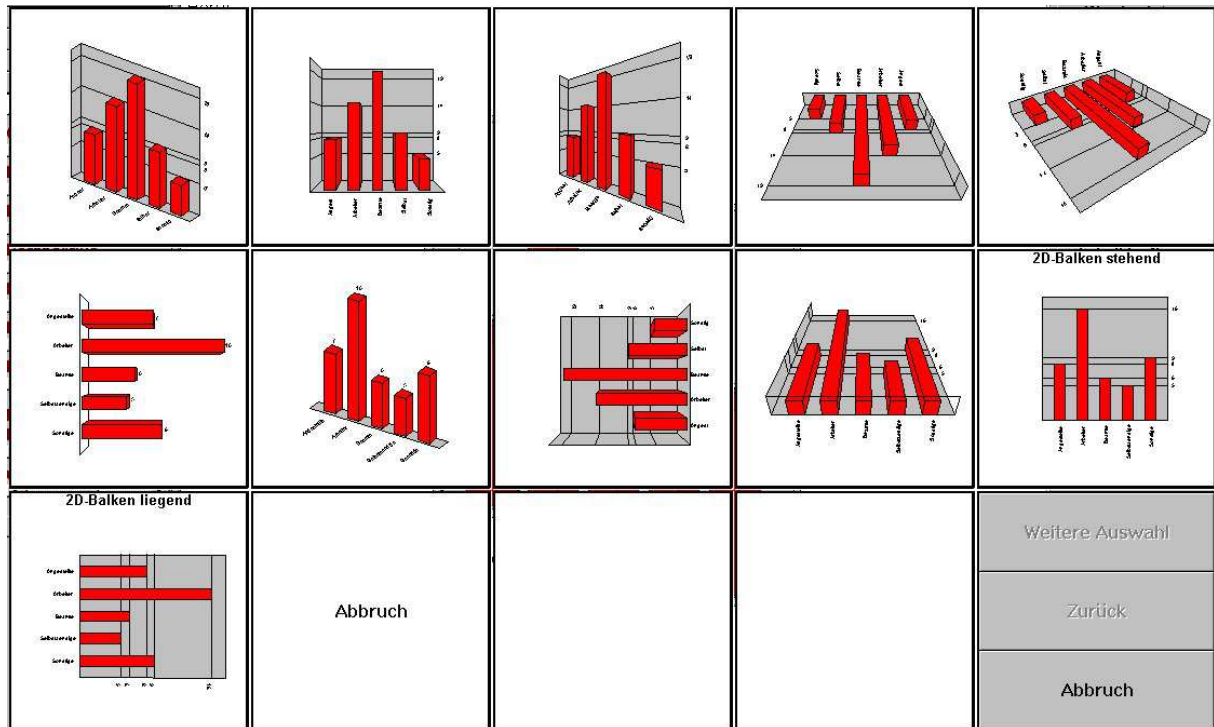
Almo liefert zuerst eine einfache Grafik im Textmodus, die dem Benutzer nur eine erste Übersicht ermöglichen soll und danach dann ein Balkendiagramm im hochauflösenden Grafikmodus.

Wenn Sie auf den Knopf "Grafik" klicken, der sich hinter dieser Tabelle befindet, dann lädt Almo den Grafikeditor und präsentiert Ihnen folgendes Balkendiagramm:



Wir wollen hier die vielfältigen Möglichkeiten des ALMO-Grafik-Editors nicht beschreiben. Eine ausführliche Darstellung ist im ALMO-Handbuch, Teil 1, "Bedienungsanleitung" enthalten. Zwei bedeutsame Gestaltungsmöglichkeiten wollen wir jedoch ausführlicher behandeln.

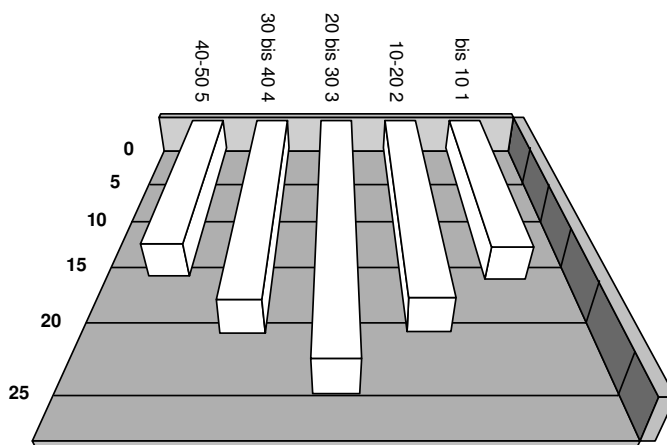
Klicken Sie auf den Knopf "Diverse Positionen". ALMO bietet Ihnen dann folgende Auswahl an:



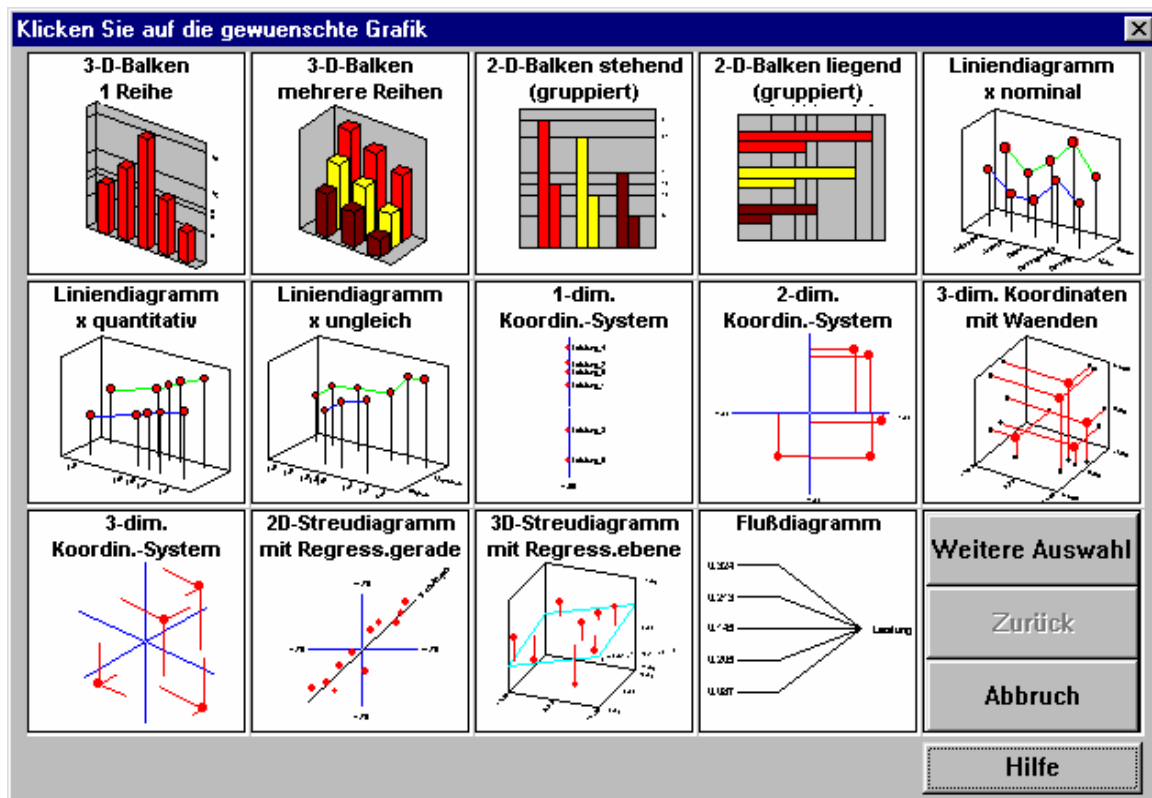
Klicken Sie auf die Position, die Sie wünschen. Also wandelt dann die aktuelle Grafik in diese Position. Also ändert dabei die Einstellung der Schieber 'Grösse', 'Perspektive', 'Drehen horizontal', 'Drehen vertikal' und 'Kippen'. Das bedeutet, daß Sie diese diversen Positionen auch selbst herstellen könnten, wenn Sie diese Schieber verstellen.

Wenn Sie beispielsweise auf die 4. Grafik in der ersten Reihe klicken, dann erzeugt ALMO folgende Grafik:

Haeufigkeitsverteilung (in %)
Einkommen

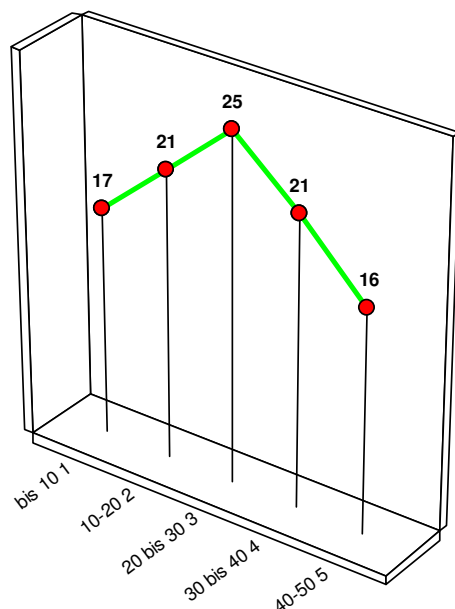


Sie können auch anstelle eines Balkendiagramms ein Liniendiagramm erzeugen. Klicken Sie auf den Knopf "Anderer Grafiktyp". Sie erhalten dann folgende Auswahl von Grafiktypen präsentiert.



Klicken Sie auf das 5. Bild in der ersten Reihe (Liniendiagramm, x nominal). ALMO wandelt dann das Balkendiagramm in folgendes Liniendiagramm:

Häufigkeitsverteilung (in %)
Einkommen



Sie können nun auch hier auf den Knopf "Diverse Positionen" klicken und das Liniendiagramm in einer anderen Position darstellen.

arithmetisches Mittel	2.994000	
Standardabweichung	1.316041	(im Nenner: n)
Standardabweichung	1.316700	(im Nenner: n-1)
Stand.fehler arithm. Mittels	0.041638	
Variabilitaetskoeffizient	43.977941	
(mit 100 multipliziert)		
geometrisches Mittel	2.651048	
harmonisches Mittel	2.274537	

******* Erläuterung:**

Diese Mittelwerte und Streuungsmaße sind nur dann verwenbar, wenn die Variable als quantitative betrachtet werden kann.

	mittlerer Wert bzw. Intervallmitte	linear interpoliert mit Wertintervall=1
1.Quartil	2.000000	1.901914
Median	3.000000	2.996032
3.Quartil	4.000000	4.082938
Quartilsabstand (mit 2 dividiert)	1.000000	1.090512

******* Erläuterung:**

Median und Quartile sind nur dann verwenbar, wenn die Variable als ordinale betrachtet werden kann.

Die Ergebnisse der Auszählung sind für uns in mehrfacher Weise bedeutsam:

1. Sind die Variablen ausgezählt, dann erkennen wir, ob sie stark disproportional verteilt sind. Wäre die Verteilung der gekauften Produkte etwa folgende,

	Fälle
Kleidung	950
Möbel	25
technische Geräte	25

dann wäre diese Variable für die Analyse kaum brauchbar.

2. Bei der Auszählung der Variablen „Beruf“ erhalten wir (so wollen wir annehmen) folgende Verteilung.

	Fälle
Hilfsarbeiter	500
Facharbeiter	350
Angestellte	75
Beamte	75

Hier ist es sinnvoll für die folgenden Analysen des Data-Mining-Prozesses Angestellte und Beamte zusammen zu fassen, da sie einzeln zu wenige Fälle umfassen und da sie substantiell auch gut in eine Kategorie passen.

3. Quantitative Variable, die sehr viele Ausprägungen haben, müssen für das Auszählen zusammengefaßt werden. Aus Prog45ml in Abschnitt P45.4 haben wir erfahren, daß das Einkommen 887 diverse Werte besitzt. Wir fassen deswegen für die Auszählung das Einkommen in 5 Gruppen zu je 10 000 Geldeinheiten zusammen. Etwa so:

Einkommen (-0 bis 10000 = 1;
	10001 bis 20000 = 2;
	20001 bis 30000 = 3;
	30001 bis 40000 = 4;
	40001 bis 50000 = 5)

Eleganter und kürzer ist folgende Umkodierungsanweisung.

Einkommen (0 Schritt 10000 bis 50000 = I)

I = heißt „Intervallkodierung“.

Selbstverständlich wäre es auch zulässig das Einkommen zu 10 oder sogar 20 Kategorien zusammenzufassen.

Für die nachfolgenden Programme des Data-Mining-Prozesses wird das Einkommen als quantitative Variable behandelt, also nicht zusammengefaßt.

Kapitel 3: Daten bereinigen

Kapitel 3 ist als eigenständiges Almo-Dokument Nr. 12 unter dem Titel "Daten-Imputation" vorhanden. Es ist im Vergleich zum folgenden "alten" Text ausführlicher und durch neue Abschnitte erweitert. Wir empfehlen anstelle des Kapitels 3 das neue Almo-Dokument 12 zu lesen.

Daten sind häufig unvollständig. Bei Befragungen beispielsweise wird von sehr vielen Befragten die Frage nach ihrem Einkommen nicht beantwortet. Die Variable „Einkommen“ besitzt also bei vielen Datensätzen – wie wir in Almo formulieren – den Wert „Kein_Wert“ (kurz: KW). Die Frage ist nun, wie kann man „Kein_Wert“ behandeln.

Oder noch grundsätzlicher gefragt: Kann man fehlende Werte überhaupt durch einen Schätzwert ersetzen?

Es ist z.B. möglich, daß der Wert für die Zielvariable aus Gründen fehlt, die nicht über irgend eine Schätzmethode erfaßt werden können. Beispielsweise könnte ein Befragter in einer Umfrage aus Scham sein Einkommen verschweigen, weil er im Verhältnis zu seinem Beruf und seinem Bildungsniveau viel zu viel oder viel zu wenig verdient.

Ob das eine oder das andere zutrifft oder ob überhaupt diese Ursache gegeben ist, ist kaum zu entscheiden.

Andererseits ist es oft plausibel anzunehmen, dass ein Befragter einem Fremden nicht die Höhe seines Einkommens mitteilen will – andere Befragte diese Hemmung aber weniger verspüren. Wenn wir dann den Beruf und das Bildungsniveau des Antwortverweigerers kennen, dann können wir versuchen sein Einkommen zu schätzen.

In Almo sind hier 2 prinzipiell verschiedene Möglichkeiten vorgesehen:

1. Mit drei speziellen Almo-Programmen werden die fehlenden Werte durch einen Schätzwert ersetzt. Die so "vervollständigten" Daten werden in eine neue Datei abgespeichert. In späteren Analysen mit den verschiedenen Almo-Programmen werden dann diese neuen "vollständigen" Dateien verwendet.
2. In allen Almo-Programmen sind spezielle Behandlungen für das Kein-Wert-Problem vorgesehen. Beim Einlesen und Verarbeiten der Daten bemerken die Almo-Programme, dass ein Wert fehlt und führen dann eine spezielle Operation aus, beispielsweise das sogenannte "paarweise Ausscheiden" oder, dass sie den Variablen-Mittelwert für den fehlenden Wert einsetzen oder dass sie einen Datensatz vollständig ausscheiden, wenn auch nur eine Analyse-Variable keinen Wert besitzt. Eine neue Datei wird nicht erstellt. Das „paarweise Ausscheiden“ werden wir in aller Ausführlichkeit in Abschnitt P45.12.4 behandeln.

Die Möglichkeit 2 wird in allen Almo-Programmen angeboten. Wir werden, wenn wir diese Programme erläutern, auch ausführlich die jeweils vorgesehenen "Kein-Wert-Behandlungen" beschreiben.

Für die 1. Vorgehensweise bietet Almo drei Programme an.

- a. Das Programm Prog45mo, das für fehlende Werte
 - bei quantitativen Variablen den Mittelwert einsetzt
 - bei ordinalen Variablen den Median
 - und bei nominalen Variablen den Erwartungswert

Das Programm erlaubt es auch, aus einem Wertebereich um den Mittelwert bzw. Median herum einen (normalverteilten) Zufallswert als Einsetzungswert zu wählen. Bei nominalen Variablen kann auch ein Zufallswert eingesetzt werden, dessen Wahrscheinlichkeit durch die Häufigkeitsverteilung bestimmt ist.

Wir werden das noch ausführlich darstellen.

- b. Das Programm Prog45mm, das mit Hilfe des Allgemeinen Linearen Modells (ALM) aus den vorhandenen Daten einen Prognosewert für den fehlenden Wert errechnet und diesen einsetzt und dann eine neue Datei erstellt. Das geht bei quantitativen und bei nominalen Variablen. Zum Prognosewert kann auch ein normalverteilter Zufallswert hinzu addiert werden.
- c. Ist die Zielvariable, deren fehlende Werte ersetzt werden sollen, nominal, dann bietet Almo als Alternative zum ALM das Programm Prog45mz an, das die fehlenden Werte über das Logitmodell zu schätzen versucht und ansonsten genau so vorgeht wie Prog45mm.

P45.6 Schritt 5a: Mittelwert für fehlende Werte einsetzen

Almo ermöglicht es, auf 4 verschiedene Weisen Einsetzungswerte für fehlende Werte einzusetzen. Wir wollen im folgenden von "Kein-Wert-Einsetzungswerten" oder kurz von "KW-Einsetzungswerten" sprechen. Diese werden

- a. bei quantitativen Variablen aus dem Mittelwert errechnet
- b. bei ordinalen Variablen aus dem Median (=dem mittleren Wert)
- c. bei nominalen Variablen aus dem Erwartungswert

Dabei gibt es jeweils verschiedene Varianten

Prog45mo.Msk
Einsetzen eines Ersatzwertes für Kein-Wert

1. Bei quantitativen Variablen wird der Mittelwert für "Kein-Wert" eingesetzt oder ein normalverteilter Zufallswert mit Mittelwert und Standardabweichung der Variablen
2. bei ordinalen Variablen der Median oder ein normalverteilter Zufallswert mit dem Median als Mittelwert und dem halben Quartilsabstand als Standardabweichung
3. bei nominalen Variablen der Erwartungswert oder der wahrscheinlichste Wert

siehe Handbuch "P45 Almo-Data-Mining", Abschnitt 45.6

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\DatMinKW.dir"

6

quantitative Zielvariable

Einkommen, Kinderzahl, Rueckrate, Laufzeit

ordinale Zielvariable

nominale Zielvariable

Wohnort, Geschlecht, Beruf, Hausbesitz, Produkt, Rueckzahl

Kein-Wert-Behandlung

↑ ↓ **5**

4 **Mittelwert-Einsetzung I** **Hilfe**
 Für Kein_Wert wird eingesetzt:
 a. bei quantitativen Variablen der Mittelwert
 b. bei ordinalen Variablen der zum Median
 nächst gelegene empirische Skalenwert
 c. bei nominalen Variablen die zum Erwartungswert
 nächst gelegene empirische Codeziffer

5 **Mittelwert-Einsetzung II** **Hilfe**
 Für Kein_Wert wird eingesetzt:
 a. bei quantitativen Variablen der zum Mittelwert
 nächst gelegene empirische Skalenwert
 b. bei ordinalen Variablen wie bei 4
 c. bei nominalen Variablen wie bei 4

6 **Mittelwert-Einsetzung III** **Hilfe**
 Für Kein_Wert wird eingesetzt:
 a. bei quantitativen Variablen
 ein normalverteilter Zufallswert mit Mittelwert
 und Standardabweichung der Variablen
 b. bei ordinalen Variablen der Median
 ein normalverteilter Zufallswert mit dem Median
 als Mittelwert und dem halben Quartilsabstand
 als Standardabweichung. Der zu diesem Zufallswert
 nächst gelegene empirische Skalenwert wird eingesetzt
 c. bei nominalen Variablen
 der wahrscheinlichste Ausprägungswert

7 **Mittelwert-Einsetzung IV** **Hilfe**
 a. bei quantitativen Variablen
 ein normalverteilter Zufallswert mit Mittelwert
 und Standardabweichung der Variablen. Der nächst
 gelegene empirische Skalenwert wird eingesetzt
 b. bei ordinalen Variablen wie bei 6
 c. bei nominalen Variablen wie bei 6

7

Startwert für Zufallsgenerator
 (für Verfahren 6 und 7)

↔ **123457** **Hilfe**

8

Kein-Wert-Angabe und Umkodierungen
 Werden Variable umkodiert, dann gehen sie
 in der umkodierten Form in die neue Datei ein

Kein-Wert-Angabe **Hilfe**
 Umkodierungen **Hilfe**

↔ ↓
 ↔ ↓
 ↔ ↓

[...] erzeuge zusätzliche Felder für Umkodierungen / Kein_Wert-Angaben

9

Neue Almo-Dateien

Für Zielvariable werden Ersatzwerte für fehlende Werte eingesetzt

📁 **"C:\Almo7\PROGS\DatMinNeu"**

Geben Sie den Dateinamen ohne Erweiterung
 an. Almo erzeugt 2 Dateien:

1. eine nicht lesbare Almo-Arbeitsdatei
 mit der Erweiterung __.dir
2. eine anschaulbare Datei im freien Format
 mit der Erweiterung __.fre

↔ **U1:10**

der Datensatz soll diese Variablen enthalten

10

P45.6.1 Erläuterungen zu den Eingabe-Boxen

Eingabe-Box 1 bis Eingabe-Box 5:

Siehe PO.1 bis PO.4.

Eingabe-Box 6: Variablen, für die ein Ersatzwert für fehlende Werte eingesetzt werden soll.

Variablen, für die ein Ersatzwert für fehlende Werte eingesetzt werden soll

quantitative Zielvariable

↔ □□ Einkommen, Kinderzahl, Rueckrate, Laufzeit

ordinale Zielvariable Hilfe

↔ □□ ■

nominale Zielvariable Hilfe

↔ □□ Wohnort, Geschlecht, Beruf, Hausbesitz, Produkt, Rueckzahl

Eingabe-Box 7: Kein-Wert-Behandlung

Kein-Wert-Behandlung

↑ ↓ 5

4 **Mittelwert-Einsetzung I** Hilfe

Für Kein_Wert wird eingesetzt:

- a. bei quantitativen Variablen der Mittelwert
- b. bei ordinalen Variablen der zum Median nächst gelegene empirische Skalenwert
- c. bei nominalen Variablen die zum Erwartungswert nächst gelegene empirische Codeziffer

5 **Mittelwert-Einsetzung II** Hilfe

Für Kein_Wert wird eingesetzt:

- a. bei quantitativen Variablen der zum Mittelwert nächst gelegene empirische Skalenwert
- b. bei ordinalen Variablen wie bei 4
- c. bei nominalen Variablen wie bei 4

6 **Mittelwert-Einsetzung III** Hilfe

Für Kein_Wert wird eingesetzt:

- a. bei quantitativen Variablen ein normalverteilter Zufallswert mit Mittelwert und Standardabweichung der Variablen
- b. bei ordinalen Variablen der Median ein normalverteilter Zufallswert mit dem Median als Mittelwert und dem halben Quartilsabstand als Standardabweichung. Der zu diesem Zufallswert nächst gelegene empirische Skalenwert wird eingesetzt
- c. bei nominalen Variablen der wahrscheinlichste Ausprägungswert

7 **Mittelwert-Einsetzung IV** Hilfe

- a. bei quantitativen Variablen ein normalverteilter Zufallswert mit Mittelwert und Standardabweichung der Variablen. Der nächst gelegene empirische Skalenwert wird eingesetzt
- b. bei ordinalen Variablen wie bei 6
- c. bei nominalen Variablen wie bei 6

Kein-Wert-Behandlung 4: Mittelwert-Einsetzung I

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

Für Kein_Wert wird eingesetzt:

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

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

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

Der Erwartungswert ist 2.1

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

Kein-Wert-Behandlung 5: Mittelwert-Einsetzung II

Für Kein_Wert wird eingesetzt:

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

Kein-Wert-Behandlung 6: Mittelwert-Einsetzung III

Für Kein_Wert wird eingesetzt:

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

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

Liegt dieser zwischen zwei empirisch vorkommenden Werten, dann wird der zum Median nächst gelegene empirische Wert verwendet.

Ist die Variable mit gleicher Schrittweite kodiert, dann wird ein Wert X errechnet, der sich ergibt aus Median +/- einem normalverteilten Zufallswert mit Mittelwert=0 und Standardabweichung in der Größe des halben Quartilsabstands der Variablen. Der zu X nächst gelegene empirische Skalenwert wird dann eingesetzt.

Bei quantitativen und bei ordinalen Variablen wird also eine normalverteilte Zufallszahl mit Mittelwert=0 generiert.

Als Standardabweichung wird bei quantitativen Variablen die der jeweiligen Variablen verwendet. Bei ordinalen Variablen wird der halbe Quartilsabstand verwendet.

Betrachten wir ein Beispiel: Die quantitative Variable sei das Lebensalter. Also errechnet für sie einen Mittelwert von 40 und eine Standardabweichung von 20. Dann wird eine normalverteilte Zufallszahl mit Mittelwert=0 und Standardabweichung=20 erzeugt. Nehmen wir an es entsteht der Zufallswert -15.25. Für den fehlenden Wert wird dann eingesetzt $X = 40 - 15.25 = 24.75$.

Bei einer ordinalen Variablen wird entsprechend verfahren. Als Standardabweichung für die Generierung der Zufallszahl wird der halbe Quartilsabstand verwendet. Der ermittelte X-Wert wird bei der ordinalen Variablen aber noch nicht als KW-Einsetzungswert verwendet. Es wird nach dem empirisch vorkommenden Wert gesucht, der am dichtesten bei X liegt. Dieser wird als KW-Einsetzungswert verwendet. So wird verhindert, daß KW-Einsetzungswerte entstehen, die empirisch nicht vorkommen.

- c. Bei nominalen Variablen wird der wahrscheinlichste Ausprägungswert eingesetzt. Die Vorgehensweise soll an einem Beispiel gezeigt werden. Die nominale Variable sei der Beruf mit den 3 Ausprägungen Arbeiter, Angestellte, Sonstige. Dabei wurden folgende Häufigkeiten ermittelt.

	Code	Häufigkeit	in %	in % kumuliert
Arbeiter	1	250	25	25
Angestellte	2	400	40	65
Sonstige	3	350	35	100

Dann wird eine gleichverteilte Zufallszahl zwischen 0 und 100 erzeugt.

Liegt sie zwischen

0 und 25, dann wird für den fehlenden Wert 1 eingesetzt	
25	65
65	100

Kein-Wert-Behandlung 7: Mittelwert-Einsetzung IV

Für Kein_Wert wird eingesetzt:

- a. Bei quantitativen Variablen:

Es wird zunächst ein Wert X errechnet, der sich ergibt aus dem Mittelwert +/- einem normalverteilten Zufallswert mit Mittelwert=0 und der Standardabweichung der Variablen. Dann wird der zu X nächst gelegene empirische Skalenwert für Kein_Wert eingesetzt. So wird verhindert, dass KW-Einsetzungswerte entstehen, die empirisch nicht vorkommen.

- b. bei ordinalen Variablen wie bei Kein-Wert-Behandlung 6
- c. Bei nominalen Variablen wie bei Kein-Wert-Behandlung 6

Kein-Wert-Behandlung 4 und 5 unterscheiden sich von 6 und 7 dadurch, dass bei 6 und 7 eine Zufallsvariation dem Mittelwert bzw. Median bzw. Erwartungswert hinzugefügt wird.

Die Kein-Wert-Behandlung 4 unterscheiden sich von 5 nur dadurch dass für die quantitativen Variablen ein Mal der Mittelwert und das andere Mal der zum Mittelwert nächste empirisch vorkommende Wert als KW-Einsetzungswert verwendet wird.

Warum Zufallswert hinzufügen?

Es muß noch folgende Frage beantwortet werden: Warum wird der Mittelwert bzw. der Median bei Kein-Wert-Behandlung 6 und 7 durch einen Zufallswert überlagert?

Wird als KW-Einsetzungswert nur der Mittelwert (bzw. der Median) verwendet, dann wird die Varianz der Variablen verringert, weil für Kein-Wert immer derselbe Wert eingesetzt wird.

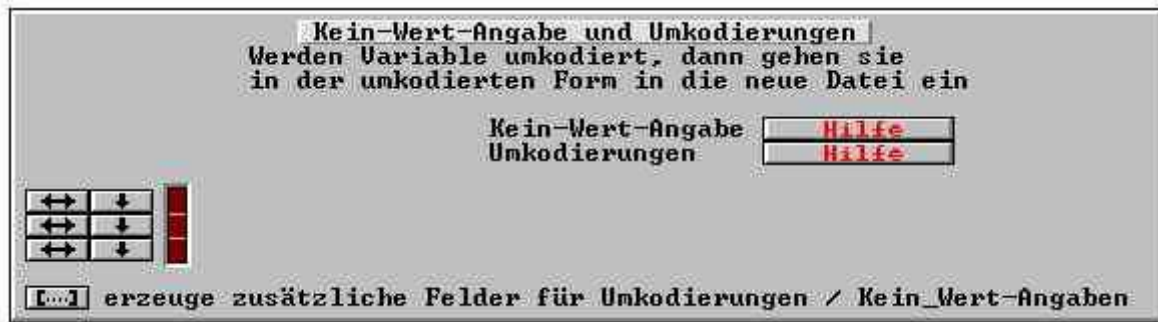
Werden mit den so erzeugten „vollständigen“ Daten beispielsweise Korrelationen errechnet, dann werden die Signifikanzen dieser Korrelationen überschätzt. Siehe dazu etwa R. J. A. Little/D. B. Rubin (1990, S. 381).

Die Überlagerung durch einen normalverteilten Zufallswert mit der Standardabweichung der Variablen bezweckt also, dass die Varianz der Variablen (fast) unverändert bleibt. Gleiches gilt auch für nominale Variable. Der Erwartungswert der Variablen ist immer derselbe. Dadurch wird die Varianz verringert. Durch den "wahrscheinlichsten" Wert bleibt die Streuung (fast) unverändert.

Eingabe-Box 8: Startwert für Zufallsgenerator

Die Zufallswerte, die in den oben beschriebenen Kein-Wert-Behandlungen 6 und 7 benötigt werden, erzeugt Almo mit einem "Zufallsgenerator". Wenn die Startzahl nicht verändert wird, dann werden bei einem 2. und jedem weiteren Lauf des Programms Prog45mo immer dieselben Zufallszahlen und damit dieselben KW-Einsetzungswerte erzeugt. Ist dies jedoch nicht erwünscht, dann muß der Benutzer die Startzahl ändern. Verwenden Sie eine 6-stellige ungerade Zahl.

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



Im "Arbeiten mit Almo-Datenanalyse-System", Abschnitt P0.5 ist ausführlich beschrieben, wie Kein-Wert-Angaben und wie Umkodierungen zu schreiben sind. Hier ist nun noch folgendes hinzuzufügen:

Almo muß selbstverständlich wissen, an welchen Codeziffern es den Kein-Wert-Fall erkennen kann. Hier gibt es 2 Vorgehensweisen:

- a. Der Benutzer hat schon eine Almo-Arbeitsdatei (im Format DIREKT) erstellt. Dabei hat er Almo mitgeteilt, welche Codeziffern den Kein-Wert-Fall bezeichnen. Der Benutzer hat beispielsweise die Ziffer 0 und bei einigen Variablen die Zahl -1 als Kein-Wert-Code verwendet. Almo hat dann die Almo-Arbeitsdatei erzeugt und dabei die Kein-Wert-Codeziffer (beispielsweise die 0 und die -1) durch einen Almo-internen Kein-Wert-Code (die riesige negative Zahl $10^{\text{hoch}} -38$) ersetzt. In diesem Fall ist jetzt eine Kein-Wert-Angabe nicht mehr notwendig.

Mit dieser Vorgehensweise haben wir die Almo-Arbeitsdatei in den Programmen Prog45md und Prog45mh in den Abschnitten P45.1 und P45.2 erzeugt.

- b. Der Benutzer hat eine Kein-Wert-Deklaration noch nicht vorgenommen. In der Arbeitsdatei stehen also noch die ursprünglichen Codes (z.B. 0 oder -1 für Kein-Wert). In diesem Fall muß der Benutzer jetzt eine Kein-Wert-Angabe vornehmen - beispielsweise so:

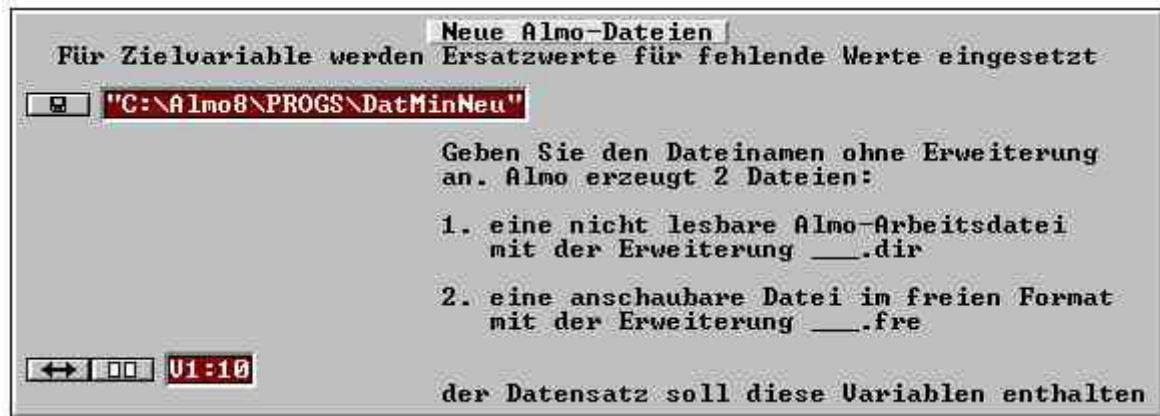
```
Beruf, Wohnort ( 0 = Kein_Wert )
Einkommen      ( -1 = Kein_Wert )
```

In der Eingabe-Box "Kein-Wert-Angabe und Umkodierungen" können auch Variable umkodiert werden. Sie gehen dann in der umkodierten Form in die neue Datei ein. Wenn der Benutzer beispielsweise die Variable Beruf dichotomisiert und schreibt

```
Beruf (0=Kein_Wert; 0 bis 5 = 1; 5 bis 10 = 2)
```

dann wird Beruf in dieser dichotomisierten Form in die neue Datei geschrieben.

Eingabe-Box 10: Neue Almo-Dateien



Eingabefeld 1: Geben Sie den Dateinamen für die Datei an, in die Sie Ihre Daten speichern wollen. Almo schreibt dabei für die als "Zielvariable" angegebenen Variablen (sofern diese Kein-Wert waren) die errechneten Einsetzungswerte. Die anderen Variablen werden mit ihren ursprünglichen Werten übernommen.

Geben Sie dabei den Dateinamen ohne Erweiterung an. Almo erzeugt dann 2 Dateien:

1. eine nicht lesbare Almo-Arbeitsdatei mit der Erweiterung __.dir
Im Beispiel: "C:\Almo6\PROGS\DatMinNeu.dir"
2. eine anschaulbare Datei im freien Format mit der Erweiterung __.fre
Im Beispiel: "C:\Almo6\PROGS\DatMinNeu.fre"

In der "anschaulbaren" Datei können Sie sich nochmals die von Almo errechneten Kein-Wert-Einsetzungswerte anschauen.

Eingabefeld 2: Geben Sie die Nummern der Variablen an, die in die neue Datei übernommen werden sollen. In der Regel wird man alle Variable aus der Ursprungsdatei übernehmen. Verwenden Sie dabei die Schreibweise "V1:...". Der Doppelpunkt heißt „bis“.

P45.7 Prognosewerte für fehlende Werte einsetzen

P45.7.1 Schritt 5b: Prognosewerte für fehlende Werte durch das Allgemeine Lineare Modell ermitteln

Betrachten wir folgende Situation: In einer Umfrage erfahren wir von 70% der Befragten die Höhe ihres Einkommens. 30% teilen ihr Einkommen nicht mit.

In der Umfrage wurde auch der Beruf und das Bildungsniveau erfragt. Diese beiden Variablen stehen mit dem Einkommen in einem engen Zusammenhang. Sie korrelieren mit dem Einkommen.

Folgende Vorgehensweise bietet sich nun an:

Wir rechnen ein Allgemeines Lineares Modell (kurz: ALM) mit Einkommen als Zielvariable und Beruf und Bildungsniveau als ursächliche Variable. Dabei erhalten wir Regressionskoeffizienten bzw. Effekte für die ursächlichen Variablen.

Almo ermittelt dabei folgende Gleichung

$$E = \beta \cdot B + a_i + \text{const}$$

E = Einkommen

B = Bildungsniveau (quantitativ)

β = Regressionskoeffizient

a_i = Effekt des Berufs i

const = Konstante

Diese Gleichung verwenden wir um das Einkommen jener Befragten, von denen wir dieses nicht kennen, zu „prognostizieren“. Voraussetzung ist natürlich, daß wir ihren Beruf und ihre Schulbildung kennen. Den so errechneten „Prognosewert“ setzen wir für diese Befragten in die Variable des Einkommens ein.

Diese Vorgehensweise ist auch möglich, wenn die Zielvariable nominal (dichotom oder polytom) ist. Betrachten wir dazu ein Beispiel: Die Zielvariable sei „Kauf auf Kredit: ja, nein“ mit den Codeziffern 1 und 2.

Unser Programm errechnet als vorläufigen Prognosewert, die Wahrscheinlichkeit für „Kreditkauf: ja“ und für „Kreditkauf: nein“. Ist beispielsweise die Wahrscheinlichkeit für „Kreditkauf: ja“ größer, dann wird als entgültiger Prognosewert „1“ für den fehlenden Wert eingesetzt (und im umgekehrten Fall die „2“).

Folgende generelle Anmerkungen zur Schätzung eines Ersatzwertes für fehlende Werte mittels des ALM sind notwendig:

Gegen die Verwendung des ALM, wenn die abhängige Variable eine nominale ist, gibt es Einwände, im wesentlichen folgende zwei:

- a. Es besteht modellbedingte Varianzheterogenität mit der Folge, daß die Schätzer nicht effizient sind.
- b. Die prognostizierten Wahrscheinlichkeiten können außerhalb des Bereichs 0-1 liegen.

Als Alternative wird das gewichtete ALM oder, noch besser, die Logit-Analyse, empfohlen. Wir werden in Abschnitt P45.15.1.0 auf diese Einwände ausführlich eingehen. Hier ist jedoch folgendes zu sagen:

Die Einwände sind u. E. nicht schwerwiegend, wenn es darum geht, innerhalb einer Stichprobe fehlende Werte zu prognostizieren – und nicht darum von einer Stichprobe auf eine Grundgesamtheit zu schließen. Wir werden trotzdem mit Prog45mz (in Abschnitt P45.7.2) ein Programm anbieten, das die Logitanalyse verwendet.

Wenn die Zielvariable, für die ein Wert fehlt, nur schwach durch die ursächlichen Variablen determiniert wird, dann ist der Prognosewert selbstverständlich nicht gut. Ob die Determination schwach oder stark ist, hängt natürlich auch davon ab, ob die relevanten ursächlichen Variablen zur Verfügung stehen und in das Modell aufgenommen wurden. Je schlechter die Determination, umso mehr nähert sich der Prognosewert dem Mittelwert der Zielvariablen an, umso mehr nähert sich unsere Vorgehensweise der Mittelwerteinsetzung in P45.6 an.

Almo bietet das Maskenprogramm Prog45mm an, bei dem der Benutzer nur einige wenige Eingaben vornehmen muß. Dieses Programm bietet auch die Möglichkeit an, den Kein-Wert-Einsetzungswert aus einem Zufallsintervall um den Prognosewert herum zu wählen.

P45.7.1.1 Wie Prog45mm rechnet

Die Einsetzung eines Ersatzwertes für Kein-Wert geschieht in folgenden 3 Schritten:

Schritt 1:

Zuerst wird mit den vorhandenen Daten ein Allgemeines Lineares Modell gerechnet. Dabei entstehen u.a. Regressionskoeffizienten und Effekte.

Schritt 2:

Dann werden mit Hilfe der dabei errechneten Koeffizienten, Prognosewerte für die Zielvariable ermittelt.

Schritt 3:

Aus diesen Prognosewerte werden dann die Einsetzungswerte für die fehlenden Werte gebildet. Die Daten werden dann in eine neue Datei abgespeichert.

Betrachten wir diese Abfolge von Rechenschritten etwas genauer:

Schritt 1: Ermittlung der Regressionskoeffizienten und Effekte

Das Programm rechnet ein Allgemeines Lineares Modell (ALM) für die Zielvariablen und die ursächlichen Variablen, die der Benutzer angibt.

Der Benutzer kann die ursächlichen Variablen umkodieren, nicht jedoch die Zielvariablen.

Almo verwendet das Verfahren der "weighted squares of means". Ist nur 1 ursächliche nominale Variable vorhanden, dann schaltet Almo selbständig auf "fitting constants" um.

Es wird die Quadratsummen-Matrix gebildet.

Aus der Quadratsummen-Matrix werden dann die Regressionskoeffizienten für die ursächlichen quantitativen/ordinalen Variablen und die Effekte für die ursächlichen nominalen Variablen (und ihre Interaktionen) ermittelt. Die Effekte sind die (transformierten) Regressionskoeffizienten der Dummies der nominalen Variablen. Siehe dazu Almo-Handbuch zu P20, Allgemeines Lineares Modell, Abschnitt P20.3.

Folgendes Problem kann sich stellen:

Die ursächlichen Variablen können auch fehlende Werte aufweisen. Die Katze beißt sich hier in den Schwanz. Um Ersatzwerte für fehlende Werte der Zielvariable zu schätzen, benötigen wir ursächliche Variable, die ihrerseits aber auch wieder fehlende Werte aufweisen können.

Es ist naheliegend, dass man in dieser Situation das ALM nur auf jene Datensätze anwendet, die in allen Analysevariablen (Zielvariable und ursächlichen Variablen) einen Wert besitzen.

Almo bietet deswegen die Methode des "vollständigen Ausscheidens" an. Dabei kann allerdings der Datenverlust sehr groß sein. Almo bietet deswegen zusätzlich die Methode des "paarweisen Ausscheidens" und des Einsetzens von Ersatzwerten an (siehe hierzu P45.7.1.3, Erläuterung zu Eingabe-Box 9).

Schritt 2: Prognosewerte für Zielvariable ermitteln

Mit Hilfe der Regressionskoeffizienten und der Effekte werden nun für jene Untersuchungseinheiten, die in der Zielvariablen keinen Wert besitzen, "Prognosewerte" ermittelt. Dabei stellt sich nun wieder das Problem, wie verfahren werden soll, wenn ursächliche Variable keinen Wert besitzen. Almo setzt in diesem Falle bei quantitativen ursächlichen Variablen den Mittelwert ein und bei nominalen ursächlichen Variablen den Erwartungswert. Zu dessen Berechnung siehe bei Prog45mo die Erläuterungen zu der Eingabe-Box "Kein-Wert-Behandlung" Abschnitt "Kein-Wert-Behandlung 4: Mittelwert-Einsetzung I"

Schritt 3: Aus den Prognosewerten die Kein-Wert-Einsetzungswerte bilden

Der errechnete Prognosewert kann, muß aber noch nicht, als KW-Einsetzungswert verwendet werden.

Beispielsweise besteht die Möglichkeit, den Kein-Wert-Einsetzungswert aus einem Intervall um den Prognosewert herum zufällig auszuwählen. Almo bietet mehrere Möglichkeiten an (in der Eingabe-Box "Transformation der Prognosewerte zu Einsetzungswerten"). Schließlich werden alle Datensätze in eine neue Datei gespeichert.

P45.7.1.2 Eingabe in das Maskenprogramm Prog45mm

Prog45mm.Msk

Einsetzungswerte für fehlende Werte ermitteln
mit Hilfe des Allgemeinen Linearen Modells (ALM)

Die Einsetzung erfolgt in folgenden 3 Schritten:

Schritt 1:
Zuerst wird mit den vorhandenen Daten ein ALM gerechnet.
Dabei entstehen u.a. Regressionskoeffizienten und Effekte

Schritt 2:
Dann werden mit Hilfe der dabei errechneten Koeffizienten
Prognosewerte für jene Zielvariable ermittelt, die
keinen Wert besitzen

Schritt 3:
Aus diesen Prognosewerten werden dann die Einsetzungswerte
für die fehlenden Werte gebildet.
Die Daten werden dann in eine neue Datei abgespeichert

siehe Handbuch "P45 Almo-Data-Mining", Abschnitt 45.7

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

5

6
für die Prognosewerte für fehlende Werte eingesetzt werden sollen

BEACHTEN: Erlaubt sind:

1. Beliebig viele quantitative und/oder
dichotome Variable
oder (exklusiv)
2. Eine nominale Variable mit beliebig
vielen Ausprägungen

 (mehrere) quantitative/dichotome Zielvariable

 (nur eine) nominale Zielvariable

7

Ursächliche Variable für die Zielvariable

ursächliche nominale Variable

Hausbesitz, Beruf

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

ursächliche quantitative Variable

Rueckrate, Laufzeit

ursächliche ordinale Variable

8

Kein-Wert-Angabe: Für ursächliche Variable und Zielvariable

Umkodierungen: Nur für ursächliche Variable
Zielvariable darf nicht umkodiert werden !!!
Umkodierungen sind temporär

Kein-Wert-Angabe
Umkodierungen

erzeuge zusätzliche Felder für Umkodierungen / Kein_Wert-Angaben

9

Kein-Wert-Behandlung der ursächlichen Variablen

bei Berechnung der Regressionskoeffizienten und Effekte
(möglich: 1 - 7; empfohlen: 3 =Vollständiges Ausscheiden)

bei Berechnung der Prognosewerte
(möglich: 4 - 7; empfohlen: 4 =Mittelwert-Einsetzung)

10

Transformation der Prognosewerte zu Kein-Wert-Einsetzungswerten

möglich 4 - 7; empfohlen: 5 oder 7

11

Startwert für Zufallsgenerator
(für Verfahren 6 und 7)

12

Neue Almo-Dateien

Für Zielvariable werden Einsetzungswerte für fehlende Werte eingesetzt

Geben Sie den Dateinamen ohne Erweiterung an. Almo erzeugt 2 Dateien:

1. eine nicht lesbare Almo-Arbeitsdatei mit der Erweiterung `__.dir`
2. eine anschauliche Datei im freien Format mit der Erweiterung `__.fre`

der Datensatz soll diese Variablen enthalten

13

Ausgabe der Ergebnisse aus ALM

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

1= Basisstatistiken ausgeben
0= nicht

Almo = Almo-Grafik ausgeben
0 = keine Grafik

P45.7.1.3 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: Zielvariable

Zielvariable
für die Prognosewerte für fehlende Werte eingesetzt werden sollen

BEACHTEN: Erlaubt sind:

1. Beliebig viele quantitative und/oder dichotome Variable oder (exklusiv)
2. Eine nominale Variable mit beliebig vielen Ausprägungen

<mehrere> quantitative/dichotome Zielvariable

<nur eine> nominale Zielvariable

Mit "Zielvariable" bezeichnen wir jene Variable, deren fehlende Werte durch Prognosewerte aus dem ALM ersetzt werden sollen.

Eingabefeld 1:

Es können beliebig viele quantitative oder dichotome Variable eingegeben werden.

Empfehlung: Man sollte nur 1 Zielvariable einsetzen. Werden mehrere eingesetzt, dann wird beim Kalkül des ALM ein zusätzlicher Datenverlust auftreten. Wir beschreiben dieses Problem, wenn wir die Eingabe-Box 9 „Kein-Wert-Behandlung der ursächlichen Variablen“ erläutern. Siehe unsere dortige Empfehlung.

Eingabefeld 2:

Es kann nur 1 nominale Variable (mit beliebig vielen Ausprägungen) angegeben werden.

BEACHTEN: Es darf nur 1 Eingabefeld benutzt werden. D.h. erlaubt sind

1. beliebig viele quantitative und dichotome Variable oder (exklusiv)
2. eine nominale Variable

Ordinale Variable können nicht als Zielvariable angegeben werden, da es problematisch ist, Prognosewerte für ordinale Zielvariable mit dem ALM zu berechnen.

Dichotome Variable können im 1. Eingabefeld (auch zusätzlich zu quantitativen Variablen) eingegeben werden. Das hat den Vorteil, daß dann beliebig viele Zielvariable angegeben werden können.

Wird eine dichotome Variable im 2. Eingabefeld eingesetzt, was selbstverständlich korrekt ist, dann ist nur diese eine Zielvariable möglich. Werden dichotome Variable

im 1. Eingabefeld eingegeben, dann muß man in der Eingabe-Box 10 „Transformation der Prognosewerte“ entweder „5“ oder „7“ einsetzen, damit für den fehlenden Wert einer der beiden empirisch vorkommender (in der Regel ganzzahligen) Werte entsteht – und nicht ein Wert entsteht, dem keine der beiden empirischen Ausprägungen der dichotomen Variablen entspricht.

Wird in Eingabe-Box 10 „5“ eingesetzt, dann entsteht für die dichotome Variable derselbe Kein-Wert-Ersetzungswert, egal ob sie im 1. oder im 2. Eingabefeld eingetragen wurde. Wird „7“ eingetragen, dann entstehen etwas verschiedene KW-Ersetzungs-Werte, was durch die unterschiedliche Hinzufügung einer Zufallsvariation verursacht wird.

Eingabe-Box 7: Ursächliche Variable

The screenshot shows a software interface titled "Ursächliche Variable für die Zielvariable". It is organized into three main sections, each with a "Hilfe" button:

- ursächliche nominale Variable:** The input field contains "Hausbesitz, Beruf". To the right of the input field are two "Hilfe" buttons.
- Interaktions-Section:** Below the first section, there is text: "Interaktionen x. Ordnung zwischen den ursächlichen nominalen Variablen bilden oder einige ausgewählte Interaktionen bilden". Below this text is a radio button labeled "Ø =keine Interaktionen bilden" and a "Hilfe" button.
- ursächliche quantitative Variable:** The input field contains "Rueckrate, Laufzeit". To the right is a "Hilfe" button.
- ursächliche ordinale Variable:** The input field is empty. To the right is a "Hilfe" button.

Geben Sie in dieser Eingabe-Box jene Variable an, von denen Sie vermuten, dass sie die Zielvariable am besten determinieren. Wenn Sie in der Eingabe-Box "Zielvariable" mehrere quantitative Variable angegeben haben, dann berechnet Almo für jede Zielvariable getrennt, die Regressionskoeffizienten und Effekte der ursächlichen Variablen. Es ist nicht möglich für jede Zielvariable einen eigenen Satz von ursächlichen Variablen anzugeben. Die ursächlichen Variablen gelten gemeinsam für alle Zielvariablen.

Zum Begriff der "ursächlichen" Variablen:

Dieser Begriff darf nicht wörtlich genommen werden. Gemeint sind Variable, die dazu verwendet werden können, die Zielvariable zu determinieren. Das können durchaus "ursächliche" sein, d.h. Variable, die (in unserem Beispiel) Ursachen des Einkommens sind (wie etwa das Bildungsniveau); aber auch "Folgevariable", wie z.B. der Besitz von mehr oder weniger hochwertigen Konsumgütern. So ist möglicherweise der Autobesitz (Kleinwagen, Mittelklassewagen, Luxuswagen) eine sichtbare Folge der Einkommensverhältnisse - also eine Variable, die als Determinante des Einkommens mitverwendet werden kann.

Eingabefeld 1:

Geben Sie hier die nominalen ursächlichen Variablen an. Zu den Messniveaus "nominal", "ordinal" und "quantitativ" siehe P45.12.

Eingabefeld 2:

Wenn Sie Interaktionen zwischen den ursächlichen nominalen Variablen miteinbeziehen wollen, dann geben Sie hier die Interaktionsordnung an. Siehe dazu die ausführliche Erläuterung für Prog45mf in P45.15.1.2, Eingabe-Box „Ursächliche Variable“.

Eingabefeld 3:

Geben Sie hier die quantitativen ursächlichen Variablen an. Siehe dazu die ausführliche Erläuterung für Prog45mf in P45.15.1.2, Eingabe-Box „Ursächliche Variable“.

Eingabefeld 4:

Geben Sie hier die ordinalen ursächlichen Variablen an. Siehe dazu die ausführliche Erläuterung für Prog45mf in P45.15.1.2, Eingabe-Box „Ursächliche Variable“.

Eingabe-Box 8: Kein-Wert-Angabe und Umkodierungen

Kein-Wert-Angabe: Für ursächliche Variable und Zielvariable

Umkodierungen: Nur für ursächliche Variable
Zielvariable darf nicht umkodiert werden !!!
Umkodierungen sind temporär

Kein-Wert-Angabe
Umkodierungen

erzeuge zusätzliche Felder für Umkodierungen / Kein_Wert-Angaben

Kein-Wert-Angaben:

Almo muß selbstverständlich wissen, an welchen Codeziffern es den Kein-Wert-Fall erkennen kann. Hier gibt es 2 Vorgehensweisen:

1. Der Benutzer hat schon eine Almo-Arbeitsdatei (im Format DIREKT) erstellt. Dabei hat er Almo mitgeteilt, welche Codeziffern den Kein-Wert-Fall bezeichnen. In der Regel wird der Benutzer die Ziffer 0 als Kein-Wert-Code verwendet haben. Almo hat dann die Almo-Arbeitsdatei erzeugt und dabei die Kein-Wert-Codeziffer (beispielsweise die 0) durch einen Almo-internen Kein-Wert-Code (die riesige negative Zahl 10 hoch -38) ersetzt. In diesem Fall ist jetzt eine Kein-Wert-Angabe nicht mehr notwendig.

Mit dieser Vorgehensweise haben wir die Almo-Arbeitsdatei in den Programmen Prog45md und Prog45mh erzeugt.

2. Der Benutzer hat eine Kein-Wert-Deklaration noch nicht vorgenommen. In der Arbeitsdatei stehen also noch die ursprünglichen Codes (z.B. 0 für Kein-Wert). In diesem Fall muß der Benutzer jetzt eine Kein-Wert-Angabe vornehmen - beispielsweise so:

```
Beruf, Wohnort ( 0 = Kein_Wert )  
Einkommen      ( -1 = Kein_Wert )
```

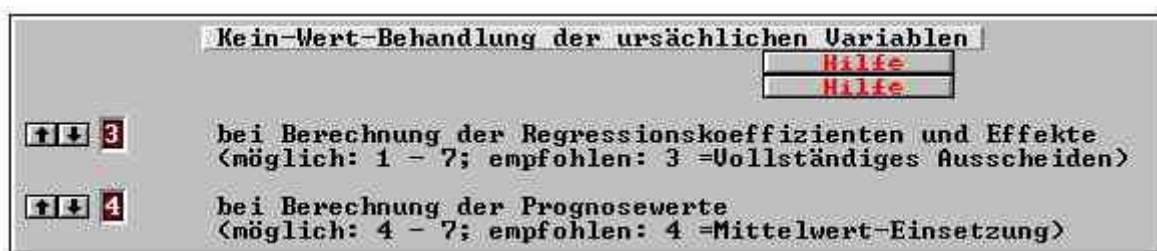
Umkodierungen:

In der Eingabe-Box "Kein-Wert-Angabe und Umkodierungen" können auch Variable umkodiert werden.

- i. Es dürfen nur die ursächlichen Variablen umkodiert werden - nicht die Zielvariablen.
- ii. Im Unterschied zum Programm Prog45mo (Mittelwert-Einsetzung) in Abschnitt P45.6 sind diese Umkodierungen nur temporär. Sie wirken nur während der Berechnung der Koeffizienten und der Prognosewerte. D.h. die eventuell umkodierten ursächlichen Variablen gehen in ihrer ursprünglichen Form in die neue Datei ein.

Wie Kein-Wert-Angaben und wie Umkodierungen zu erzeugen bzw. zu schreiben sind ist ausführlich in P0.5 beschrieben worden.

Eingabe-Box 9: Kein-Wert-Behandlung der ursächlichen Variablen



Eingabefeld 1: Um Prognosewerte für die Zielvariable zu bilden rechnet Almo ein Allgemeines Lineares Modell (ALM) für die Zielvariablen und die ursächlichen Variablen, die der Benutzer angibt.

Als Ergebnis dieses ALM entstehen u.a. Regressionskoeffizienten und Effekte der ursächlichen Variablen hinsichtlich der Zielvariablen. Dann werden mit Hilfe dieser Koeffizienten, Prognosewerte für jene Zielvariable ermittelt, die keinen Wert besitzen.

Die ursächlichen Variablen können nun auch fehlende Werte aufweisen. Die Katze beißt sich hier in den Schwanz. Um Ersatzwerte für fehlende Werte der Zielvariable zu schätzen, benötigen wir ursächliche Variable, die ihrerseits aber auch wieder fehlende Werte aufweisen können.

Prinzipiell gilt: Für die Berechnung der Regressionskoeffizienten und Effekte der ursächlichen Variablen hinsichtlich der Zielvariablen werden nur die Datensätze verwendet, bei denen die Zielvariable einen validen Wert besitzt. Besitzt ein Datensatz in der Zielvariablen keinen Wert, dann wird er aus der Berechnung der genannten Koeffizienten ausgeschlossen.

Besitzen in einem Datensatz eine (oder mehrere) ursächliche Variable keinen Wert (aber sehr wohl die Zielvariable), dann bietet Almo hier 7 "Kein-Wert-Behandlungs-Methoden" an.

Wenn der Benutzer auf den nachfolgenden Hilfeknopf in der Eingabe-Box klickt, dann werden ihm diese 7 Methoden gezeigt.

Empfehlung 1:

Wenn nur 1 Zielvariable vorhanden ist, dann sollte man die Kein-Wert-Behandlung 3, das "vollständige Ausscheiden" einsetzen. Ein Datensatz wird ausgeschlossen, wenn er auch nur in einer ursächlichen Variablen keinen Wert besitzt. Ist der

Datenverlust dabei zu groß, dann sollte man vorzugsweise die Kein-Wert-Behandlung 1, das "paarweise Ausscheiden" einsetzen.

Empfehlung 2:

Sind 2 oder mehrere quantitative bzw. dichotome Zielvariable vorhanden, dann kann ein großer Datenverlust eintreten. Wenn beispielsweise die drei Variablen y_1 , y_2 , y_3 als Zielvariable angegeben wurden, dann werden für die Berechnung der Koeffizienten für y_1 auch unnötigerweise die Datensätze ausgeschlossen, in denen y_2 oder y_3 keinen Wert besitzt. Wir empfehlen deswegen nur eine Zielvariable einzusetzen.

Eingabefeld 2: Oben wurde ausgeführt, dass mit Hilfe der Koeffizienten, die das ALM für die ursächlichen Variablen hinsichtlich der Zielvariablen liefert, Prognosewerte für jene Zielvariable ermittelt werden, die keinen Wert besitzen.

Bei der Berechnung der Prognosewerte stellt sich nun dasselbe Problem. Wie soll verfahren werden, wenn eine der ursächlichen Variablen keinen Wert besitzt? Das "vollständige Ausscheiden" eines Datensatzes, wenn auch nur eine ursächliche Variable keinen Wert besitzt, ist hier nicht möglich, da wir ja dann unsere Aufgabe, Ersatzwerte für fehlende Werte in der Zielvariablen zu erzeugen, nicht erfüllen könnten. Hier sind deswegen nur die "Kein-Wert-Behandlungs-Methoden" 4 bis 7 möglich, bei denen der Mittelwert (bzw. Median, bzw. Erwartungswert) der ursächlichen Variablen (eventuell mit einer "Zufalls-Überlagerung") eingesetzt wird.

Wenn der Benutzer auf den nachfolgenden Hilfefknopf in der Eingabe-Box klickt, dann werden ihm diese Methoden gezeigt.

Wir empfehlen deswegen nur eine Zielvariable einzusetzen.

Kein-Wert-Behandlung 1: "Paarweises Ausscheiden"

Wir werden dieses Verfahren in Abschnitt P45.12.4 sehr ausführlich darstellen. Hier wollen wir es nur kurz beschreiben.

Betrachten wir die Matrix der Abweichungsquadratsummen (kurz: Quadratsummen-matrix) zwischen den 3 Variablen V_1 , V_2 und V_3 .

	V_1	V_2	V_3
V_1	SS_{11}	SS_{12}	SS_{13}
V_2		SS_{22}	SS_{23}
V_3			SS_{33}

Die Quadratsumme SS_{12} zwischen den Variablen V_1 und V_2 wird aus den Datensätzen ermittelt, die in diesen beiden Variablen valide Werte besitzen. Entsprechend wird SS_{13} und SS_{23} berechnet. Die Folge dieser Vorgehensweise ist, dass die 3 Quadratsummen auf jeweils verschiedenen n_{ij} (Zahl der Untersuchungseinheiten) beruhen. In die Diagonale wird die Quadratsumme der Variablen selbst eingesetzt. SS_{11} ist also die Quadratsumme für die Variable V_1 , die sich aus den Untersuchungseinheiten ergibt, die in V_1 einen validen Wert besitzen. Entsprechend wird auch SS_{22} und SS_{33} gebildet. Dann wird jede Zelle der Quadratsummenmatrix zuerst durch das zu ihr gehörende n_{ij} dividiert. Dadurch entsteht die Kovarianzmatrix. Sie ist also die „durchschnittliche“ Quadratsummenmatrix. Also ermittelt nun das harmonische Mittel n_h aus den

unterschiedlichen n_{ij} des oberen Dreiecks der Matrix (ohne Diagonale). Die Kovarianzmatrix wird dann mit n_h multipliziert. Damit entsteht wieder eine Quadratsummenmatrix, diese Mal mit gleichen n_{ij} . Dieses Hochrechnen der Kovarianzmatrix zu einer neuen Quadratsummenmatrix könnte auch unterbleiben. Die Regressionskoeffizienten und Effekte sind die gleichen, egal ob wir für den Kalkül des ALM die Kovarianzmatrix oder die „hochgerechnete“ Quadratsummenmatrix verwenden. Dabei ist es sogar gleichgültig mit welchem n multipliziert wurde. Um die Signifikanzen ermitteln zu können, muß allerdings eine Entscheidung für ein bestimmtes n getroffen werden. Also entscheidet sich hier bei Programm Prgo45mm für das harmonische Mittel n_h . Bei anderen Almo-Programmen kann der Benutzer zwischen mehreren Möglichkeiten wählen. Wir stellen das in Abschnitt P45.12.4 dar.

Beachte: Wenn die Zielvariable selbst keinen Wert besitzt, dann wird hier in Prog45mm der gesamte Datensatz ausgeschlossen. Es findet kein paarweises Ausscheiden statt.

Kein-Wert-Behandlung 2: „Paarweises Ausscheiden II“

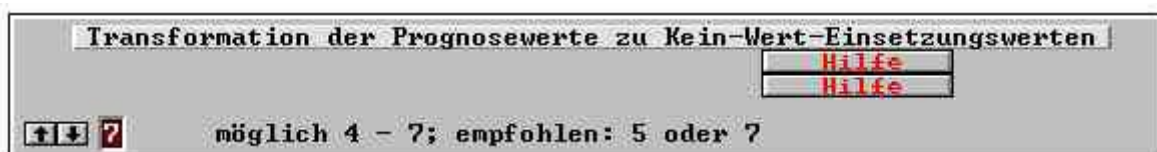
- a. Paarweises Ausscheiden bei ursächlichen quantitativen und ordinalen Variablen.
- b. Vollständiges Ausscheiden bei ursächlichen nominalen Variablen und deren Interaktionen, wenn auch nur eine der nominalen Analyse-Variablen den Wert "Kein_Wert" besitzt

Kein-Wert-Behandlung 3: „Vollständiges Ausscheiden“

Vollständiges Ausscheiden des gesamten Datensatzes, wenn auch nur eine der ursächlichen Analyse-Variable "Kein_Wert" ist.

Kein-Wert-Behandlung 4 bis 7 sind verschiedene Varianten der Mittelwert-Einsetzung. Sie sind in den Erläuterungen zu den Eingabe-Boxen des Programms P45mo dargestellt. Siehe P45.6.1, Erläuterung zu Eingabe-Box 7.

Eingabe-Box 10: Transformation der Prognosewerte zu Kein-Wert-Einsetzungswerten



Der errechnete Prognosewert kann, muß aber noch nicht, als KW-Einsetzungswert verwendet werden.

Beispielsweise besteht die Möglichkeit, den Kein-Wert-Einsetzungswert aus einem Intervall um den Prognosewert herum zufällig auszuwählen. Almo bietet hier folgende Möglichkeiten an.

Prognosewert-Behandlung = 4

- a. bei quantitativen Zielvariablen
Der Prognosewert wird unverändert für Kein_Wert eingesetzt.
- b. bei nominalen Variablen
Die Codeziffer mit dem größten Prognosewert (Wahrscheinlichkeit) wird eingesetzt.

Prognosewert-Behandlung = 5

a. bei quantitativen Zielvariablen

Der zum Prognosewert nächst gelegene empirische Wert wird für Kein_Wert eingesetzt.

b. bei nominalen Variablen

Die Codeziffer mit dem größten Prognosewert (Wahrscheinlichkeit) wird eingesetzt.

Prognosewert-Behandlung = 6

a. bei quantitativen Zielvariablen

Für Kein_Wert wird eingesetzt: der Prognosewert +/- einem normalverteilten Zufallswert mit Mittelwert=0 und Standardabweichung der Residuen der Variablen. Residuen sind die Differenz zwischen Prognosewert und empirischen Wert.

b. Bei nominalen Variablen wird der Prognosewert zufällig verändert.

Die Vorgehensweise soll an einem Beispiel gezeigt werden. Die nominale Variable sei der Beruf mit den 3 Ausprägungen Arbeiter, Angestellte, Sonstige. Dabei wurden für eine Untersuchungseinheit, die in der Zielvariablen keinen Wert besitzt, folgende Prognosewerte (Wahrscheinlichkeiten) ermittelt.

	Code	Prognosewert (Wahrscheinlichkeit)	kumuliert
Arbeiter	1	0.25	0.25
Angestellte	2	0.40	0.65
Sonstige	3	0.35	1.00

Die größte Wahrscheinlichkeit mit $p=0.4$ besitzt die Ausprägung 2 "Angestellter". Bei der Prognosewert-Behandlung 4 und 5 würde deswegen "2" als KW-Einsatzwert verwendet werden.

Bei der Prognosewert-Behandlung 6 und 7 wird nun eine gleichverteilte Zufallszahl zwischen 0 und 1 erzeugt.

Liegt sie zwischen

0	und 0.25,	dann wird für den fehlenden Wert 1 eingesetzt
0.25	und 0.65	2
0.65	und 1.00	3

Prognosewert-Behandlung = 7

a. bei quantitativen Zielvariablen wie bei 6.

Es wird aber der nächst gelegene empirische Wert eingesetzt.

Es wird zunächst ein Wert X errechnet, der sich ergibt aus dem Prognosewert +/- einem normalverteilten Zufallswert mit Mittelwert=0 und Standardabweichung der Residuen der Variablen. Residuen sind die Differenz zwischen Prognosewert und empirischem Wert. Dann wird der zu X nächst gelegene empirische Skalenwert für Kein_Wert eingesetzt.

b. Bei nominalen Variablen wie bei Prognosewert-Behandlung = 6

Wurden in Eingabe-Box 6 „Zielvariable“ **dichotome Variable** im Eingabefeld 1 eingetragen (und nicht als nominale Variable in Eingabefeld 2), dann muß Prognosewert-Behandlung 5 oder 7 verwendet werden. Siehe dazu die Begründung in den Erläuterungen zu Eingabe-Box 6.


Überlagerung durch Zufallswert

Es muß noch folgende Frage beantwortet werden:

Warum wird bei quantitativen Zielvariablen der Prognosewert bei Kein-Wert-Behandlung 6 und 7 durch einen Zufallswert überlagert?

Wird als KW-Einsetzungswert nur der Prognosewert verwendet, dann wird die Varianz der Variablen verringert. Werden mit den so erzeugten „vollständigen“ Daten beispielsweise Korrelationen errechnet, dann werden die Signifikanzen dieser Korrelationen überschätzt. Siehe dazu etwa R. J. A. Little/D. B. Rubin (1990, S. 381). Die Überlagerung durch einen normalverteilten Zufallswert mit der Standardabweichung der Residuen der Variablen bezweckt also, dass die Varianz der Variablen (fast) unverändert bleibt. Die Standardabweichung der Residuen ist identisch mit der Fehlerstreuung (ausgedrückt als Standardabweichung) des Gesamtmodells.

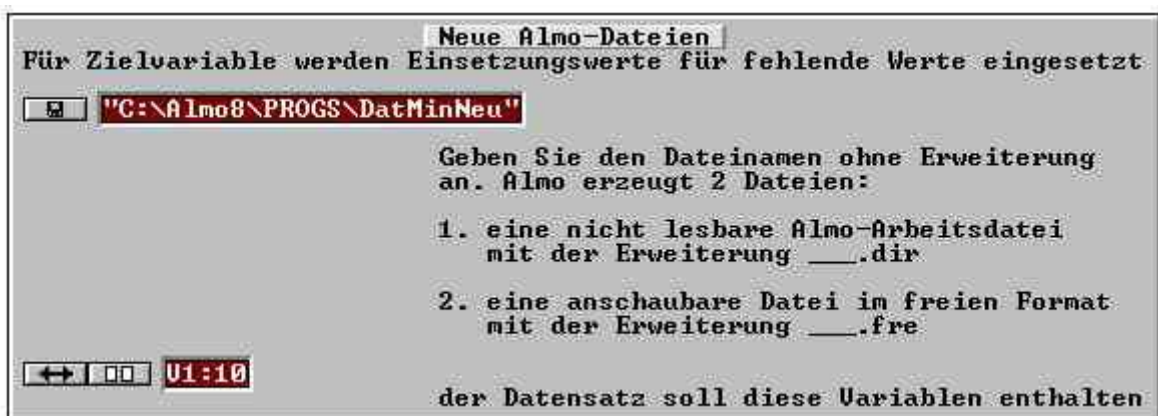
Eingabe-Box 11: Startwert für Zufallsgenerator (für Verfahren 6 und 7)



Bei der Kein-Wert-Behandlung 6 und 7 der ursächlichen Variablen und bei der Prognosewert-Transformation 6 und 7 errechnet Almo einen Zufallswert. Dieser wird durch den Almo-Zufallsgenerator erzeugt. Wenn Sie den Startwert für den Zufallsgenerator nicht ändern, dann wird exakt dieselbe Folge von Zufallszahlen erzeugt, wenn das Programm ein 2. und weitere Male gerechnet wird. Es werden also exakt dieselben Kein-Wert-Einsetzungswerte erzeugt.

Wird der Startwert verändert, dann entsteht eine andere Folge von Zufallszahlen und somit eine andere Folge von Kein-Wert-Einsetzungswerten.

Eingabe-Box 12: Neue Almo-Dateien



Eingabefeld 1: Geben Sie den Dateinamen für die Datei an, in die Sie Ihre Daten speichern wollen. Almo schreibt dabei für die als "Zielvariable" angegebenen Variablen die errechneten Einsetzungswerte. Die anderen Variablen werden mit ihren ursprünglichen Werten übernommen.

Geben Sie dabei den Dateinamen ohne Erweiterung an. Almo erzeugt dann 2 Dateien:

1. eine nicht lesbare Almo-Arbeitsdatei mit der Erweiterung __.dir

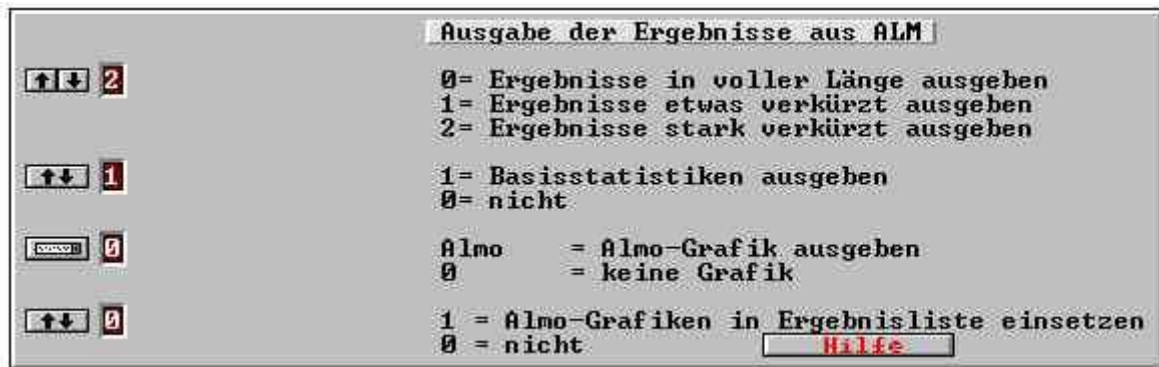
Im Beispiel: "C:\Almo6\PROGS\DatMinNeu.dir"

2. eine anschaubare Datei im freien Format mit der Erweiterung __.fre
Im Beispiel: "C:\Almo6\PROGS\DatMinNeu.fre"

In der "anschubaren" Datei können Sie sich nochmals die von Almo errechneten Kein-Wert-Einsetzungswerte anschauen.

Eingabefeld 2: Geben Sie die Nummern der Variablen an, die in die neue Datei übernommen werden sollen. In der Regel wird man alle Variable aus der Ursprungsdatei übernehmen. Verwenden Sie dabei die Schreibweise "V1:...".

Eingabe-Box 13: Ausgabe der Ergebnisse aus ALM



Eingabefeld 1: Almo rechnet, wenn mehrere Zielvariable vorhanden sind, eine multivariate Analyse und gibt, wenn auf "2" (stark verkürzte Ausgabe) eingestellt wurde nur eine zusammenfassende Ergebnistabelle aus, die keine Information über die einzelnen Zielvariablen enthält. Wir empfehlen deswegen auf "1" (etwas verkürzte Ausgabe) zu stellen, wenn mehrere quantitative (oder dichotome) Zielvariable vorhanden sind. Ist nur 1 vorhanden, dann sollte man auf "2" einstellen.

Eingabefeld 3: Man sollte auf "0" (keine Grafik) einstellen, da man sonst einen sehr umfangreichen Output erhält.

P45.7.1.4 Ausgabe aus Prog45mm

Almo hat zuerst das Allgemeine Lineare Modell gerechnet. Dabei wird folgendes ausgegeben (gekürzt):

```
Zahl der insgesamt eingelesenen Einheiten      1000  
Zahl der in die Analyse einbezogenen Einheiten  619  
=====
```

******* Erläuterung:**

Zur Ermittlung der Regressionskoeffizienten und Effekte haben wir die Kein-Wert-Behandlung des "vollständigen Ausscheidens" gewählt. Es wurden nur jene Datensätze verwendet, die in der Zielvariablen und in allen ursächlichen Variablen valide Werte besaßen. 1000 Datensätze wurden eingelesen. Es konnten jedoch nur 619 ausgewertet werden. Wir haben also einen Datenverlust von ca. 40 %.

Zusammenfassung

Streuungsquelle

Korrel Signifikanz

	Koeff.	p	(1-p) 100
alle unabh. Var. zusammen	0.4701	0.0000	99.9995
quant./ordin. Var. zusammen	0.4670	0.0000	99.9995
nominale Variable und ihre Interaktionen zusammen	0.0516	0.6559	34.4092
V7 Rueckrate	-0.4413	0.0000	99.9995
V8 Laufzeit	0.0157	0.6980	30.1981
V6 Hausbesitz	0.0125	0.7581	24.1878
V3 Beruf	0.0053	0.8951	10.4857
V6*V3	0.0223	0.5811	41.8924

******* Erläuterung:**

Ein sehr wichtiger Koeffizient ist (in der 1. Zeile) die multiple Korrelation und ihre Signifikanz. Sie beträgt 0.4701 und ist mit 99.99% signifikant. Die Determination der Zielvariablen des Einkommens ist also nicht gut. Wir müssen einen „ordentlichen“ multiplen Korrelationskoeffizienten verlangen, soll die Kein-Wert-Einsetzung erfolgreich sein. Ist er das nicht, dann können auch die Prognosewerte für die Kein-Wert-Fälle nicht gut sein. Geht die multiple Korrelation gegen 0, dann nähert sich der Prognosewert dem Mittelwert der Variablen an.

Die obige Tabelle ermöglicht es auch, festzustellen, welche ursächlichen Variablen die Zielvariable nur schwach determinieren. Das sind in unserem Beispiel alle nominalen ursächlichen Variablen (Hausbesitz, Beruf und deren Interaktion), sowie die quantitative ursächliche Variable der Laufzeit. Man erkennt das an den niedrigen Korrelationskoeffizienten und vor allem an der (nicht vorhandenen) Signifikanz. Es wäre zu überlegen, ob man ursächliche Variable, die nur sehr schwach die Zielvariable determinieren, aus der Analyse herausnimmt. Man muß dies aber nicht.

Wird eine ursächliche Variable herausgenommen, dann wird in der Regel der multiple Korrelationskoeffizient niedriger. Die Verringerung wird aber umso kleiner, je unbedeutender die Wirkung der ursächlichen Variablen ist.

Beachte: Werden 2 oder mehr quantitative Zielvariable angegeben, dann rechnet Almo eine „multivariate“ Analyse. Die obige zusammenfassende Tabelle wird dann durch eine andere äquivalente ersetzt.

Zusammenfassung: Effekte und Regressionskoeffizienten
und ihre Signifikanzen
hinsichtlich der abhaengigen Variablen
Einkommen

	Effekte Regress.koeff	Signifikanz (1-p) *100
A1 kein Haus	301.954540	24.183332
A2 hat Haus	-301.954540	24.183332
B1 Selbst	127.474186	10.481236
B2 Unselbst	-127.474186	10.481236
A1 B1	-538.230314	41.887943
A1 B2	538.230314	41.887943
A2 B1	538.230314	41.887943
A2 B2	-538.230314	41.887943
Rueckrate	-4.770476	99.999500
Laufzeit	37.171341	30.198060

******* Erläuterung:**

Dies sind die Effekte und Regressionskoeffizienten, die benötigt werden um die Prognosewerte zu errechnen. Die Konstante wird an dieser Stelle nicht ausgegeben.

Almo berechnet nun die Prognosewerte für alle Datensätze, also für die Datensätze mit validem Wert in der Zielvariablen und für die Datensätze, deren Zielvariable keinen Wert besitzt. Sie werden zwischengespeichert aber nicht ausgegeben.

Berechnung der Prognosewerte bzw. Schätzwerte fuer Variable mit Kein_Wert

```
-----
Mittelwert und Standardabweichung der Residuen
                Mittelwert   Standardabweichung
V4 Einkommen      401.281      12375.2
```

******* Erläuterung:**

Almo ermittelt für die Datensätze mit validem Wert in der Zielvariablen die Residuen. Residuen sind die Differenzen zwischen Prognosewert und empirischem Wert. Der Mittelwert der Residuen ist nahe 0, in unserem Beispiel mit 401.281 relativ hoch (der Mittelwert der Variablen selbst ist 24822.5904). Die Standardabweichung der Residuen wird verwendet um einen Streuungsbereich um den Prognosewert herum zu erzeugen.

Ursaechliche Variable, die Kein_Wert waren und durch Schaetzwerte ersetzt wurden

```
V6 Hausbesitz      in 92 Datensuetzen
V3 Beruf           in 102 Datensuetzen
V7 Rueckrate       in 81 Datensuetzen
V8 Laufzeit        in 105 Datensuetzen
```

******* Erläuterung:**

Die ursächliche Variable "Hausbesitz" besaß in 92 Datensätzen keinen Wert usw.

Almo ermittelt nun aus den Prognosewerten die Kein-Wert-Einsetzungswerte

Fuer "Kein_Wert" eingesetzter Schaetzwert

```
-----
Datensatz  Variable  Prognosewert  eingesetzter Wert
-----
1          V4        16631.1       16645
6          V4        22339.7       22296
16         V4        32181.8       32187
19         V4        22624.2       22604
26         V4        15483.8       15535
33         V4        29690.7       29695
54         V4        31097.1       31041
57         V4        17665.3       17649
60         V4        24560.9       24553
.          .         .             .
.          .         .             .
.          .         .             .
.          .         .             .
```

******* Erläuterung:**

Wir haben in der Eingabe-Box „Transformation der Prognosewerte...“ auf „5“ eingestellt. Almo übernimmt deswegen nicht den in der 3. Spalte angegebenen Prognosewert als KW_Einsetzungswert. Es sucht in den Datensätze mit validem Wert in der Zielvariablen den zum Prognosewert nächst gelegenen empirisch vorkommenden Wert.

Dieser wird als KW_Einsetzungswert verwendet. Beispiel: Im Datensatz 1 ist die Zielvariable V4 gleich Kein_Wert.

Der Prognosewert ist 16631.1. Der nächste empirisch vorkommende Wert ist 16645. Dieser wird als KW_Einsetzungswert verwendet.

Zahl der Datensätze mit Kein_Wert in Zielvariable

V4 Einkommen 87 Kein-Wert-Faelle

******* Erläuterung:**

In 87 Datensätzen besaß die Zielvariable "Einkommen" keinen Wert. D.h. in 87 Fällen wurden Kein-Wert-Einsetzungswerte eingesetzt.

```
***** MITTEILUNG
Lesen oder Schreiben korrekt beendet in Datei
"C:\Almo6\PROGS\DatMinNeu.fre"
```

```
***** MITTEILUNG
Lesen oder Schreiben korrekt beendet in Datei
"C:\Almo6\PROGS\DatMinNeu.dir"
```

******* Erläuterung:**

Mit diesen beiden Mitteilungen gibt Almo kund, dass es 2 neue Dateien erzeugt hat, eine im Format FREI (Erweiterung: ...fre) und eine im Format DIREKT (Erweiterung: ...dir). Alle in der Zielvariablen fehlenden Werte sind durch KW_Einsetzungswerte ersetzt.

Hätten wir in der Eingabe-Box „Transformation der Prognosewerte zu Kein-Wert-Einsetzungswerten“ auf „6“ eingestellt, dann hätten wir folgendes Ergebnis bekommen:

Fuer "Kein_Wert" eingesetzter Schaetzwert

```
-----
```

Datensatz	Variable	Prognosewert	eingesetzter Wert
1	V4	16631.1	16412.9
6	V4	22339.7	14153
16	V4	32181.8	39092.2
19	V4	22624.2	21026.1
26	V4	15483.8	11242.9
33	V4	29690.7	-711.204
54	V4	31097.1	38308.4
57	V4	17665.3	24808.7
60	V4	24560.9	3080.45
.	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.

Hier wird der Prognosewert nicht unmittelbar als Kein-Wert-Einsetzungswert übernommen. Dieser wird aus einem normalverteilten Zufallsbereich um den Prognosewert herum entnommen. Als Standardabweichung für diesen Zufallsbereich wird die Standardabweichung der Residuen verwendet. Wie man sieht weichen Prognosewert und der Kein-Wert-Einsetzungswert teilweise stark voneinander ab. Die Ursache dafür ist, daß die Determination der Zielvariablen

durch die ursächlichen Variablen in unserem Beispiel schlecht ist – wodurch die Standardabweichung der Residuen und demzufolge des Zufallsbereichs um den Prognosewert herum sehr groß ist.

P45.7.2 Schritt 5c: Prognosewerte für fehlende Werte durch Logitanalyse ermitteln

Ist die Zielvariable, für die wir Prognosewerte für fehlende Werte einsetzen wollen, nominal (dichotom oder polytom), dann verwenden wir vorzugsweise die Logitanalyse. Siehe dazu die Begründung in Abschnitt P45.15.1.

Im nachfolgenden Programm Prog45mz kann allerdings nur 1 nominale Zielvariable eingegeben werden. Hat man mehrere nominale Zielvariable, für die man Prognosewerte für fehlende Werte einsetzen möchte, dann muß man nacheinander mehrere Analysen rechnen.

Das Modell der Logitanalyse wird ausführlich in Abschnitt P45.16 beschrieben.

Wir verwenden für das folgende Programm Prog45mz dieselben Daten "DatMinKW.dir", wie wir sie auch für Prog45mm verwendet haben. Als Zielvariable, deren fehlende Werte geschätzt werden sollen, setzen wir die nominal-polytome Variable "gekauft Produkt" ein.

P45.7.2.1 Eingabe in Prog45mz

Prog45mz.Msk

Einsetzungswerte für fehlende Werte ermitteln
mit Hilfe der Logit-Analyse

Die Einsetzung erfolgt in folgenden 3 Schritten:

Schritt 1:
Zuerst wird eine Logitanalyse gerechnet.
Dabei entstehen u.a. Regressionskoeffizienten

Schritt 2:
Dann werden mit Hilfe der dabei errechneten Koeffizienten
Prognosewerte für jene Zielvariable ermittelt, die
keinen Wert besitzt

Schritt 3:
Aus diesen Prognosewerte werden dann die Einsetzungswerte
für die fehlenden Werte gebildet.
Die Daten werden dann in eine neue Datei abgespeichert

siehe Handbuch "P45 Almo-Data-Mining", Abschnitt 45.7

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\DatMinKW.dir"

6

für die Prognosewerte für fehlende Werte eingesetzt werden sollen

 <nur eine> nominale Zielvariable

ursächliche nominale Variable

ursächliche quantitative Variable

8

Kein-Wert-Angabe: Für ursächliche Variable und Zielvariable

Umkodierungen: Nur für ursächliche Variable
Zielvariable darf nicht umkodiert werden !!!
Umkodierungen sind temporär

Kein-Wert-Angabe

Umkodierungen

erzeuge zusätzliche Felder für Umkodierungen / Kein_Wert-Angaben

9

Kein-Wert-Behandlung der ursächlichen Variablen

3 bei Berechnung der Regressionskoeffizienten und Effekte
nur 3 = Vollständiges Ausscheiden möglich

4 bei Berechnung der Prognosewerte
(möglich: 4 - 7; empfohlen: 4 =Mittelwert-Einsetzung)

10

Transformation der Prognosewerte zu Kein-Wert-Einsetzungswerten

7 möglich 4 - 7; empfohlen: 7

11

Startwert für Zufallsgenerator

(für Verfahren 6 und 7)

123457

12

Neue Almo-Dateien

Für Zielvariable werden Einsetzungswerte für fehlende Werte eingesetzt

"C:\Almo7\PROGS\DatMinNeu"

Geben Sie den Dateinamen ohne Erweiterung an. Almo erzeugt 2 Dateien:

- eine nicht lesbare Almo-Arbeitsdatei mit der Erweiterung **__.dir**
- eine anschaulbare Datei im freien Format mit der Erweiterung **__.fre**

U1:10

der Datensatz soll diese Variablen enthalten

P45.7.2.2 Erläuterungen zu den Eingabe-Boxen

Prog45mz entspricht weitgehend dem in Abschnitt P45.7.1.3 bereits erläuterten Prog45mm, so daß wir hier nur 3 Eingabe-Boxen erläutern müssen.

Eingabe-Box 6: Zielvariable

Eingabe-Box 7: Ursächliche Variable für die Zielvariable

Zielvariable für die Prognosewerte für fehlende Werte eingesetzt werden sollen	
<input type="text" value="Produkt"/>	\
	(nur eine) nominale Zielvariable
Ursächliche Variable für die Zielvariable	
<input type="text" value="Wohnort, Geschlecht, Hausbesitz"/>	\
	ursächliche nominale Variable
<input type="text" value="Einkommen"/>	\
	ursächliche quantitative Variable

Es ist nur 1 nominale Zielvariable erlaubt. Sie kann dichotom oder polytom sein. Als ursächliche Variable sind nur nominale (dichotom und polytom) und/oder quantitative Variable erlaubt.

Die Eingabe-Box 9: Kein-Wert-Behandlung der ursächlichen Variablen

Kein-Wert-Behandlung der ursächlichen Variablen	
	<input type="text" value="Hilfe"/>
<input type="text" value="3"/>	\
	bei Berechnung der Regressionskoeffizienten und Effekte nur 3 = Vollständiges Ausscheiden möglich
<input type="text" value="4"/>	\
	bei Berechnung der Prognosewerte (möglich: 4 - 7; empfohlen: 4 =Mittelwert-Einsetzung)

Im Unterschied zu Prog 45mm wird bei der Logitanalyse nur und ausschließlich das "vollständige Ausscheiden" durchgeführt. Der Benutzer kann das nicht beeinflussen. Wenn auch nur eine Analyse-Variable Kein-Wert ist, dann wird der gesamte Datensatz aus der Analyse ausgeschlossen. Hingegen kann bei der Berechnung der Prognosewerte die Kein-Wert-Behandlung wie bei Prog45mm gewählt werden.

P45.7.2.3 Ausgabe aus Prog45mz

Almo gibt zuerst die Ergebnisse aus der Logitanalyse aus. Wir werden diese nur insofern erläutern, als sie für unser Thema der Kein-Wert-Einsetzung von Belang sind. In Abschnitt P45.16.1.3 wird die Ausgabe aus der Logitanalyse im Detail behandelt.

Modellspezifikation: mehrdimensionales Logit-Modell

unabhaengige nominale Variablen:

```
-----
      V1      Wohnort      Werte-Untergrenze = 1 Obergrenze = 2
      V2      Geschlecht  Werte-Untergrenze = 1 Obergrenze = 2
      V6      Hausbesitz   Werte-Untergrenze = 1 Obergrenze = 2
```

Beachte:

Fuer die unabhaengigen nominalen Variablen wird die 0,1,-1 Dummy-Kodierung verwendet.

unabhaengige quantitative Variablen:

```
-----
      V4      Einkommen
```

abhaengige nominale Variable:

```
-----
      V9      Produkt      Werte-Untergrenze = 1 Obergrenze = 3
```

Beachte:

Zur Schaetzung wird die 1. Auspraegung der abhaengigen Variablen als Referenz verwendet

Datensatz wird wegen fehlender Werte
oder negativer Haeufigkeiten eliminiert

```
Datensatz 1
Datensatz 2
Datensatz 3
Datensatz 6
Datensatz 7
Datensatz 9
Datensatz 11
.
.
.
.
Datensatz 990
Datensatz 991
Datensatz 993
Datensatz 996
Datensatz 999
```

Zahl der eingelesenen Datensaeetze = 1000

Zahl der in Analyse einbezogenen Datensaeetze = 606

***** Erläuterung:

Es wurden insgesamt 1000 Datensätze eingelesen, es konnten jedoch nur 606 Datensätze analysiert werden, da 394 in mindestens einer der Analysevariablen keinen Wert aufwiesen. Selbstverständlich sind dabei die Datensätze inkludiert, die in der abhängigen Variablen „Produkt“ keinen Wert hatten.

Maximum-Likelihood-Schaetzer der Koeffizienten:

Ergebnisse fuer 2. Auspraegung "Möbel" der abhaengigen Variablen V9 Produkt
(als Referenz wird die 1. Auspraegung "Kleidung" verwendet)

unabhaengige Variable			Regress. koeff.β	Risiko epx(β)	relatives Risiko	Signifikanz (1-p)*100	partielle Korrelation
A1	Wohnort:	Stadt	-0.12471	0.88276	-11.72430	72.82	-0.02477
A2	Wohnort:	Land	0.12471	1.13281	13.28146	72.82	0.02477

B1	Geschlec:	m	0.20752	1.23062	23.06234	94.08	0.03475
B2	Geschlec:	w	-0.20752	0.81260	-18.74037	94.08	-0.03475
C1	Hausbesi:kein Hau		-0.06005	0.94172	-5.82839	32.07	-0.03764
C2	Hausbesi:hat Haus		0.06005	1.06189	6.18911	32.07	0.03764
V4	Einkommen		0.00000	1.00000	0.00008	8.31	0.03926

Ergebnisse fuer 3. Auspraegung "Technik" der abhaengigen Variablen V9 Produkt
(als Referenz wird die 1. Auspraegung "Kleidung" verwendet)

unabhaengige Variable			Regress. koeff.β	Risiko epx(β)	relatives Risiko	Signifikanz (1-p)*100	partielle Korrelation
A1	Wohnort:	Stadt	-0.07492	0.92781	-7.21865	49.81	-0.03464
A2	Wohnort:	Land	0.07492	1.07780	7.78028	49.81	0.03464
B1	Geschlec:	m	0.08187	1.08532	8.53162	54.86	0.03331
B2	Geschlec:	w	-0.08187	0.92139	-7.86095	54.86	-0.03331
C1	Hausbesi:kein Hau		-0.00191	0.99809	-0.19081	0.85	-0.03936
C2	Hausbesi:hat Haus		0.00191	1.00191	0.19118	0.85	0.03936
V4	Einkommen		-0.00000	1.00000	-0.00011	10.82	-0.03918

******* Erläuterung:**

Almo teilt die Regressionskoeffizienten (und weitere Koeffizienten) der ursächlichen Variablen mit. Almo verwendet diese Regressionskoeffizienten für 2 Zwecke:

2. Um Prognosewerte für jene Personen zu rechnen, für die die Werte aller Analysevariablen vorhanden sind.
3. Um Prognosewerte für jene Personen zu rechnen, die in der Zielvariablen keinen Werte besitzen - und diesen dann als Kein-Wert-Einsetzungswert einzusetzen.

Trefferhaeufigkeiten bei Individualdaten
fuer abhaengige Variable V9 Produkt

		tatsaechlich			prognostiziert absolut		
		1	2	3	1	2	3
		Kleidu	Möbel	Techni	Kleidu	Möbel	Techni
Kleidung	1	131	0	0	0	47	84
Möbel	2	0	232	0	0	109	123
Technik	3	0	0	243	0	94	149

		prognostiziert relativ			erwartet Zufall		
		1	2	3	1	2	3
		Kleidu	Möbel	Techni	Kleidu	Möbel	Techni
Kleidung	1	28.8	49.4	52.7	28.3	50.2	52.5
Möbel	2	49.4	89.9	92.7	50.2	88.8	93.0
Technik	3	52.7	92.7	97.6	52.5	93.0	97.4

absolut: Chi-Quadrat(4) =191.585 Signifikanz 100*(1-p) = 100.000
relativ: Chi-Quadrat(4) = 0.046 Signifikanz 100*(1-p) = 0.136

******* Erläuterung:**

Für jene Personen, für die in allen ursächlichen Variablen und in der Zielvariablen valide Werte vorhanden waren, prognostiziert Almo welches Produkt sie gekauft haben. Diese Prognose wird dann mit dem tatsächlichen Produktkauf verglichen. So kann die Trefferhäufigkeit festgestellt werden. Betrachten wir aus diese Tabelle der Trefferhäufigkeiten die oberen beiden Teiltabellen und die Teiltabelle unten rechts. Dabei genügt es die Diagonalen anzuschauen:

	tatsächlich	richtig prognostiziert	zufällig richtig prognostiziert
Kleidung	131	0	28
Möbel	232	109	88
Technik	243	149	97

Möbel und Technik konnte durch die Logitanalyse besser prognostiziert werden, wie wenn man zufällig die Personen den Produkten zugewiesen hätte. Bei Kleidung ist unser Modell aber schlechter als der Zufall. Die Ursache für dieses wenig befriedigende Ergebnis ist, daß die ursächlichen Variablen die Zielvariable nur schwach determinieren.

Berechnung der Prognosewerte bzw. Schätzwerte fuer Variable mit Kein_Wert

```

***** MITTEILUNG
      Sind unabhaengige Variable, die für die Berechnung
des Prognosewerts benoetigt werden, gleich "Kein_Wert"
dann wird fuer sie "Kein-Wert-Behandlung = 4" durchgefuehrt

```

Mittelwert und Standardabweichung der Residuen
fuer Variable V9 Produkt: Kleidung, Möbel, Technik

	Mittelwert	Standardabweichung
Gruppe 1 Kleidung	-0.00410058	0.404612
Gruppe 2 Möbel	-0.00250691	0.485784
Gruppe 3 Technik	0.00660748	0.491555

******* Erläuterung:**

Almo ermittelt für die Datensätze mit validem Wert in der Zielvariablen die Residuen. Residuen sind die Differenz zwischen Prognosewert und empirischem Wert. Der Mittelwert der Residuen ist nahe 0. Die Standardabweichung der Residuen wird verwendet, wenn der Benutzer in Eingabe-Box 10 "Transformation der Prognosewerte zu Kein-Wert-Einsatzwerten" die Methode 6 oder 7 einträgt. Diese beiden Methoden erzeugen für die Personen, die in der Zielvariablen keinen Wert haben, einen KW-Einsatzwert, der aus einer normalverteilten Zufallsvariation des Prognosewertes mit der oben angegebenen Standardabweichung hervorgeht.

Ursächliche Variable, die Kein_Wert waren
und durch Schätzwerte ersetzt wurden

V1	Wohnort	in	101	Datensätzen
V2	Geschlecht	in	104	Datensätzen
V6	Hausbesitz	in	92	Datensätzen
V4	Einkommen	in	87	Datensätzen

******* Erläuterung:**

Beispiel: Die ursächliche Variable "Wohnort" besaß in 101 Datensätzen keinen Wert usw. Da wir in Eingabe-Box 9 "Kein-Wert-Behandlung der ursächlichen Variablen" Eingabefeld 2 ("bei Berechnung der Prognosewerte") als "Kein-Wert-Behandlung = 4" eingesetzt hatten, wird in diesen Variablen der Mittelwert (bei quantitativen Variablen) bzw. der Erwartungswert (bei nominalen Variablen) als Ersatzwert eingesetzt.

Fuer "Kein_Wert" eingesetzter Schaetzwert

***** MITTEILUNG
Der fuer fehlende Werte eingesetzte Schaetzwert
wird noch der "Prognosewert-Behandlung = 7" unterworfen

Datensatz	Variable	eingesetzter Wert
1	V9 Produkt	3
7	V9 Produkt	2
19	V9 Produkt	2
26	V9 Produkt	3
32	V9 Produkt	3
34	V9 Produkt	3
.	.	.
.	.	.
.	.	.
.	.	.
.	.	.
.	.	.
914	V9 Produkt	3
924	V9 Produkt	3
926	V9 Produkt	3
957	V9 Produkt	2
989	V9 Produkt	3
990	V9 Produkt	3

Zahl der Datensaeetze mit Kein_Wert in Zielvariable

V9 Produkt 102 Kein-Wert-Faelle

******* Erläuterung:**

Die Datensätze, für die ein Kein-Wert-Einsetzungswert eingesetzt wurde, werden aufgelistet. Im Datensatz 1 beispielsweise wurde für die Zielvariable "V9 Produkt" der Wert 3 (=Technik) für Kein-Wert eingesetzt.

***** MITTEILUNG
Lesen oder Schreiben korrekt beendet in Datei
"C:\Almo6\PROGS\DatMinNeu.fre"

***** MITTEILUNG
Lesen oder Schreiben korrekt beendet in Datei
"C:\Almo6\PROGS\DatMinNeu.dir"

******* Erläuterung:**

Almo teilt abschliessend noch mit, dass es die neue Datei "DatMinNeu" angelegt hat, einmal im Format FREI (Erweiterung: .fre) und einmal im Format DIREKT (Erweiterung: .dir). Letztere ist eine Almo-Arbeitsdatei, die in allen Data-Mining-Programmen eingesetzt werden kann. Die neuen Dateien enthalten nun für die Variable "Produkt" keine Kein-Wert-Fälle mehr.

P45.7.3 Schritt 5d: Multiple Imputation

Prog45mm kann verwendet werden, um die von Rubin vorgeschlagene "multiple Imputation" durchzuführen. Siehe dazu Rubin (1987).

Das Verfahren ist - kurz beschrieben - folgendes: Aus einer Datei mit fehlenden Werten werden mehrere (etwa 3) Dateien erzeugt, in denen die fehlenden Werte (mit einem zufallsgesteuerten Verfahren) ersetzt sind. Dazu wird unser Prog45mm verwendet. Mit diesen 3 Dateien wird dann die gewünschte statistische Analyse gerechnet. So entstehen 3 etwas verschiedene Ergebnisse. Die interessierenden Parameter werden gemittelt.

Wir wollen die multiplen Imputation am Beispiel einer Regressionsanalyse mit unvollständigen Daten kurz darstellen. Siehe auch die kompakte Darstellung und die Simulation bei Paul D. Allison (2000).

1. Schritt: Erzeuge 3 (oder mehrere) Dateien mit vollständigen Daten
 - a. Prog45mm wird 3 Mal auf dieselbe unvollständige Datei angewendet.
 - b. Dabei wird in Eingabe-Box "Transformation der Prognosewerte" auf 6 oder 7 gestellt. Dadurch wird der Kein-Wert-Einsetzungswert aus einem normalverteilten Zufallsbereich um den Prognosewert herum entnommen (bei quantitativer Zielvariablen). Bei "7" wird der Kein-Wert-Einsetzungswert noch nicht verwendet. Es wird der zu ihm nächste empirisch vorkommende Wert verwendet.
 - c. Die Startzahl für den Zufallsgenerator in Eingabe-Box 11 muss jedes Mal eine andere sein. So entstehen 3 neue Dateien mit verschiedenen Kein-Wert-Einsetzungswerten.
2. Schritt: Mit den 3 Dateien wird je eine Regressionsanalyse gerechnet (mit Prog45mf oder Prog20mo)
3. Schritt: Die Regressionskoeffizienten werden über die 3 Analysen gemittelt. Der so entstehende Mittelwert ist der gesuchte Schätzer.
4. Schritt: Der Standardfehler des so gefundenen Regressionskoeffizienten i wird gemäß folgender Formel errechnet:

$$\sqrt{\frac{1}{M} \sum_k s_k^2 + \left(1 + \frac{1}{M}\right) \left(\frac{1}{M-1}\right) \sum_k (b_k - b^*)^2}$$

M = Zahl der Dateien

k = Index für Dateien

b_k = Regressionskoeffizient i aus Datei k

s_k = Standardfehler für Regressionskoeffizient i aus Datei k

b^* = Mittelwert aus den Regressionskoeffizienten i aus den k Dateien

Ein Beispiel

Wir wollen eine Regressionsanalyse mit y als Zielvariablen und x_1 und x_2 als ursächlicher Variablen rechnen. Die Datei sei folgende: ("kw" steht für Kein-Wert)

x1	x2	y
3	2	2
5	kw	5
3	4	5
kw	2	2
4	3	3
2	5	kw
3	5	4
5	3	3
3	3	3
3	1	kw
4	4	4
1	2	2
4	4	4
kw	4	3
4	3	kw
3	0	0
4	2	2
3	2	3
5	kw	3
4	2	3

Die Datei ist in Almo unter dem Namen ".\Testdat\KWSim.fre" enthalten. Die Zielvariable y weist in 3 Datensätzen Kein-Wert auf. Auch für die ursächliche Variable haben wir - um realistisch zu sein - je 2 Mal Kein-Wert eingesetzt.

Die Vorgehensweise ist nun folgende:

1. Schritt: Mit Prog45mh erzeugen wir eine Almo-Arbeitsdatei im Format DIREKT. Sie ist in Almo unter ".\Testdat\KWSim.dir" zu finden.
2. Schritt: Mit Prog45mm rechnen wir 3 Ananalysen mit 3 verschiedenen Startzahlen für den Zufallsgenerator. Wir zeigen das Prog45mm für die 1. neue Datei

Variablenamen

Datei der Variablenamen

zeige = Namensdatei in Output zeigen
 leer = nicht zeigen

Freie Namensfelder

Leere alle Eingabefelder dieser Sub-Box

Name1=x1
 Name2=x2
 Name3=y

erzeuge zusätzliche Namensfelder

Variablenamen in Datei speichern

Eingabefeld leer = nicht speichern

Datei aus der gelesen wird

"C:\Almo15\TESTDAT\KWSim.dir"

direkt Format der Daten

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

Zielvariable

für die Prognosewerte für fehlende Werte eingesetzt werden sollen

BEACHTTE: Erlaubt sind:

1. Beliebige viele quantitative und/oder dichotome Variable oder (exklusiv)
2. Eine nominale Variable mit beliebig vielen Ausprägungen

   y

(mehrere) quantitative/dichotome Zielvariable

(nur eine) nominale Zielvariable

[Hilfe](#)

Ursächliche Variable für die Zielvariable

ursächliche nominale Variable

[Hilfe](#)

[Hilfe](#)

  0

Interaktionen x. Ordnung zwischen den ursächlichen nominalen Variablen bilden

oder einige ausgewählte Interaktionen bilden

0 =keine Interaktionen bilden

[Hilfe](#)

ursächliche quantitative Variable

[Hilfe](#)

  x1,x2

ursächliche ordinale Variable

[Hilfe](#)


  

Kein-Wert-Angabe: Für ursächliche Variable und Zielvariable

Umkodierungen: Nur für ursächliche Variable
Zielvariable darf nicht umkodiert werden !!!
Umkodierungen sind temporär


Kein-Wert-Angabe


Umkodierungen




erzeuge zusätzliche Felder für Umkodierungen / Kein_Wert-Angaben

Kein-Wert-Behandlung der ursächlichen Variablen

 **3** bei Berechnung der Regressionskoeffizienten und Effekte
(möglich: 1 - 7; empfohlen: 3 =Vollständiges Ausscheiden)


 **4** bei Berechnung der Prognosewerte
(möglich: 4 - 7; empfohlen: 4 =Mittelwert-Einsetzung)

Transformation der Prognosewerte zu Kein-Wert-Einsetzungswerten

 **6** möglich 4 - 7; empfohlen: 5 oder 7


Startwert für Zufallsgenerator

(für Verfahren 6 und 7)

 **123457**


Neue Almo-Dateien

Für Zielvariable werden Einsetzungswerte für fehlende Werte eingesetzt

 **"C:\Almo15\PROGS\KWneu1"**

Geben Sie den Dateinamen ohne Erweiterung an. Almo erzeugt 2 Dateien:

1. eine nicht lesbare Almo-Arbeitsdatei mit der Erweiterung `__.dir`
2. eine anschaulbare Datei im freien Format mit der Erweiterung `__.fre`

 **v1:3**

der Datensatz soll diese Variablen enthalten

Beachte:

- In der Eingabe-Box "Kein-Wert-Angabe: Für ursächliche Variable und Zielvariable" wird nichts eingegeben, da schon bei der Erzeugung der Datei

- ".\Testdat\KWSim.fre" der Kein-Wert-Code in die Datensätze eingeschrieben wurde.
- In der Eingabe-Box "Kein-Wert-Behandlung der ursächlichen Variablen" wird festgelegt, dass Datensätze für die Berechnung der Regressionskoeffizienten vollständig ausgeschieden werden, wenn auch nur eine der ursächlichen Variablen oder der Zielvariablen keinen Wert besitzen. Um einen zu großen Datenverlust zu vermeiden wäre hier auch das "paarweise Ausscheiden" zulässig.
 - Das Programm erzeugt die neue nunmehr vollständige Datei ".\Progs\KWneu1.fre" (im Format FREI) und ".\Progs\KWneu1.dir" (im Format DIREKT).

Ergebnis aus 1. Lauf von Prog45mm

Prog45mm durchläuft insgesamt 3 Rechenschritte: Zuerst werden die Ergebnisse aus der Berechnung der Regressionskoeffizienten (mit Hilfe des Allgemeinen Linearen Modells) gezeigt:

Zahl der insgesamt eingelesenen Einheiten	20		
Zahl der in die Analyse einbezogenen Einheiten	13		
Streuungsquelle	Korrel Koeff.	Signifikanz p	(1-p)100
-----	-----	-----	-----
alle unabh. Var. zusammen	0.9075	0.0004	99.9644
x1	0.1046	0.7461	25.3887
x2	0.9019	0.0000	99.9995

Wir erkennen, dass „alle unabh. Var. zusammen“ einen sehr hohen multiplen Korrelationskoeffizienten von 0.9075 besitzen. Unser Modell ist also sehr gut. Wir sehen aber auch, dass x_1 die Zielvariable nicht signifikant determiniert. Es wäre zu überlegen, ob man sie aus der Analyse ausscheiden sollte.

Der 2. Rechenschritt besteht in der Berechnung der Prognosewerte. Also stellt fest, daß in jeweils 2 Datensätzen, die ursächlichen Variablen keinen Wert besitzen. Es setzt für diese den Mittelwert ein. Es muß eine Einsetzung vorgenommen werden, da sonst für die Zielvariable kein Prognosewert ermittelt werden kann.

Ursaechliche Variable, die Kein_Wert waren und durch Schaetzwerte ersetzt wurden			
x1	in	2	Datensaetzen
x2	in	2	Datensaetzen

Also berechnet noch den Mittelwert und die Standardabweichung der Residuen. „Residuum“ ist die Differenz zwischen empirischen y-Wert und Prognosewert (natürlich nur berechenbar, wenn der empirische y-Wert nicht fehlt)

	Mittelwert	Standardabweichung
y	0.0348661	0.689983

Die Standardabweichung wird verwendet, um (im nachfolgend beschriebenen 3. Rechenschritt) das normalverteilte Zufallsintervall zu bilden.

Im 3. Rechenschritt bildet Also aus den Prognosewerten die Kein-Wert-Einsetzungswerte und erzeugt die neue Datei. Prog45mm hat in der 1. neuen Datei ".\Progs\KWneu1.dir" folgende Kein-Wert-Einsetzungswerte gebildet:

Fuer "Kein_Wert" eingesetzter Schaetzwert

Datensatz	Variable	Prognosewert	eingesetzter Wert
6	y	4.77281	4.76065
10	y	1.36851	0.912058
15	y	3.15921	3.5445

Wir haben in der Eingabe-Box "Transformation der Prognosewerte" auf 6 gestellt. Dadurch wird der Kein-Wert-Einsetzungswert aus einem normalverteilten Zufallsbereich um den Prognosewert herum entnommen (bei quantitativer Zielvariablen). Hätten wir auf "7" gestellt, dann würde der zu oben angegebenen "eingesetzten Wert" nächste empirisch vorkommende Wert verwendet. Das wären dann die Werte 5, 1, 4.

Ergebnis aus 2. Lauf von Prog45mm

Die 2. neue Datei erzeugen wir mit der Startzahl 323457 für den Zufallsgenerator.

Es entsteht die 2. neue Datei ".\Progs\KWneu2.dir" mit folgende Kein-Wert-Einsetzungswerten

Fuer "Kein_Wert" eingesetzter Schaetzwert

Datensatz	Variable	Prognosewert	eingesetzter Wert
6	y	4.77281	5.09578
10	y	1.36851	1.022
15	y	3.15921	3.95517

Ergebnis aus 3. Lauf von Prog45mm

Die 3. neue Datei erzeugen wir mit der Startzahl 373457 für den Zufallsgenerator. Es entsteht die 2. neue Datei ".\Progs\KWneu3.dir" mit folgenden Kein-Wert-Einsetzungswerten

Fuer "Kein_Wert" eingesetzter Schaetzwert

Datensatz	Variable	Prognosewert	eingesetzter Wert
6	y	4.77281	4.70025
10	y	1.36851	2.56192
15	y	3.15921	3.50267

Der Prognosewert ist selbstverständlich in den 3 Analysen derselbe. Da der Kein-Wert-Einsetzungswert aus einem normalverteilten Zufallsintervall um den Prognosewert herum entnommen ist, ist er in den 3 Analysen jeweils ein anderer.

3. Schritt: Mit Prog45mf rechnen wir nun für die 3 neuen „vollständigen“ Dateien je eine Regressionsanalyse für die abhängige Variable y. Wir erhalten folgende Ergebnisse:

Variable	1. Analyse		2. Analyse		3. Analyse	
	Regr. koeff.	Standard fehler	Regr. koeff.	Standard fehler	Regr. koeff.	Standard fehler
x1	0.2309	0.1355	0.2107	0.1409	0.1959	0.1530

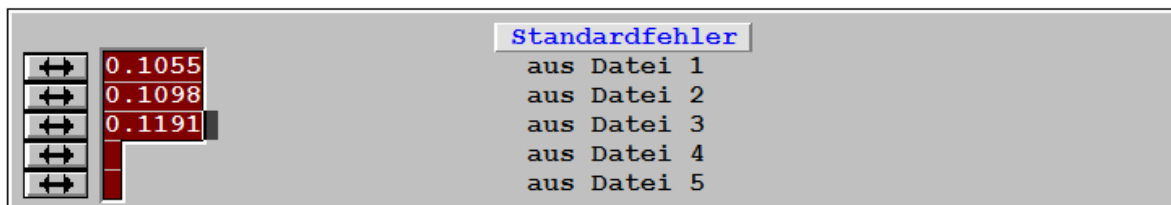
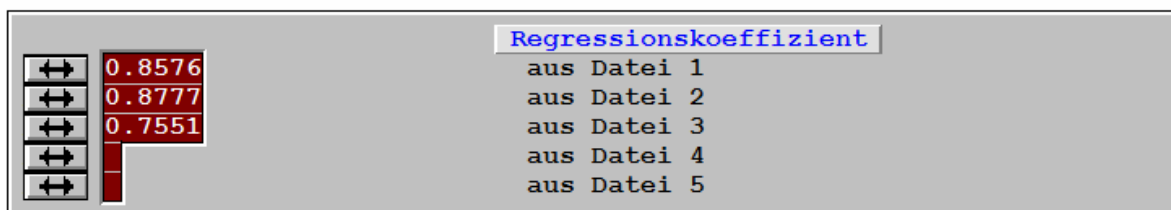
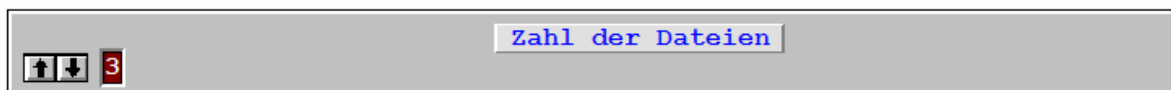
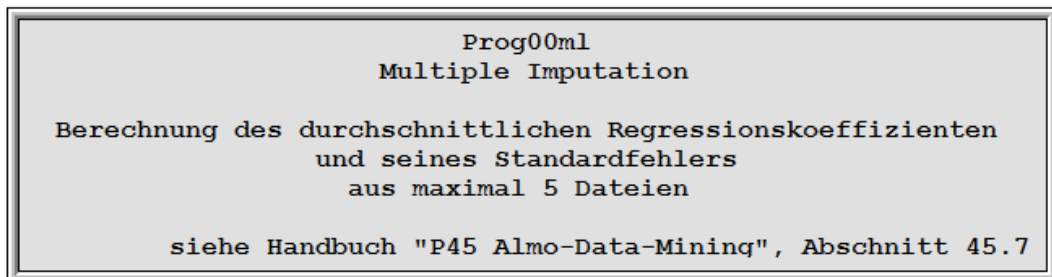
x2 0.8576 0.1055 0.8777 0.1098 0.7551 0.1191

4. Schritt: Wir berechnen "von Hand" die durchschnittlichen Regressionskoeffizienten und die Standardfehler gemäß oben angegebener Formel:

Variable	durchschnittlicher Regressionskoeff.	Standard fehler
x1	0.2125	0.14475
x2	0.8301	0.13499

Der Koeffizient für x1 hat einen 95%-Konfidenzbereich nach oben und unten von $2 \cdot 0.14475 = 0.2895$. Der Wert .0 wird also überschritten. Die Signifikanz $(1-p)100$ des Koeffizienten von x1 liegt demzufolge unter 95.5%.

Um das Rechnen "von Hand" zu erleichtern haben wir das kleine Maskenprogramm Prog00ml entwickelt, in das die Regressionskoeffizienten und ihre Standardfehler aus maximal 5 Dateien einzugeben sind. Also errechnet dann den mittleren Regressionskoeffizienten und seinen Standardfehler. Wir zeigen dieses Programm mit den Einträgen für die Variable x1.



P45.7.4 Ein Experiment

Der österreichische soziale Survey ist eine repräsentative Befragung der österreichischen Bevölkerung. Siehe Haller/Holm u.a. (1996). 2011 Befragte wurden durch eine Zufallsstichprobe ausgewählt. Aus den Daten dieses Surveys haben wir nun eine Unterdatei gebildet, die folgende Personen umfasst:

1. Personen, die voll berufstätig sind
2. ein eigenes Einkommen beziehen und dieses in der Befragung auch angegeben haben
3. eine Parteipräferenz angegeben haben

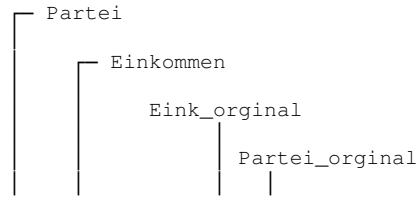
Von den sehr vielen Variablen des Surveys wurden folgende Variable in die Unterdatei übernommen:

```
Name 1=Bildung:Pflichts ohne Lehre,  
                Pflichts mit Lehre,  
                mittlere Schule,  
                Gymnasium,  
                Hochschule;  
Name 2=Beruf:  Bauern,  
                Selbständig,  
                Arbeiter,  
                Facharbeiter,  
                Angest/Beamte,  
                leitende Angest/Beamte;  
Name 3=FrauNichtBeruf:stimmt,stimmt nicht;  
Name 4=Todesstrafe:dafür,bedingt,dagegen;  
Name 5=Gewerkschaftsmitglied:ja,nein;  
Name 6=Parteimitglied:ja,nein;  
Name 7=Kirchgang:nie,  
                selten,  
                mehrmalJahr,  
                1malMonat,  
                2-3malMonat,  
                min1malWoche;  
Name 8=ZeitungLesen;  
Name 9=GastarbeiterRechte:zu wenig,ausreichend,zuviel;  
Name 10=Lebenszufriedenheit;  
Name 11=Partei:SPÖ,ÖVP,FPÖ,Grüne;  
Name 12=Einkommen;  
Name 13=Alter;  
Name 14=Geschlecht:m,w;  
Name 15=Eink_original;  
Name 16=Partei_original:SPÖ,ÖVP,FPÖ,Grüne;
```

Das Einkommen wurde in Intervallen von öS 2000.- gemessen und mit den Ziffern 1 bis 20 kodiert. Die ersten 4 Einkommensintervalle wurden aus der Unterdatei ausgeschlossen, da anzunehmen war, dass diese Personen nicht voll berufstätig waren.

Die Unterdatei ist unter dem Namen "SozSurv.fre" (im lesbaren freien Format) und "SozSurv.dir" (im nicht lesbaren direkten Format) im Almo-Ordner "Testdat" enthalten. Die oben angegebenen Dateinamen sind in der Datei "SozSurv.nam" ebenfalls im Ordner "Testdat" enthalten.

Die Datei „SozSurv.fre“ ist folgende



BefragterNr	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	V16
1	2	6	2	1	1	2	1	1	3	2	1	kw	43	1	9	1
2	5	6	2	2	1	2	6	1	2	2	3	kw	56	1	17	3
3	2	6	1	2	1	2	6	2	2	2	1	kw	50	2	13	1
.
.
.
229	2	5	2	3	1	1	1	1	1	1	2	kw	27	1	7	2
230	2	4	2	2	1	1	1	1	1	1	1	kw	24	1	10	1
231	4	5	2	3	1	2	1	1	2	1	1	kw	30	1	10	1

232	1	6	1	3	2	2	3	1	2	2	kw	6	40	2	6	1
233	5	6	2	3	1	2	1	1	1	3	kw	14	41	2	14	2
234	3	6	1	3	1	1	6	1	2	2	kw	10	56	1	10	2
.
.
.
463	4	6	2	3	2	2	2	1	1	3	kw	10	43	2	10	4
464	5	6	2	3	2	2	3	1	1	1	kw	11	38	1	11	4
465	2	5	2	2	1	2	1	2	2	2	kw	8	24	1	8	3

Die Datei umfasst insgesamt 465 Befragte. Die Aufeinanderfolge der Befragten ist zufällig. In V12 ist das Einkommen enthalten. Bei den Befragten 1 bis 231 (oberhalb des Trennstrichs) wurde es auf Kein-Wert (kurz: kw) gesetzt - zuvor jedoch in V15 gerettet. V15 erhält den Variablennamen "Eink_original". Für diese 1. Gruppe der Befragten wollen wir nun mit dem Programm Prog45mm die fehlenden Werte in V12 schätzen, indem wir für die 2. Gruppe (unterhalb des Trennstrichs) ein ALM rechnen. Da wir das tatsächliche Einkommen der 1. Gruppe in V15 gerettet haben, können wir nach der Schätzung den tatsächlichen Wert in V15 mit dem in V12 eingesetzten Schätzwert korrelieren. Auf diese Weise erhalten wir eine Information über die Güte der Schätzung. Als ursächliche Variable haben wir in Prog45mm eingesetzt:

Ursächliche Variable für die Zielvariable

ursächliche nominale Variable Hilfe

Hilfe

↔ ☐☐ **Beruf, Geschlecht**

↑↓ 0

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

ursächliche quantitative Variable Hilfe

Hilfe

↔ ☐☐ **Bildung, Alter**

ursächliche ordinale Variable Hilfe

Hilfe

↔ ☐☐ █

Das voll ausgefüllte Prog45mm ist unter dem Namen "KWEinset.Alm" als Beispielprogramm in Almo enthalten. Klicken Sie in der Knopfleiste auf den Knopf "Beispiel" und scrollen. Sie dann weit hinunter bis zur Überschrift "Daten-Imputation und Daten-Fusion".

Mit dem Beispielprogramm "KWEinKorr.Alm" wird dann die Korrelation zwischen dem in V12 eingesetztem Wert und dem tatsächlichen Wert in V15 berechnet. Wir erhalten

$$r = 0.6268$$

Das ist ein ordentlicher Wert.

Sie können mit dem Beispielprogramm "KWEinset.Alm" experimentieren. Wird beispielsweise als Kein-Wert-Behandlung bei der Berechnung der Regressionskoeffizienten und Effekte "1", das "paarweise Ausscheiden" eingesetzt, dann wird die Streuungs-Submatrix der ursächlichen Variablen aus allen 465 Befragten errechnet. Es entsteht dann eine geringfügig höhere Korrelation zwischen Schätzwert und tatsächlichem Wert von

$$r = 0.6273$$

Wird als Kein-Wert-Behandlung das "vollständige Ausscheiden" eingesetzt und bei der Prognosewert-Behandlung "6", die Überlagerung durch eine Zufallsvariation, dann entsteht eine deutlich schlechtere Korrelation von

$$r = 0.3754$$

Schätzung der fehlenden Werte in Variable V11 Partei

In V11 ist die Parteipräferenz enthalten. Dies ist eine nominal-polytome Variable. Bei den Befragten 232 bis 465 (unterhalb des Trennstrichs) wurde sie auf Kein-Wert (kurz: kw) gesetzt - zuvor jedoch in V16 gerettet. V16 erhält den Variablennamen "Partei_ordinal". Für diese 2. Gruppe der Befragten wollen wir nun mit dem Programm Prog45mz die fehlenden Werte in V11 schätzen, indem wir für die 1. Gruppe (oberhalb des Trennstrichs) eine Logitanalyse rechnen. Da wir die tatsächliche Parteipräferenz der 2. Gruppe in V16 gerettet haben, können wir nach der Schätzung den tatsächlichen Wert in V16 mit dem in V11 eingesetzten Schätzwert korrelieren. Auf diese Weise erhalten wir eine Information über die Güte der Schätzung. Als ursächliche Variable haben wir in Prog45mz eingesetzt:

Zielvariable

für die Prognosewerte für fehlende Werte eingesetzt werden sollen

↔ **Partei**

(nur eine) nominale Zielvariable

Ursächliche Variable für die Zielvariable

ursächliche nominale Variable

↔ **V2:6,9,14**

ursächliche quantitative Variable

↔ **Bildung,Kirchgang,Lebenszufriedenheit,Alter,Eink_ordinal**

Das voll ausgefüllte Prog45mz ist unter dem Namen "KWEinLog.Alm" als Beispielprogramm in Almo enthalten.

Mit dem Beispielprogramm "KWEinTab.Alm" wird dann die Korrelation zwischen dem in V11 eingesetztem Wert und dem tatsächlichen Wert in V16 berechnet. Wir erhalten folgende Tabelle und folgende Korrelationskoeffizienten:

	Parteiordinal				Summe
	SPÖ 1	ÖVP 2	FPÖ 3	Grüne 4	
Partei SPÖ	62	37	14	10	123
ÖVP	17	35	7	1	60
FPÖ	5	5	5	1	16
Grüne	6	9	2	10	27
Summe	90	86	28	22	226

Kontingenzkoeffizient C(cor) = 0.480
 Cramers V = 0.264

Unsere Schätzung wäre perfekt, wenn in obiger Tabelle nur die Diagonale besetzt wäre. Von den beiden Korrelationskoeffizienten, die Almo ausgibt, ist Cramers V vorzuziehen. Dies ist ein Koeffizient, der im Rahmen des ALM und der kanonischen Korrelationsanalyse aus „Pillais Spur“ entsteht. Siehe dazu Handbuch, Teil 3,

Abschnitt P19.2. Cramers V ist mit 0.264 kein befriedigender Wert. Die Kein-Wert-Einsetzung ist eher als gescheitert zu betrachten.

Mit dem Beispielprogramm "KWEinNom.Alm" kann man alternativ zum Logitmodell ein ALM für nominal-polytome Zielvariable rechnen (Einwände zu dieser Vorgehensweise siehe Abschnitt P45.15.1.0). Es liefert ein Cramers V von 0.227.

Kapitel 4: Dateien vereinen

Daß man verschiedene Dateien vereinen möchte, dürfte eher selten sein. Wenn es jedoch erforderlich ist, dann ist dies meist eine mühsame Arbeit. Almo bietet hier nun mehrere sehr komfortable Programme an, die nahezu jede Vereinigung von Dateien ermöglichen.

In Almo sind folgende Programme zum Vereinen von Dateien enthalten

1. Prog00md: Datensätze aus einer Datei A an eine Datei B anhängen
2. Prog00me: Zwei parallele Dateien zusammenfügen
3. Prog00mf+Prog00mg: Zwei Dateien über eine Verbindungs-Variable zusammenfügen
4. Prog45mw: Zwei verschiedene Dateien mit einigen gemeinsamen Variablen vereinen
5. DATBAN11.ALM Zwei Dateien vollständig verschmelzen

Die ersten 3 Programme findet man, wenn man auf den Knopf "Verfahren" klickt und dann "Datei-Operationen" selektiert. Das 4. Programm werden wir nachfolgend in Abschnitt P45.8.3 darstellen. Das 5. Programm findet man, wenn man auf den Knopf "Beispiel" klickt und dann (fast) bis zum Ende der Listbox scrollt. Eine ausführliche Beschreibung dieses in der Almo-Programmiersprache geschriebenen Programms ist im Handbuch, Teil 2, Abschnitt 46.9 enthalten.

Eine besondere (und auch umstrittene) Form des Vereinens von Dateien ist die "Datenfusion". Wir stellen sie im folgenden dar.

P45.8 Datenfusion

Von "Datenfusion" spricht man, wenn eine Datei A und eine Datei B, die verschiedene Personen enthalten und die einige Variable gemeinsam besitzen, zu einer einheitlichen Datei verknüpft werden – und wenn dabei Variable, die in A vorhanden sind, aber nicht in B, nach einem bestimmten Kalkül von A an B „gespendet“ werden.

Wir wollen diesen Begriff an einem Beispiel aus der Umfrageforschung erläutern. Dieses Beispiel werden wir auch im nachfolgenden Fusionsprogramm Prog45mw verwenden. Wir haben zwei Befragungen (mit verschiedenen Personen) durchgeführt, aus denen wir nun über eine Datei A und eine Datei B verfügen. Die beiden Dateien umfassen verschiedene Personen, denen zum Teil dieselben und zum Teil verschiedene Fragen gestellt worden sind. Betrachten wir die Variablen, die in den beiden Dateien enthalten sind.

Datei A

Name 1=Bildung:Pflichts ohne Lehre, Pflichts mit Lehre, mittlere Schule,
Gymnasium,Hochschule;
Name 2=Beruf:Bauern,
Selbständig,
Arbeiter,
Facharbeiter,
Angest/Beamte,
leitende Angest/Beamte;
Name 3=Einkommen;
Name 4=Alter;
Name 5=Geschlecht:m,w;
Name 6=Gewerkschaft:ja,nein;
Name 7=Kirchgang:1 mal pro Woche,
2-3 mal im Monat,
1 mal im Monat,
mehrmals im Jahr,
selten,
nie;
Name 8=GastarbeiterRechte:zu wenig,ausreichend,zuviel;
Name 9=Lebenszufriedenheit:sehr zufrieden,
ziemlich,
eher zufrieden,
eher unzufrieden,
ziemlich unzufrieden;
Name 10=FrauNichtBeruf:stimmt,stimmt nicht;
Name 11=Todesstrafe:dafür,bedingt dafür,dagegen;

Datei B

Name 1=Parteipraeferenz:SPÖ,ÖVP,FPÖ,Grüne,keine;
Name 2=Parteimitglied:ja,nein;
Name 3=Bildung:Pflichts ohne Lehre, Pflichts mit Lehre, mittlere Schule,
Gymnasium,Hochschule;
Name 4=Beruf:Bauern,
Selbständig,
Arbeiter,
Facharbeiter,
Angest/Beamte,
leitende Angest/Beamte;
Name 5=Alter;
Name 6=Geschlecht:m,w;
Name 7=ZeitungLesen:regelmässig,nicht regelmässig;
Name 8=Gewerkschaft:ja,nein;
Name 9=Kirchgang:1 mal pro Woche,
2-3 mal im Monat,
1 mal im Monat,
mehrmals im Jahr,
selten,
nie;
Name 10=GastarbeiterRechte:zu wenig,ausreichend,zuviel;

Offensichtlich sind in beiden Befragungen teilweise dieselben, teilweise aber auch verschiedene Variablen enthalten. So wurde etwa das Bildungsniveau in beiden Befragungen erkundet, während das Einkommen nur in der 1. Befragung erkundet wurde. Wir sortieren nun die Variablen nach (1) gemeinsamen Variablen, (2) zu "spendenden" Variablen und (3) nach spezifischen Variablen.

Datei A	Datei B	
V1 Bildungsniveau V2 Beruf V4 Alter V5 Geschlecht V6 Gewerkschaftsmitglied V7 Kirchengangshäufigkeit V8 Einstellung zu Gastarbeitern	V3 Bildungsniveau V4 Beruf V5 Alter V6 Geschlecht V8 Gewerkschaftsmitglied V9 Kirchengangshäufigkeit V10 Einstellung zu Gastarbeitern	gemeinsame Variable
V3 Einkommen --- V9 Lebenszufriedenheit	--- V1 Parteipräferenz ---	zu "spendende" Variable
V10 FrauNichtBerufstätig V11 Einstellung zu Todesstrafe --- ---	--- --- V2 Parteimitglied V7 ZeitungLesen	spezifische Variable

Datei A und Datei B besitzen mehrere gemeinsame Variable. Die Variable des Bildungsniveaus ist beispielsweise in beiden Dateien vorhanden.

Die Besonderheit ist nun folgende: Die Variablen "Einkommen" und "Lebenszufriedenheit" sind in Datei A vorhanden, nicht jedoch in Datei B. Wir haben das in Datei B durch 3 Striche markiert. Gesucht ist nun ein Verfahren, das es ermöglicht, diese beiden Variablen von Datei A an Datei B zu übertragen (zu "spenden"). Genauer formuliert: Die Personen in Datei B, die andere sind als in der Datei A, sollen Werte in diesen beiden Variablen zugewiesen bekommen, die von den Personen aus Datei A "gespendet" werden. Eine naheliegende Methode, die aber in Almo nicht verwendet wird, wäre, für jede Person in Datei B einen statistischen Zwilling in Datei A zu suchen, der dann seinen Wert in den Variablen "Einkommen" und "Lebenszufriedenheit" an die Person aus Datei B spendet.

Umgekehrt ist in Datei B die Variable "Parteipräferenz" vorhanden, die jedoch in Datei A fehlt. Wir haben das in Datei A durch 3 Striche markiert. Auch hier wünschen wir, daß diese Variable von Datei B an die Datei A "gespendet" werden könnte.

Mit dem Begriff "Datenfusion" ist diese einseitige oder auch gegenseitige "Spende" von Variablen gemeint.

Im 3. Teil der obigen Variablenliste sind die Variable angegeben, die jeweils nur in Datei A bzw. Datei B enthalten sind - die uns aber nicht weiter interessieren. Wir nennen sie "spezifische Variable".

So können wir nun 3 Arten von Variablen unterscheiden:

1. Die gemeinsamen Variablen. Sie sind in beiden Dateien enthalten.
2. Die zu "spendenden" Variablen. Das sind diejenigen, von denen wir wünschen, daß sie von der einen Datei an die andere "gespendet" werden.
3. Die spezifischen Variablen. Sie sind nur in einer Datei vorhanden – interessieren uns aber weiter nicht.

Unsere Beispieldaten

Unsere Beispieldaten sind empirische Daten. Sie sind Unterstichproben aus dem österreichischen sozialen Survey 1993 (Haller, Holm u.a., 1996). Die in Datei A fehlende Variable V1 Parteipräferenz und die in Datei B fehlenden Variablen V3 Einkommen und V9 Lebenszufriedenheit sind in Wirklichkeit vorhanden. Dadurch wird es möglich, zu überprüfen, wie gut diese Variablen von der einen Datei an die andere "gespendet" wurden.

Wir wollen das Ergebnis gleich vorwegnehmen: Die Korrelationen betragen

- 1) zwischen der "gespendeten" Variable des Einkommens und der wirklichen 0.519
- 2) zwischen der "gespendeten" Variable der Lebenszufriedenheit und der wirklichen 0.023
- 3) zwischen der "gespendeten" Variable der Parteipräferenz und der wirklichen 0.304

Weiter unten (bei Eingabe-Box 12 "Transformation der Prognosewerte für zu spendende Variable") werden wir zeigen, daß man den Wert der "gespendeten" Variablen mit einer Zufallsvariation überlagern kann. In Abschnitt P45.7.1.3 bei der Erläuterung zur Eingabe-Box 10 haben wir ausgeführt, daß dies sinnvoll ist. Wird diese Zufallsüberlagerung durchgeführt, dann verringert sich verständlicher Weise die Korrelation zu (1) auf 0.372. Die anderen beiden Korrelationen verändern sich nur minimal.

Die Koeffizienten zu (1) und (3) sind mit $(1-p)*100 = 99.9\%$ signifikant ($df=172$).

Der Koeffizient zu (2) ist nicht signifikant.

Die Datenfusion zu (3) ist mit dem Logitmodell gerechnet worden, da die "zu spendende" Variable nominal ist. Die beiden anderen wurden mit dem ALM gerechnet.

Die Korrelationskoeffizienten von (1) und (2) sind Produkt-Moment r .

Der Korrelationskoeffizienten bei (3) ist Cramer's V. Wird hier der (korrigierte)

Kontingenzkoeffizient gerechnet, dann entsteht sogar 0.5803.

Der Korrelationskoeffizient zwischen der "gespendeten" Variable der Lebenszufriedenheit und der wirklichen ist sehr niedrig. Die gespendete Variable ist unbrauchbar. Das werden wir auch erkennen, wenn wir nachfolgend das ALM rechnen, um diese Variable zu "spenden".

Die Korrelation zu (1) ist "ordentlich", die zu (3) "nicht gerade begeisternd".

Daten zum Experimentieren

Die Dateien A und B sind ein 2. Mal in Almo (im Verzeichnis "Testdat") enthalten. Ihre Namen lauten DatenA2.fre bzw. DatenA2.dir und DatenB2.fre bzw. DatenB2.dir. In DatenA2 ist die Parteipräferenz als V12 enthalten. In DatenB2 ist das Einkommen und die Lebenszufriedenheit als V11 und V12 enthalten. Der Benutzer kann mit diesen Dateien experimentieren und dabei überprüfen, wie gut die Variablenspende funktioniert hat. Wir stellen zu diesem Zweck die Beispielprogramme

QFusion.Alm
NFusion.Alm
QuantKor.Alm
NomKorr.Alm

zur Verfügung. Der Benutzer findet diese Programme, wenn er auf den Knopf "Beispiel" in der Knopfleiste klickt und in der dann erscheinenden Listbox sehr weit nach unten scrollt bis zu der Überschrift "Daten-Imputation und Daten-Fusion".

P45.8.1 Schritt 6a: Datenfusion mit dem Allgemeinen Linearen Modell

In Almo wird folgende Vorgehensweise für die Datenfusion gewählt:

Betrachten wir Datei A als "Spenderdatei", die an Datei B (die "Empfängerdatei") die "zu spendenden" Variablen Einkommen und Lebenszufriedenheit übergeben soll.

Mit Datei A wird ein Allgemeines Lineares Modell (ALM) gerechnet. Siehe dazu Abschnitt P45.15. Die Zielvariablen dieses ALM sind die zu spendenden Variablen Einkommen und Lebenszufriedenheit. Die ursächlichen Variablen sind die gemeinsamen Variablen, bzw. jene aus diesen, von denen wir annehmen, daß sie die Zielvariable signifikant determinieren. Man beachte: Das ALM wird nur mit den Personen der Spenderdatei A gerechnet.

Almo liefert uns aus dieser Analyse Regressionskoeffizienten und Effekte der ursächlichen Variablen. Almo ermittelt dabei folgende Gleichung (wir betrachten der Einfachheit halber nur das Einkommen als Zielvariable):

$$E = \beta_1 * B + \beta_2 * A + \dots + a_i + b_j + \text{const}$$

E = Einkommen (Prognosewert)
B = Bildungsniveau
 β_1 = Regressionskoeffizient für B
A = Alter
 β_2 = Regressionskoeffizient für A
 a_i = Effekte von Beruf
 b_j = Effekte von Geschlecht
const = Konstante

Da in der Empfängerdatei die ursächlichen Variablen Bildung, Alter etc. vorhanden sind, können wir diese Gleichung verwenden, um für jede Person aus der Empfängerdatei B einen Prognosewert hinsichtlich der (in Datei B ja nicht vorhandenen) Variablen des Einkommens und der Lebenszufriedenheit zu errechnen. Siehe dazu Abschnitt P45.17.

Die Prognosewerte dieser beiden Variablen werden in der Datei B bei jeder Person angehängt. Die Datei B ist damit um 2 Variable, die bei den Personen nicht unmittelbar erhoben wurden, vergrößert worden. Wir haben eine einseitige Datenfusion geleistet.

Diese Vorgehensweise ist auch möglich, wenn die "zu spendende Variable" nominal (dichotom oder polytom) ist. Gegen die Anwendung des ALM auf nominale Zielvariable sind verschiedene Einwände erhoben worden. Wir referieren sie ausführlich in Abschnitt P45.15.1.

Als Alternative wird das gewichtete ALM oder noch besser die Logitanalyse empfohlen. Diese Einwände sind u.E. nicht schwerwiegend, wenn es nicht darum geht, von Stichprobenergebnissen auf eine Grundgesamtheit zu schließen. Trotzdem werden wir mit Prog45my ein Programm anbieten, das die Datenfusion

mit Hilfe der Logitanalyse durchführt. Der Benutzer kann dann selbst entscheiden, ob er das ALM oder die Logitanalyse verwenden möchte.

Unsere Vorgehensweise ist sehr ähnlich derjenigen, die wir für das Einsetzen von Ersatzwerten bei fehlenden Werten gewählt haben. Tatsächlich sind das nachfolgende Programm Prog45mw zur Datenfusion und das in Abschnitt P45.7 dargestellte Programm Prog45mm zur Einsetzung von Prognosewerten für fehlende Werte nahezu identisch. Auch die auf der Logitanalyse beruhenden Programme Prog45my (für die Datenfusion) und Prog45mz (für die Kein-Wert-Einsetzung) entsprechen sich.

Natürlich können wir mit derselben Vorgehensweise, die in Datei A fehlende Variable der Parteipräferenz aus der Datei B als "Spenderdatei" gewinnen. Wenn wir die beiden Dateien A und B dann noch in einer gewissen Weise vereinigen, dann haben wir eine gegenseitige Datenfusion geleistet.

Der in Almo verwendete ALM-Ansatz der Datenfusion ist nicht der einzig mögliche. Gängig ist auch die Datenfusion über die Clusteranalyse. Zu einem späteren Zeitpunkt wird Johann Bacher ein derartiges Programm zur Verfügung stellen.

Um die Datenfusion ist teilweise heftig gestritten worden. Wir wollen dem Almo-Benutzer zumindest einen Hinweis geben: Die Determination der Zielvariablen (der zu "spendenden" Variablen) durch die ursächlichen Variablen (die gemeinsamen Variablen) ist an der multiplen Korrelation ablesbar. Siehe dazu Abschnitt P45.15.1.3. Ist diese schwach oder gar insignifikant, dann nähert sich der Prognosewert dem Mittelwert der Zielvariablen an. Die Datenfusion ist dann sinnlos.

P45.8.1.1 Eingabe in Programm Prog45mw zur einseitigen Datenfusion

Prog45mw leistet eine einseitige Datenfusion. Von einer Datei A (der "Spenderdatei") werden Variable an die Datei B (die "Empfängerdatei") übertragen.

Prog45mw.Msk

Datenfusion
mit Hilfe des Allgemeines Lineares Modells <ALM>

Die Datenfusion erfolgt in folgenden 3 Schritten:

Schritt 1:
Zuerst wird mit den Daten der Spenderdatei für die "zu spendende" Variable als Zielvariable ein ALM gerechnet. Als ursächliche Variable werden die Variablen verwendet, die die Spenderdatei gemeinsam mit der Empfängerdatei besitzt.

Schritt 2:
Dann werden mit Hilfe der dabei errechneten Koeffizienten aus den "gemeinsamen" Variablen als ursächlichen Variablen hinsichtlich der "zu spendenden" Variablen Prognosewerte für die Empfängerdatei ermittelt.

Schritt 3:
Diese Prognosewerte können dann noch durch Zufallsvariation verändert werden.
Die so ergänzte Empfängerdatei wird dann in eine neue Datei abgespeichert

siehe Handbuch "P45 Almo-Data-Mining", Abschnitt P45.8

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

Speicher fuer x Variable

1 Vereinbare Variable= ; Vereinbaren Sie mindestens so viele Variable, wie Spender- u. Empfängerdatei zusammen besitzen (+ Reserve)

2 Option: Weitere Vereinbarungen - nur wenn Almo dazu auffordert

Datei der Variablennamen aus Spenderdatei

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

Freie Namensfelder für Spenderdatei

4
 erzeuge zusätzliche Namensfelder

Spenderdatei

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

6

Empfängerdatei [Hilfe](#)

7

gemeinsame Variable aus Spender- und Empfängerdatei [Hilfe](#)

Variablennummern aus Spenderdatei

Variablennummern aus Empfängerdatei

8

Zu spendende Variable aus der Spenderdatei

BEACHTEN: Erlaubt sind:

1. Beliebige quantitative und/oder dichotome Variable
- oder (exklusiv)
2. Eine nominale Variable mit beliebig vielen Ausprägungen

(mehrere) quantitative/dichotome Variable

(nur eine) nominale Variable [Hilfe](#)

Zahl der zu spendenden Variablen

9

Ursächliche Variable

für die zu spendende Variable in der Spenderdatei

ursächliche nominale Variable [Hilfe](#)

[Hilfe](#)

Interaktionen x. Ordnung zwischen den ursächlichen nominalen Variablen bilden
oder einige ausgewählte Interaktionen bilden
0 = keine Interaktionen bilden [Hilfe](#)

ursächliche quantitative Variable [Hilfe](#)

ursächliche ordinale Variable [Hilfe](#)


10

Kein-Wert-Angaben und Umkodierungen

Kein-Wert-Angabe: Für zu spendende Variable
und ihre ursächlichen Variablen

Umkodierungen: Nur für die ursächlichen Variablen.
Zu spendende Variable darf nicht umkodiert werden
Umkodierungen sind temporär


Kein-Wert-Angabe
Umkodierungen




erzeuge zusätzliche Felder für Umkodierungen / Kein_Wert-Angaben

11

Kein-Wert-Behandlung der ursächlichen Variablen aus Spenderdatei

 **3** bei Berechnung der Regressionskoeffizienten und Effekte
möglich: 1 - 7; empfohlen: 3 =Vollständiges Ausscheiden
1 =Paarweises Ausscheiden

 **4** bei Berechnung der Prognosewerte für zu spendende Variable
möglich: 4 - 7; empfohlen: 4 =Mittelwert-Einsetzung

12

Transformation der Prognosewerte für zu spendende Variable

 **5** möglich: 4 - 7; empfohlen: 5 oder 7


13

Startwert für Zufallsgenerator
(für Verfahren 6 und 7)

 **123457**

14

Neue Empfängerdatei

 **"C:\Almo7\Progs\NeuDatB"**


Geben Sie den Dateinamen ohne Erweiterung
an. Almo erzeugt 2 Dateien:


1. eine nicht lesbare Almo-Arbeitsdatei
mit der Erweiterung **__.dir**
2. eine anschauliche Datei im freien Format
mit der Erweiterung **__.fre**


Ein Datensatz der neuen Empfängerdatei
enthält jetzt die Variablen der alten Datei
(im Beispiel sind das U1:U10)
plus
die gespendeten Variablen, die angehängt werden
(im Beispiel wurden 2 Variable gespendet,
die als U11 und U12 angehängt werden)

15

Ausgabe der Ergebnisse aus ALM

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

 **0** 1= Basisstatistiken ausgeben
0= nicht

 **0** Almo = Almo-Grafik ausgeben
0 = keine Grafik

P45.8.1.2 Erläuterungen zu den Eingabe-Boxen

Eingabe-Box 1 und **Eingabe-Box 2:**

Siehe "Arbeiten mit Almo-Datenanalyse-System", Abschnitt P0.1 und P0.2.

Eingabe-Box 3: Datei der Variablennamen aus Spenderdatei

Datei der Variablennamen aus Spenderdatei Hilfe

 zeige = Namensdatei in Output zeigen
leer = nicht

Geben Sie hier die Datei an, in die Sie die Variablennamen der Spenderdatei geschrieben haben. In unserem Beispiel sieht diese Datei folgendermaßen aus:

```
Name 1=Bildung:Volk ohne Lehre, Volk mit Lehre, mittlere Schule,  
          Gymnasium,Hochschule;  
Name 2=Beruf:Bauern,  
          Selbständig,  
          Arbeiter,  
          Facharbeiter,  
          Angest/Beamte,  
          leitende Angest/Beamte;  
Name 3=Einkommen;  
.  
.  
.
```

Die Variablennamen der Empfängerdatei werden nicht angegeben.

Wie Variablennamen geschrieben werden ist ausführlich in "Arbeiten mit Almo-Datenanalyse-System", Abschnitt P0.3 dargestellt.

Eingabe-Box 4: Freie Namensfelder für Spenderdatei

Zur Funktion dieser Eingabe-Box siehe "Arbeiten mit Almo-Datenanalyse-System", Abschnitt P0.3.

Beachte: Es werden keine Variablennamen für die Empfängerdatei angegeben.

Eingabe-Box 5: Spenderdatei

Eingabe-Box 6: Empfängerdatei

Spenderdatei Hilfe

Empfängerdatei Hilfe

Eingabefeld 1: Geben Sie zuerst den Dateinamen der Spenderdatei an. Die Datei muß im Format "direkt" vorliegen (mit der Erweiterung "xxx.dir"). Die in der Datei

enthaltenen Variablen müssen lückenlos fortlaufend die Variablennummern 1,2,3,4,..... besitzen.

Eingabefeld 2: Geben Sie an, wie viele Variable in der Datei enthalten sind. Dies muß auch gleichzeitig die Nummer der letzten Variablen sein.

Für die Empfängerdatei wird entsprechend verfahren.

Eingabe-Box 7: Gemeinsame Variable aus Spender- und Empfängerdatei

Am besten arbeitet man hier mit Bleistift und Papier. Zuerst schreibt man sich die Variablen aus der Spenderdatei auf, die diese gemeinsam mit der Empfängerdatei hat. Dann schreibt man die entsprechenden Variablennummern aus der Empfängerdatei daneben. Das sieht dann so aus:

gemeinsame Variable aus der Spenderdatei	Nummern der gemeinsamen Variablen aus der Empfängerdatei
-----	-----
V1 Bildungsniveau	V3
V2 Beruf	V4
V4 Alter	V5
V5 Geschlecht	V6
V6 Gewerkschaftsmitglied	V8
V7 Kirchengangshäufigkeit	V9
V8 Gastarbeitern	V10

Eingabefeld 1: Die in obiger Tabelle links stehenden Variablennummern aus der Spenderdatei werden in das 1. Eingabefeld geschrieben. Sie müssen in folgender Form geschrieben werden:

V1,2,4,5,6,7,8

Vorne steht 'V' darauf folgen die Nummern - getrennt durch Beistrich.

Eingabefeld 2: Die oben rechts stehenden Variablennummern aus der Empfängerdatei werden in das 2. Eingabefeld geschrieben. Sie müssen in folgender Form geschrieben werden:

V3,4,5,6,8,9,10

Vorne steht 'V' darauf folgen die Nummern - getrennt durch Beistrich.

Am besten schreibt man die beiden Nummernreihen in den beiden Eingabefeldern exakt untereinander, so daß man optisch kontrollieren kann, welche Variable aus der Spenderdatei welcher Variablen aus der Empfängerdatei entspricht.

Beachte: Es genügt jene Variablen als gemeinsame in dieser Eingabe-Box zu deklarieren, die in der übernächsten Eingabe-Box 9 als ursächliche Variable eingetragen werden. Anders formuliert: Variable, die in Eingabe-Box 9 nicht als

ursächliche Variable verwendet werden, müssen nicht in dieser Eingabe-Box 7 als gemeinsame deklariert werden.

Eingabefeld 3: Geben Sie an, wie viele Variable Sie im 1. Eingabefeld (bzw. im 2.) geschrieben haben.

Eingabe-Box 8: Zu spendende Variable aus der Spenderdatei

Zu spendende Variable aus der Spenderdatei

BEACHTEN: Erlaubt sind:

1. Beliebige viele quantitative und/oder dichotome Variable

oder (exklusiv)

2. Eine nominale Variable mit beliebig vielen Ausprägungen

(mehrere) quantitative/dichotome Variable

Einkommen, Lebenszufriedenheit

(nur eine) nominale Variable Hilfe

2 Zahl der zu spendenden Variablen

In dieser Eingabe-Box sind jene Variable einzutragen, die von der Spenderdatei an die Empfängerdatei übertragen werden sollen.

Eingabefeld 1: Es können beliebig viele quantitative und dichotome Variable eingegeben werden.

Eingabefeld 2: Es kann nur 1 nominale Variable (mit beliebig vielen Ausprägungen) angegeben werden.

BEACHTEN: Es darf nur 1 Eingabefeld benutzt werden. D.h. erlaubt sind

1. beliebig viele quantitativen und dichotome Variable oder (exklusiv)
2. eine nominale Variable

Ordinale Variable können nicht als Zielvariable angegeben werden, da es problematisch ist, Prognosewerte für ordinale Zielvariable mit dem ALM zu berechnen.

Dichotome Variable werden im ALM (als Zielvariable) behandelt wie quantitative Variable. Das ist der Grund dafür, daß sie auch im 1. Eingabefeld eingegeben

werden können. Das hat den Vorteil, daß dann beliebig viele "zu spendende" Variable angegeben werden können. Wird eine dichotome Variable im 2. Eingabefeld (als nominale Variable) eingesetzt - was selbstverständlich auch korrekt ist - dann ist nur diese eine als "zu spendende" Variable möglich. Im 2. Eingabefeld darf nur 1 Variable eingetragen werden (und im 1. dann überhaupt keine).

Werden dichotome Variable im 1. Eingabefeld eingetragen, dann muß man in der Eingabe-Box 12 "Transformation der Prognosewerte" entweder "5" oder "7" als Prognosewert-Behandlung einsetzen. Dann entsteht als Wert der "zu spendenden" Variablen einer der beiden empirisch vorkommenden (in der Regel ganzzahligen) Werte - und nicht ein Wert, dem keine der beiden empirischen Ausprägungen der dichotomen Variablen entspricht. Wird in Eingabe-Box 12 "5" eingesetzt, dann entsteht für die dichotome "zu spendende" Variable derselbe Wert, egal ob sie im 1. oder 2. Eingabefeld der Eingabe-Box 8 eingetragen wurde. Wird "7" eingesetzt, dann entstehen etwas verschiedene Werte, was durch die unterschiedliche Hinzufügung einer Zufallsvariation verursacht wird.

Eingabe-Box 9: Ursächliche Variable für die zu spendende Variable in der Spenderdatei

Ursächliche Variable
für die zu spendende Variable in der Spenderdatei

ursächliche nominale Variable

Beruf, Geschlecht

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

ursächliche quantitative Variable

Bildung, Alter, Kirchengang

ursächliche ordinale Variable

Geben Sie in dieser Eingabe-Box jene Variable aus der Spenderdatei an, von denen Sie vermuten, daß sie die "zu spendenden Variablen" am besten determinieren. Wenn Sie in der Eingabe-Box 8 "zu spendende Variablen" mehrere quantitative und dichotome Variable angebegeben haben, dann berechnet Almo für jede dieser Zielvariablen getrennt, die Regressionskoeffizienten und Effekte der ursächlichen Variablen. Es ist nicht möglich für jede einzelne der "zu spendenden Variablen" einen eigenen Satz von ursächlichen Variablen anzugeben. Die ursächlichen Variablen gelten gemeinsam für alle "zu spendenden Variablen".

Als ursächliche Variable können nur Variable verwendet werden, die oben in Eingabe-Box 7: "Gemeinsame Variable" im 1. Eingabefeld als "gemeinsame Variable" deklariert wurden.

Zum Begriff der "ursächlichen" Variablen:

Dieser Begriff darf nicht wörtlich genommen werden. Gemeint sind Variable, die dazu verwendet werden können, die Zielvariable zu determinieren. Das können durchaus "ursächliche" sein, d.h. Variable, die (in unserem Beispiel) Ursachen des Einkommens sind (wie etwa das Bildungsniveau); aber auch "Folgevariable", wie z.B. der Besitz von mehr oder weniger hochwertigen Konsumgütern. So ist möglicherweise der Autobesitz (Kleinwagen, Mittelklassewagen, Luxuswagen) eine sichtbare Folge der Einkommensverhältnisse - also eine Variable, die als Determinante des Einkommens mitverwendet werden kann.

Eingabefeld 1: Geben Sie hier die nominalen ursächlichen Variablen an. Zu den Messniveaus "nominal", "ordinal" und "quantitativ" siehe P45.12.0.

Eingabefeld 2: Wenn Sie Interaktionen zwischen den ursächlichen nominalen Variablen miteinbeziehen wollen, dann geben Sie hier die Interaktionsordnung an. Siehe dazu die ausführliche Erläuterung für Prog45mf in P45.15.1.2, Erläuterung zu Eingabe-Box 6 "Ursächliche Variable".

Eingabefeld 3: Geben Sie hier die quantitativen ursächlichen Variablen an. Siehe dazu die ausführliche Erläuterung für Prog45mf ebenfalls in P45.15.1.2.

Eingabefeld 4: Geben Sie hier die ordinalen ursächlichen Variablen an. Siehe dazu die ausführliche Erläuterung für Prog45mf ebenfalls in P45.15.1.2.

Eingabe-Box 10: Kein-Wert-Angabe und Umkodierungen

Kein-Wert-Angaben und Umkodierungen

Kein-Wert-Angabe: Für zu spendende Variable
und ihre ursächlichen Variablen

Umkodierungen: Nur für die ursächlichen Variablen.
Zu spendende Variable darf nicht umkodiert werden
Umkodierungen sind temporär

Kein-Wert-Angabe

Umkodierungen

erzeuge zusätzliche Felder für Umkodierungen / Kein_Wert-Angaben

Kein-Wert-Angaben:

Almo muß selbstverständlich wissen, an welchen Codeziffern es den Kein-Wert-Fall in den in Eingabe-Box 8 und 9 angegebenen Variablen aus der Spenderdatei erkennen kann. Hier gibt es 2 Vorgehensweisen:

1. Der Benutzer hat schon eine Almo-Arbeitsdatei (im Format DIREKT) erstellt. Dabei hat er Almo mitgeteilt, welche Codeziffern den Kein-Wert-Fall bezeichnen.

Almo hat dann die Almo-Arbeitsdatei erzeugt und dabei die vom Benutzer definierten Kein-Wert-Codeziffern (beispielsweise die 0) durch einen Almo-internen Kein-Wert-Code ersetzt. In diesem Fall ist jetzt eine Kein-Wert-Angabe nicht mehr notwendig.

Diese Vorgehensweise haben wir zur Erzeugung der Almo-Arbeitsdatei in den Programmen Prog45md und Prog45mh in Abschnitt P45.1 und P45.2 gewählt.

2. Der Benutzer hat eine Kein-Wert-Deklaration noch nicht vorgenommen. In der Arbeitsdatei stehen also noch die ursprünglichen Codes (z.B. 0 für Kein-Wert). In diesem Fall muß der Benutzer jetzt eine Kein-Wert-Angabe vornehmen - beispielsweise so:

```
Beruf, Wohnort ( 0 = Kein_Wert )
Einkommen      ( -1 = Kein_Wert )
```

Umkodierungen:

In der Eingabe-Box "Kein-Wert-Angabe und Umkodierungen" können auch Variable umkodiert werden.

- a. Es dürfen nur die ursächlichen Variablen umkodiert werden - nicht die "zu spendende Variable"
- b. Diese Umkodierungen sind temporär. Sie wirken nur während der Berechnung der Prognosewerte. D.h. die eventuell umkodierten ursächlichen Variablen gehen in ihrer ursprünglichen Form in die neue Empfängerdatei ein (siehe Eingabe-Box 14).

Wie Kein-Wert-Angaben und wie Umkodierungen zu erzeugen bzw. zu schreiben sind ist ausführlich in Abschnitt P0.5 beschrieben worden.

Eingabe-Box 11: Kein-Wert-Behandlung der ursächlichen Variablen aus Spenderdatei

Kein-Wert-Behandlung

Kein-Wert-Behandlung der ursächlichen Variablen in Spenderdatei

Methoden der Kein-Wert-Behandlung ---> Hilfe
Hilfe

↑ ↓ 3 bei Berechnung der Regressionskoeffizienten und Effekte
möglich: 1 - 7; empfohlen: 3 =Vollständiges Ausscheiden
1 =Paarweises Ausscheiden

Kein-Wert-Behandlung der ursächlichen Variablen in Empfängerdatei

↑ ↓ 4 bei Berechnung der Prognosewerte für zu spendende Variable
möglich: 4 - 7; empfohlen: 4 =Mittelwert-Einsetzung

Eingabefeld 1: Um Prognosewerte für die "zu spendende" Variable zu bilden, rechnet Almo ein ALM für die Zielvariable (d.h. die "zu spendende" Variable) und die ursächlichen Variablen. Diese Variable können fehlende Werte aufweisen. Wie soll hier verfahren werden?

Almo geht folgendermaßen vor: Besitzt ein Datensatz in der Zielvariablen keinen Wert, dann wird er aus der Analyse ausgeschlossen. Fehlen in den ursächlichen

Variablen Werte, dann kann eine von 7 Kein-Wert-Behandlungen verwendet werden.

Die 7 Kein-Wert-Behandlungen sind ausführlich im Abschnitt P45.7.3 über die Einsetzung fehlender Wert mit Hilfe des ALM, bei der Erläuterung der Eingabe-Box 9 "Kein-Wert-Behandlung der ursächlichen Variablen" beschrieben. Der Benutzer lese unsere Ausführungen in diesem Abschnitt, insbesondere unsere Empfehlungen zur Wahl einer Kein-Wert-Behandlung.

Wir wollen hier unsere Empfehlung nochmals kurz wiederholen:

Wenn nur 1 Zielvariable (d.h. 1 "zu spendende" Variable) vorhanden ist, dann sollte man die Kein-Wert-Behandlung 3, das "vollständige Ausscheiden" wählen. Das ist die klarste und beste Lösung des Kein-Wert-Problems. Ein Datensatz wird ausgeschlossen, wenn er auch nur in einer Analyse-Variablen keinen Wert besitzt.

Sind 2 oder mehrere Zielvariable vorhanden, dann addieren sich die Kein-Wert-Fälle zu einer höheren Gesamtzahl der Ausfälle, wie wenn nur 1 Zielvariable vorhanden ist. Ist diese Erhöhung gering, dann kann man die Kein-Wert-Behandlung 3, das "vollständige Ausscheiden" beibehalten. Wenn nicht, dann sollte man die Kein-Wert-Behandlung 1, das "paarweise Ausscheiden" wählen oder sich dazu entschließen nur eine Zielvariable zu verwenden.

Eingabefeld 2: Mit Hilfe der durch das ALM errechneten Koeffizienten werden aus den "gemeinsamen" Variablen als ursächlichen Variablen hinsichtlich der "zu spendenden" Variablen Prognosewerte für die Empfängerdatei ermittelt.

Bei der Berechnung der Prognosewerte stellt sich nun dasselbe Problem. Wie soll verfahren werden, wenn eine der ursächlichen Variablen keinen Wert besitzt ? Das "vollständige Ausscheiden" eines Datensatzes, wenn auch nur eine ursächliche Variable keinen Wert besitzt, ist hier nicht möglich, da wir ja dann unsere Aufgabe, Prognosewerte für die Empfängerdatei zu erzeugen, nicht erfüllen könnten. Hier sind deswegen nur die "Kein-Wert-Behandlungs-Methoden" 4 bis 7 möglich, bei denen der Mittelwert (bzw. Median, bzw. Erwartungswert) der ursächlichen Variablen (eventuell mit einer "Zufalls-Überlagerung) eingesetzt wird.

Wenn der Benutzer auf den nachfolgenden Hilfeknopf in der Eingabe-Box klickt, dann werden ihm diese Methoden gezeigt.

Eingabe-Box 12: Transformation der Prognosewerte für zu spendende Variable

Eingabe-Box 13: Startwert für Zufallsgenerator

Transformation der Prognosewerte für zu spendende Variable

Methoden der Transformation --->

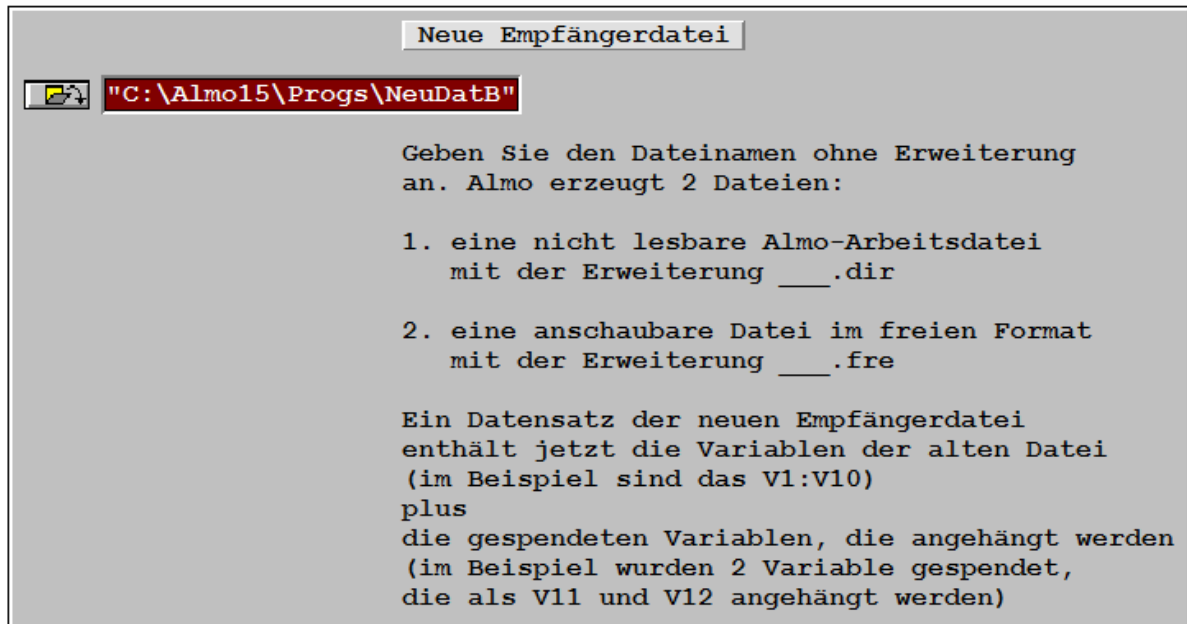
möglich: 4 - 7; empfohlen: 5 oder 7

Startwert für Zufallsgenerator
(für Verfahren 6 und 7)

Diese 2 Eingabe-Boxen entsprechen den bereits in Abschnitt P45.7.1.3 dargestellten Eingabe-Boxen 10 und 11, so daß wir hier nur eine kurze Ergänzung vornehmen wollen.

Zu Eingabe-Box 12 "Transformation der Prognosewerte" ist folgendes anzumerken: Ist die zu spendende Variable eine dichotome, dann muß als Prognosewert-Behandlung 5 oder 7 eingesetzt werden. Wir haben darauf bereits oben bei Eingabe-Box 8: "Zu spendende Variable" hingewiesen.

Eingabe-Box 14: Neue Empfängerdatei



Neue Empfängerdatei

"C:\Almo15\Progs\NeuDatB"

Geben Sie den Dateinamen ohne Erweiterung an. Almo erzeugt 2 Dateien:

1. eine nicht lesbare Almo-Arbeitsdatei mit der Erweiterung ____.dir
2. eine anschaulbare Datei im freien Format mit der Erweiterung ____.fre

Ein Datensatz der neuen Empfängerdatei enthält jetzt die Variablen der alten Datei (im Beispiel sind das V1:V10) plus die gespendeten Variablen, die angehängt werden (im Beispiel wurden 2 Variable gespendet, die als V11 und V12 angehängt werden)

Eingabefeld 1: Geben Sie den Dateinamen für die Datei an, in die Sie die um die zu spendenden

Variablen vergrößerte Empfängerdatei speichern wollen. Almo schreibt dabei für die in der Eingabe-Box 8 als "zu spendenden Variable" angegebenen Variablen die errechneten Einsetzungswerte. Die anderen Variablen werden mit ihren ursprünglichen Werten übernommen.

Geben Sie dabei den Dateinamen ohne Erweiterung an. Almo erzeugt dann 2 Dateien:

1. eine nicht lesbare Almo-Arbeitsdatei mit der Erweiterung ____.dir
Im Beispiel: "C:\Almo6\PROGS\NeuDatenB.dir"
2. eine anschaulbare Datei im freien Format mit der Erweiterung ____.fre
Im Beispiel: "C:\Almo6\PROGS\NeuDatenB.fre"

In der "anschaulbaren" Datei können Sie sich nochmals die von Almo errechneten Einsetzungswerte anschauen.

Eingabe-Box 15: Ausgabe der Ergebnisse aus ALM

Ausgabe der Ergebnisse aus ALM		
<input type="checkbox"/> <input checked="" type="checkbox"/> <input type="checkbox"/>	1	0= Ergebnisse in voller Länge ausgeben 1= Ergebnisse etwas verkürzt ausgeben 2= Ergebnisse stark verkürzt ausgeben
<input type="checkbox"/> <input checked="" type="checkbox"/> <input type="checkbox"/>	0	1= Basisstatistiken ausgeben 0= nicht
<input type="checkbox"/> <input checked="" type="checkbox"/> <input type="checkbox"/>	0	Almo = Almo-Grafik ausgeben 0 = keine Grafik
<input type="checkbox"/> <input checked="" type="checkbox"/> <input type="checkbox"/>	0	1 = Almo-Grafiken in Ergebnisliste einsetzen 0 = nicht
		<input type="button" value="Hilfe"/>

Eingabefeld 1: Almo rechnet, wenn mehrere "zu spendende" Variable vorhanden sind, eine multivariate Analyse und gibt, wenn auf "2" (stark verkürzte Ausgabe) eingestellt wurde nur eine zusammenfassende Ergebnistabelle aus, die keine Information über die einzelnen "zu spendenden" Variablen enthält.

Wir empfehlen deswegen auf "1" (etwas verkürzte Ausgabe) zu stellen, wenn mehrere quantitative (oder dichotome) "zu spendende" Variable vorhanden sind. Ist nur 1 vorhanden, dann sollte man auf "2" einstellen.

P45.8.1.3 Ausgabe aus Prog45mw

Almo gibt zuerst die Ergebnisse aus dem Allgemeinen Linearen Modell aus. Wir werden die einzelnen Teile der Ausgabe aus dem ALM nur insofern erläutern, als sie für

unser Thema der Datenfusion von Belang sind. In Eingabe-Box 15 "Ausgabe der Ergebnisse aus ALM" wurde auf 1 (=Ergebnisse etwas verkürzt) eingestellt. Die Ausgabe ist anders wenn in Eingabe-Box 15 auf 0 oder 2 eingestellt wird. Siehe dazu unsere ausführliche Darstellung in Abschnitt P45.15.1.3 (stark verkürzte Ausgabe), P45.15.1.4 (etwas verkürzte Ausgabe), P45.15.1.5 (volle Ausgabe).

Zahl der insgesamt eingelesenen Einheiten 378
Zahl der in die Analyse einbezogenen Einheiten 144

***** Erläuterung:

Die Zahl 378 ist die Summe der Personen aus Spender- und Empfängerdatei. 144 ist die Zahl der Personen in der Spenderdatei, d.h. die Zahl der Personen, die zur Errechnung der Regressionskoeffizienten und Effekte verwendet wurden.

Koeffizienten fuer quantitat./ordinale Variable aus univariater Analyse

hinsichtlich der abhaeng. Var. V3 Einkommen

Variable	Regr. koeff.	part. Korrel.	Signifikanz p	(1-p)100
V1 Bildung	0.7479	0.2671	0.0018	99.82
V4 Alter	-0.0389	-0.1759	0.0404	95.96
V7 Kirchgang	-0.3068	-0.2076	0.0153	98.47

hinsichtlich der abhaeng. Var. V9 Lebenszufriedenh

V1 Bildung	0.0311	0.0384	0.6571	34.29
V4 Alter	0.0018	0.0282	0.7453	25.47
V7 Kirchgang	-0.0046	-0.0107	0.9004	9.96

******* Erläuterung:**

Wir erkennen, daß die 3 quantitativen Variablen die Zielvariablen des Einkommens signifikant determinieren. Hingegen wirken sie hinsichtlich der Zielvariablen der Lebenszufriedenheit insignifikant. Man sollte sich in dieser Situation entscheiden, die Analyse hinsichtlich der Lebenszufriedenheit ohne diese 3 ursächlichen Variablen zu wiederholen. Da aber, wie weiter unten zu sehen ist, auch die nominalen ursächlichen Variablen hinsichtlich der Lebenszufriedenheit keine signifikanten Determinanten sind, sollte man darauf verzichten, die Lebenszufriedenheit als "zu spendende Variable" von der Spenderdatei auf die Empfängerdatei zu übertragen.

"multivariate" partielle Korrelation zwischen der Menge der abhaengigen Variablen und den einzelnen unabhaengigen quantitat./ordinalen Variablen

Variable	part. Korrel	Signifikanz p	(1-p)100
V1 Bildung	0.2755	0.006	99.44
V4 Alter	0.1760	0.121	87.88
V7 Kirchgang	0.2106	0.048	95.24

******* Erläuterung:**

Da wir 2 abhängige Variable haben, rechnet Almo eine multivariate Analyse. Diese ist für die Datenfusion irrelevant.

Koeffizienten der Dummies
hinsichtlich der abh. Var. V3 Einkommen

Effekte von A Beruf

	Effekte	partielle Korrelat.	Signifikanz p	(1-p)100
A1 Bauern	-1.3172	-0.0882	0.3073	69.27%
A2 Selbstä	0.7895	0.0740	0.3918	60.82%
A3 Arbeite	0.2939	0.0492	0.5688	43.12%
A4 Facharb	-0.6828	-0.0924	0.2847	71.53%
A5 Angest/	-0.5318	-0.0716	0.4077	59.23%
A6 leitend	1.4483	0.2468	0.0039	99.61%

Effekte von B Geschlecht

	Effekte	partielle Korrelat.	Signifikanz p	(1-p)100
B1 m	0.6335	0.2357	0.0056	99.44%
B2 w	-0.6335	-0.2357	0.0056	99.44%

******* Erläuterung:**

Das Geschlecht ist eine hoch signifikante Determinante des Einkommens. Beim Beruf ist lediglich bei der Dummy-Variablen A6, den leitenden Angestellten/Beamten eine Signifikanz feststellbar. Daß bei A4, den Facharbeiter, mit -0.6828 ein (im Vergleich zum Durchschnitt aus allen 144 Berufsausübenden) negativer Effekt hinsichtlich des Einkommens, bei A3, den Arbeitern, hingegen ein positiver Effekt vorhanden ist, ist nicht plausibel. Die Effekte sind ohnehin nicht

signifikant. Es wäre zu überlegen, ob man die Analyse für die Zielvariable des Einkommens ohne Beruf als ursächliche Variable wiederholt.

Koeffizienten der Dummies
hinsichtlich der abh. Var. V9 Lebenszufriedenh

Effekte von A Beruf

	Effekte	partielle	Signifikanz	
	Korrelat.	p	(1-p)100	
A1 Bauern	-0.4127	-0.0920	0.2865	71.35%
A2 Selbstä	0.5485	0.1693	0.0486	95.14%
A3 Arbeite	0.1005	0.0561	0.5161	48.39%
A4 Facharb	0.0345	0.0156	0.8573	14.27%
A5 Angest/	-0.1851	-0.0829	0.3370	66.30%
A6 leitend	-0.0858	-0.0502	0.5615	43.85%

Effekte von B Geschlecht

	Effekte	partielle	Signifikanz	
	Korrelat.	p	(1-p)100	
B1 m	0.0345	0.0439	0.6113	38.87%
B2 w	-0.0345	-0.0439	0.6113	38.87%

******* Erläuterung:**

Die Determination der Lebenszufriedenheit durch Beruf und Geschlecht ist nicht signifikant (mit der Ausnahme, daß bei A2, den Selbständigen ein signifikanter Effekt vorhanden ist). Da aber, wie wir oben schon gesehen haben, auch die quantitativen ursächlichen Variablen nicht signifikant wirken, sollte man darauf verzichten, die Lebenszufriedenheit als "zu spendende Variable" von der Spenderdatei auf die Empfängerdatei zu übertragen.

Multiple Korrelation aus univariater Analyse
hinsichtlich der abhaengigen Variablen V3 Einkommen

Fehlerstreuung	846.139609
Durch alle unabhaeng. Variablen erklärte Streuung	339.860391
Multiples Bestimmtheitsmass	0.286560
Multiple Korrelation	0.535313
F-Wert f. erklarte Streuung	5.980270
Freiheitsgrade Nenner = 9	
Zaehler= 134	
Signifikanz: p	0.000014
Signifikanz: (1-p)*100	99.998561 %
Teststaerke von F	0.999905

Multiple Korrelation aus univariater Analyse
hinsichtlich der abhaengigen Variablen V9 Lebenszufriedenh

Fehlerstreuung	76.220244
Durch alle unabhaeng. Variablen erklärte Streuung	3.751978
Multiples Bestimmtheitsmass	0.046916
Multiple Korrelation	0.216601
F-Wert f. erklarte Streuung	0.732913
Freiheitsgrade Nenner = 9	
Zaehler= 134	
Signifikanz: p	0.679626
Signifikanz: (1-p)*100	32.037380 %
Teststaerke von F	0.350438

******* Erläuterung:**

Die multiple Korrelation und ihre Signifikanz sind sehr wichtige Koeffizienten, die es erlauben abzuschätzen, ob die "Variablenspende" überhaupt einen Sinn hat. Die multiple Korrelation hinsichtlich des Einkommens ist 0.535, die Signifikanz 99.99 %. Das sind "ordentliche" Werte. Sie bedeuten, daß es gelungen ist, die für das Einkommen (der Personen der Spenderdatei) relevanten ursächlichen Variablen in die Analyse einzuführen.

Die multiple Korrelation hinsichtlich der Lebenszufriedenheit ist 0.217, die Signifikanz 32 %. Das sind sehr schlechte Werte. Sie bedeuten, daß es nicht gelungen ist, die für die Lebenszufriedenheit (der Personen der Spenderdatei) relevanten ursächlichen Variablen in die Analyse einzuführen, bzw. daß diese in der Datei gar nicht vorhanden sind. Es ist dann sinnlos, für die Personen der Empfängerdatei Prognosewerte für die Lebenszufriedenheit zu errechnen.

Die multiple Korrelation wird in obiger Form ausgegeben, wenn 2 oder mehrere Zielvariable angegeben wurden. Wird nur 1 Zielvariable angegeben, dann ist die Ausgabe etwas anders. Wir wollen annehmen, wir hätten nur das Einkommen als "zu spendende" Variable in Eingabe-Box 8 eingesetzt. Also würde dann an Stelle der obigen Ausgabe der multiple Korrelation folgende zusammenfassende Tabelle bringen:

Zusammenfassung

Streuungsquelle	Streuung	Korrel Koeff.	F-Wert	df	Signifikanz p	(1-p)100
Gesamtstreuung	1186.1103					
Fehlerstreuung	846.3134			135		
alle unabh. Var. zusammen	339.7970	0.5352	6.0225	9	0.0000	99.9986
quant./ordin. Var. zusammen	147.6438	0.3854	7.8505	3	0.0002	99.9801
nominale Variable zusammen	106.8443	0.3348	2.8406	6	0.0123	98.7651
V1 Bildung	66.1153	0.2692	10.5464	1	0.0016	99.8395
V4 Alter	26.8537	-0.1754	4.2836	1	0.0402	95.9764
V7 Kirchengang	38.1052	-0.2076	6.0784	1	0.0149	98.5085
V2 Beruf	81.6007	0.2965	2.6033	5	0.0275	97.2528
V5 Geschlecht	50.5979	0.2375	8.0711	1	0.0051	99.4887

Die multiple Korrelation und ihre Signifikanz finden wir in der Zeile "alle unabh. Var. zusammen". Sie ist 0.535. Die Signifikanz ist 99.99 %.

Berechnung der Prognosewerte bzw. Schätzwerte fuer zu "spendende" Variable

***** MITTEILUNG
 Sind unabhaengige Variable, die für die Berechnung
 des Prognosewerts benoetigt werden, gleich "Kein_Wert"
 dann wird fuer sie "Kein-Wert-Behandlung = 4" durchgefuehrt

Mittelwert und Standardabweichung der Residuen

	Mittelwert	Standardabweichung
V3 Einkommen	-0.00292378	2.41592
V9 Lebenszufrie	0.023612	0.785067

******* Erläuterung:**

Die Standardabweichungen der Residuen für die zu "spendenden" Variablen werden für die Prognosewert-Behandlung 6 und 7 verwendet, sofern der Benutzer diese in der Eingabe-Box 12 "Transformation der Prognosewerte für zu spendende Variable" eingesetzt hat. Dadurch wird für die Personen der Empfängerdatei eine

normalverteilte Zufallvariation des Prognosewertes mit der oben angegebenen Standardabweichung vorgenommen.

Ursachliche Variable, die Kein_Wert waren
und durch Schaetzwerte ersetzt wurden

V2	Beruf	in	0 Datensätzen
V5	Geschlecht	in	0 Datensätzen
V1	Bildung	in	0 Datensätzen
V4	Alter	in	0 Datensätzen
V7	Kirchgang	in	4 Datensätzen

******* Erläuterung:**

Beispiel: Die ursächliche Variable "Kirchgang" besaß in 4 Datensätzen keinen Wert usw. Diese Angaben beziehen sich auf die Personen aus der Spenderdatei. Da wir in Eingabe-Box 11 "Kein-Wert-Behandlung der ursächlichen Variablen aus Spenderdatei" als "Kein-Wert-Behandlung = 4" eingesetzt hatten, wird in diesen Variablen der Mittelwert (bei quantitativen Variablen) bzw. der Erwartungswert (bei nominalen Variablen) als Ersatzwert eingesetzt.

In "Empfaengerdatei" fuer "gespendete" Variable eingesetzter Schaetzwert

***** MITTEILUNG
Der fuer die "gespendete" Variable eingesetzte Schaetzwert
wird noch der "Prognosewert-Behandlung = 7" unterworfen

└─ Die Variablennummer ist die aus der Spenderdatei

Datensatz	Variable	Prognosewert	eingesetzter Wert
1	V3 Einkomme	6.2868	5
1	V9 Lebenszu	1.83824	3
2	V3 Einkomme	8.14958	5
2	V9 Lebenszu	1.90065	1
3	V3 Einkomme	7.43636	8
3	V9 Lebenszu	1.68287	1
4	V3 Einkomme	7.99913	5
4	V9 Lebenszu	1.69945	1
.		.	.
.		.	.
.		.	.
.		.	.
191	V3 Einkomme	8.4568	11
191	V9 Lebenszu	1.70106	1
192	V3 Einkomme	9.58078	10
192	V9 Lebenszu	1.82415	3

***** **Erläuterung:**

Almo teilt die Werte mit, die in der Empfängerdatei für die beiden zu "spendenden" Variablen eingesetzt werden.

***** MITTEILUNG
Lesen oder Schreiben korrekt beendet in Datei
"C:\Almo6\Progs\NeuDatenB.fre"

***** MITTEILUNG
Lesen oder Schreiben korrekt beendet in Datei
"C:\Almo6\Progs\NeuDatenB.dir"

***** **Erläuterung:**

Almo teilt abschliessend noch mit, daß es die neue Datei "NeuDatenB" angelegt hat, einmal im Format FREI (Erweiterung: .fre) und einmal im Format DIREKT (Erweiterung: .dir). Letztere ist eine Almo-Arbeitsdatei, die in allen Data-Mining-Programmen eingesetzt werden kann. Die neuen Dateien enthalten nun auch die beiden Variable des Einkommens und der Lebenszufriedenheit.

P45.8.2 Schritt 6b: Datenfusion mit der Logitanalyse

Ist die "zu spendende Variable", für die wir Prognosewerte einsetzen wollen, nominal (dichotom oder polytom), dann verwenden wir vorzugsweise die Logitanalyse. Siehe dazu die Begründung in Abschnitt P45.15.1.

Im nachfolgenden Programm Prog45my kann allerdings nur 1 nominale "zu spendende Variable" eingegeben werden. Hat man mehrere nominale Variable, für die man Werte einsetzen möchte, dann muß man nacheinander mehrere Analysen rechnen.

Das Modell der Logitanalyse wird ausführlich in Abschnitt P45.16 beschrieben.

P45.8.2.1 Eingabe in Prog45my

Gegenüber dem oben dargestellten ALM-Programm Prog45mw vertauschen wir nun die Rollen. Die Datei "DatenB.dir" wird zur Spenderdatei, die die nominale Variable der Parteipräferenz an die Empfängerdatei "DatenA.dir" spendet. Das eröffnet uns dann auch die Möglichkeit, im späteren Abschnitt P45.8.3 die beiden ergänzten Dateien zu einer gemeinsamen Datei zu fusionieren.

Prog45my.Msk

Datenfusion
mit Hilfe der Logit-Analyse
für nominale "zu spendende" Variable

Die Datenfusion erfolgt in folgenden 3 Schritten:

Schritt 1:
Zuerst wird mit den Daten der Spenderdatei für die "zu spendende" Variable als Zielvariable ein Logitmodell gerechnet. Als ursächliche Variable werden die Variablen verwendet, die die Spenderdatei gemeinsam mit der Empfängerdatei besitzt.

Schritt 2:
Dann werden mit Hilfe der dabei errechneten Koeffizienten aus den "gemeinsamen" Variablen als ursächlichen Variablen hinsichtlich der "zu spendenden" Variablen Prognosewerte für die Empfängerdatei ermittelt.

Schritt 3:
Diese Prognosewerte können dann noch durch Zufallsvariation verändert werden.
Die so ergänzte Empfängerdatei wird dann in eine neue Datei abgespeichert

siehe Handbuch "P45 Almo-Data-Mining", Abschnitt P45.8

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

1

Vereinbare Variable= ;

2 Option: Weitere Vereinbarungen - nur wenn Almo dazu auffordert

3

"C:\Almo7\Testdat\DatenB.nam"

 zeige = Namensdatei in Output zeigen
leer = nicht

4

5

"C:\Almo7\Testdat\DatenB.dir"

6

"C:\Almo7\Testdat\DatenA.dir"

7

 Variablennummern aus Spenderdatei

 Variablennummern aus Empfängerdatei

Neue Empfängerdatei

"C:\Almo7\Progs\NeuDatA"

Geben Sie den Dateinamen ohne Erweiterung an. Almo erzeugt 2 Dateien:

1. eine nicht lesbare Almo-Arbeitsdatei mit der Erweiterung `__.dir`
2. eine anschauliche Datei im freien Format mit der Erweiterung `__.fre`

Ein Datensatz der neuen Empfängerdatei enthält jetzt die Variablen der alten Datei (im Beispiel sind das U1:U11) plus die gespendete Variable, die angehängt wird (im Beispiel wird die gespendete Variable als U12 angehängt)

P45.8.2.2 Erläuterungen zu den Eingabe-Boxen

Prog45my entspricht weitgehend dem in Abschnitt P45.8.1.2 bereits erläuterten Prog45mw, so daß wir hier nur 3 Eingabe-Boxen erläutern müssen. Der Benutzer sollte beachten, daß in der Eingabe-Box 7 "gemeinsame Variable aus Spender- und Empfängerdatei" im 1. Eingabefeld die gemeinsamen Variablen aus der Spenderdatei einzutragen sind - und dies ist nun, da die Rollen getauscht wurden, die Datei "DatenB.dir".

Eingabe-Box 8: Zu spendende Variable aus der Spenderdatei

Eingabe-Box 9: Ursächliche Variable für die zu spendende Variable in der Spenderdatei

Zu spendende Variable aus der Spenderdatei

(nur eine) nominale Variable

Parteipraeferenz

Ursächliche Variable

für die zu spendende Variable in der Spenderdatei

ursächliche nominale Variable **Hilfe**

Beruf, Geschlecht, Gewerkschaft, GastarbeiterRechte **Hilfe**

ursächliche quantitative Variable **Hilfe**

Alter, Kirchgang

Es ist nur 1 nominale "zu spendende" Variable erlaubt. Sie kann dichotom oder polytom sein. Als ursächliche Variable sind nur nominale (dichotom und polytom) und/oder quantitative Variable erlaubt. Was unter „ursächlichen“ Variablen zu verstehen ist, haben wir bei der Erläuterung zu Eingabe-Box 9 in Abschnitt P45.8.1.2 vorgetragen.

Eingabe-Box 11: Kein-Wert-Behandlung der ursächlichen Variablen aus der Spenderdatei

Kein-Wert-Behandlung

Kein-Wert-Behandlung der ursächlichen Variablen in Spenderdatei

3 bei Berechnung der Regressionskoeffizienten und Effekte nur 3 = Vollständiges Ausscheiden möglich

Kein-Wert-Behandlung der ursächlichen Variablen in Empfängerdatei

4 bei Berechnung der Prognosewerte für zu spendende Variable möglich: 4 - 7; empfohlen: 4 =Mittelwert-Einsetzung

Im Unterschied zu Prog 45mw wird bei der Logitanalyse nur und ausschließlich das "vollständige Ausscheiden" durchgeführt. Der Benutzer kann das nicht beeinflussen. Wenn auch nur eine Analyse-Variable Kein-Wert ist, dann wird der gesamte Datensatz aus der Analyse ausgeschlossen. Hingegen kann bei der Berechnung der Prognosewerte die Kein-Wert-Behandlung wie bei Prog45mw gewählt werden.

P45.8.2.3 Ausgabe aus Prog45my

Almo gibt zuerst die Ergebnisse aus der Logitanalyse aus. Wir werden diese nur insofern erläutern, als sie für unser Thema der Datenfusion von Belang sind. In Abschnitt P45.16.1.3 wird die Ausgabe aus der Logitanalyse im Detail behandelt.

Modellspezifikation: mehrdimensionales Logit-Modell

Analysevariablen:

unabhaengige nominale Variablen:

V4	Beruf	Werte-Untergrenze = 1	Obergrenze = 6
V6	Geschlecht	Werte-Untergrenze = 1	Obergrenze = 2
V8	Gewerkschaft	Werte-Untergrenze = 1	Obergrenze = 2
V10	GastarbeiterRech	Werte-Untergrenze = 1	Obergrenze = 3

Beachte:

Fuer die unabhaengigen nominalen Variablen wird die 0,1,-1 Dummy-Kodierung verwendet.

unabhaengige quantitative Variablen:

V5	Alter
V9	Kirchgang

abhaengige nominale Variable:

V1	Parteipraeferenz	Werte-Untergrenze = 1	Obergrenze = 5
----	------------------	-----------------------	----------------

Beachte:

Zur Schaetzung wird die 1. Auspraegung der abhaengigen Variablen als Referenz verwendet

***** WARNUNG
Datensatz wird wegen fehlender Werte
oder negativer Haeufigkeiten eliminiert

keine 5 | 0 0 0 0 41 | 14 5 1 0 21 |

		prognostiziert relativ					erwartet Zufall				
		1	2	3	4	5	1	2	3	4	5
		SPÖ	ÖVP	FPÖ	Grüne	keine	SPÖ	ÖVP	FPÖ	Grüne	keine
SPÖ	1	24.5	10.4	6.7	0.5	10.9	16.7	15.5	6.3	1.6	12.9
ÖVP	2	10.7	25.6	3.0	0.9	8.8	15.5	14.3	5.8	1.5	12.0
FPÖ	3	6.1	3.3	4.2	0.3	6.1	6.3	5.8	2.4	0.6	4.9
Grüne	4	0.7	1.0	0.3	2.6	0.4	1.6	1.5	0.6	0.1	1.2
keine	5	11.1	8.7	5.7	0.7	14.8	12.9	12.0	4.9	1.2	10.0

absolut: Chi-Quadrat(16) =158.550 Signifikanz 100*(1-p) = 100.000
relativ: Chi-Quadrat(16) = 67.186 Signifikanz 100*(1-p) = 100.000

******* Erläuterung:**

Diese Tabelle bezieht sich nur auf die Personen der Spenderdatei - und dabei auch nur auf jene die in allen ursächlichen Variablen und in der Zielvariablen valide Werte besitzen. Für diese prognostiziert ALMO welche Partei sie präferieren. Diese Prognose wird dann mit der tatsächlichen Parteipräferenz verglichen. So kann die Trefferhäufigkeit festgestellt werden. Betrachten wir aus diese Tabelle der Trefferhäufigkeiten die oberen beiden Teiltabellen und die Teiltabelle unten rechts. Dabei genügt es die Diagonalen anzuschauen:

	tatsächlich	richtig prognostiziert	zufällig richtig prognostiziert
SPÖ	53	37	17
ÖVP	49	33	14
FPÖ	20	2	2
Grüne	5	3	0
keine	41	21	10

Alle Parteien außer der FPÖ konnten durch die Logitanalyse besser prognostiziert werden, wie wenn man zufällig die Personen den Parteien zugewiesen hätte.

Berechnung der Prognosewerte bzw. Schätzwerte fuer Variable mit Kein_Wert

***** MITTEILUNG
Sind unabhaengige Variable, die für die Berechnung des Prognosewerts benoetigt werden, gleich "Kein_Wert" dann wird fuer sie "Kein-Wert-Behandlung = 4" durchgefuehrt

Mittelwert und Standardabweichung der Residuen fuer Variable V1 Parteipraeferenz: SPÖ, ÖVP, FPÖ, Grüne, keine

Gruppe	Mittelwert	Standardabweichung
Gruppe 1 SPÖ	0.00167462	0.415207
Gruppe 2 ÖVP	-0.0108029	0.38019
Gruppe 3 FPÖ	0.00883044	0.313126
Gruppe 4 Grüne	-0.000251497	0.111522
Gruppe 5 keine	0.000549294	0.392639

******* Erläuterung:**

ALMO rechnet für die Personen der Spenderdatei (die ausschließlich valide Werte in den Analyse-Variablen besitzen) die Residuen als Differenz zwischen dem Prognosewert und dem tatsächlichen Wert. Der Mittelwert dieser Residuen ist (fast) 0. Die Standardabweichung der Residuen wird verwendet, wenn der Benutzer in Eingabe-Box 12 "Transformation der Prognosewerte für zu spendende Variable" die Methode 6 oder 7 einträgt. Diese beiden Methoden erzeugen für die Personen der Empfängerdatei einen Wert für die "zu spendende" Variable, der aus einer normalverteilten Zufallsvariation des Prognosewertes mit der oben angegebenen Standardabweichung hervorgeht.

Ursachliche Variable, die Kein_Wert waren und durch Schaetzwerte ersetzt wurden

```
V4 Beruf          in    0 Datensatzen
V6 Geschlecht    in    0 Datensatzen
V8 Gewerkschaft in    5 Datensatzen
V10 Gstarbeiter  in    3 Datensatzen
V5 Alter         in    0 Datensatzen
V9 Kirchgang     in    4 Datensatzen
```

******* Erläuterung:**

Beispiel: Die ursächliche Variable "Gewerkschaft" besaß in 5 Datensätzen keinen Wert usw. Diese Angaben beziehen sich auf die Personen aus der Spenderdatei. Da wir Eingabe-Box 11 "Kein-Wert-Behandlung der ursächlichen Variablen aus Spenderdatei" Eingabefeld 2 ("bei Berechnung der Prognosewerte für zu spendende Variable") als "Kein-Wert-Behandlung = 4" eingesetzt hatten, wird in diesen Variablen der Mittelwert (bei quantitativen Variablen) bzw. der Erwartungswert (bei nominalen Variablen) als Ersatzwert eingesetzt.

In "Empfaengerdatei" fuer "gespendete" Variable eingesetzter Schaetzwert

```
-----
***** MITTEILUNG
Der fuer die "gespendete" Variable eingesetzte Schaetzwert
wird noch der "Prognosewert-Behandlung = 7" unterworfen
```

└─ Die Variablennummer ist die aus der Spenderdatei

Datensatz	Variable	eingesetzter Wert
1	V1 Parteipr	2
2	V1 Parteipr	3
3	V1 Parteipr	2
4	V1 Parteipr	5
5	V1 Parteipr	2
.	.	.
.	.	.
.	.	.
182	V1 Parteipr	1
183	V1 Parteipr	1
184	V1 Parteipr	4
185	V1 Parteipr	2
186	V1 Parteipr	1

******* Erläuterung:**

Almo teilt die Werte mit, die in der Empfängerdatei für die zu "spendenden" Variablen der Parteipräferenz eingesetzt werden.

```
***** MITTEILUNG
Lesen oder Schreiben korrekt beendet in Datei
"C:\Almo6\Progs\NeuDatenA.fre"

***** MITTEILUNG
Lesen oder Schreiben korrekt beendet in Datei
"C:\Almo6\Progs\NeuDatenA.dir"
```

******* Erläuterung:**

Almo teilt abschliessend noch mit, daß es die neue Datei "NeuDatenA" angelegt hat, einmal im Format FREI (Erweiterung: .fre) und einmal im Format DIREKT (Erweiterung: .dir). Letztere ist eine Almo-Arbeitsdatei, die in allen Data-Mining-

Programmen eingesetzt werden kann. Die neuen Dateien enthalten nun auch die Variable der Parteipräferenz

P45.8.3 Schritt 6c: Fusionierte Dateien vereinen

Wir haben nun mit Prog45mw von der Datei A an die Datei B 2 Variable "gespendet" und umgekehrt haben wir mit Prog45my von Datei B an Datei A eine Variable gespendet. Nunmehr können wir die beiden so vergrößerten Dateien zusammenfassen. Die Art der Zusammenfassung wollen wir zuerst an zwei kleinen Dateien vorführen.

Datei A			Datei B			
Beruf	Alter	V3	Beruf	V2	Alter	V4
V1	V2	V3	V1	V2	V3	V4
1	21	1.5	2	46.2	24	155
3	23	3.3	2	23.8	27	268
3	27	7.0	3	10.0	21	888
1	22	2.6	1	9.7	29	342
1	29	4.1	1	97.1	25	98
4	31	2.1				

Die Datei A besteht aus 3 Variablen: V1 Beruf, V2 Alter und V3 Irgendetwas.
 Die Datei B besteht aus 4 Variablen: V1 Beruf, V2 Irgendetwas, V3 Alter und V4 Irgendetwas.
 In Datei A sind 5 Datensätze enthalten und in Datei B 4.

Gemeinsame Variable sind:
 der Beruf, das ist V1 aus Datei A und V1 aus Datei B
 das Alter, das ist V2 aus Datei A und V3 aus Datei B

Spezifische Variable sind:
 in Datei A: V3
 in Datei B: V2, V4

Nach der Zusammenfassung ist die neue Datei zunächst (und vorübergehend) folgende:

	Beruf Alter			Beruf Alter			
	V1	V2	V3	V4	V5	V6	V7
Datei A	1	21	1.5	1	kw	21	kw
	3	23	3.3	3	kw	23	kw
	3	27	7.0	3	kw	27	kw
	1	22	2.6	1	kw	22	kw
	1	29	4.1	1	kw	29	kw
	4	31	2.1	4	kw	31	kw
	2	24	kw	2	46.2	24	155
	2	27	kw	2	23.8	27	268
	3	21	kw	3	10.0	21	888
	1	29	kw	1	9.7	29	342
	1	25	kw	1	97.1	25	98

Die beiden Dateien A und B sind jetzt die Submatrizen in der Diagonalen. Die Variablen der Datei B haben somit die Nummern V4 bis V6. Die Submatrizen in der Gegendiagonalen werden zunächst auf Kein-Wert ("kw") gesetzt. Danach werden die

gemeinsamen Variablen nachgetragen. Damit sind jetzt die gemeinsamen Variablen doppelt vorhanden. Die Spalte V1 ist identisch mit der Spalte V4 und Spalte V2 mit Spalte V6. Es ist deswegen sinnvoll, die Spalte V4 und V6 nicht in die entgültige zusammengefasste Datei zu übernehmen.

Die entgültige zusammengefasste Datei ist dann folgende:

Beruf			Alter	
V1	V2	V3	V4	V5
1	21	1.5	kw	kw
3	23	3.3	kw	kw
3	27	7.0	kw	kw
1	22	2.6	kw	kw
1	29	4.1	kw	kw
4	31	2.1	kw	kw
2	24	kw	46.2	155
2	27	kw	23.8	268
3	21	kw	10.0	888
1	29	kw	9.7	342
1	25	kw	97.1	98

P45.8.3.1 Eingabe in Prog45mx

Prog45mx führt die oben dargestellte Zusammenfassung von 2 Dateien automatisch durch. Der Benutzer muß trotzdem sorgfältig darauf achten, daß er mit den Variablennummern nicht durcheinander gerät.

Prog45mx.Msk

Zusammenfassen von
2 verschiedene Dateien,
die einige Variable gemeinsam besitzen

Beispiel:

Datei A			Datei B			
Beruf	Alter		Beruf	Alter		
U1	U2	U3	U1	U2	U3	U4
1	21	1.5	2	46.2	24	155
3	23	3.3	2	23.8	27	268
3	27	7.0	3	10.0	21	888
1	22	2.6	1	9.7	29	342
1	29	4.1	1	97.1	25	98
4	31	2.1				

Gemeinsame Variable sind:
der Beruf, das ist U1 in Datei A und U1 in Datei B
das Alter, das ist U2 in Datei A und U3 in Datei B

Spezifische Variable sind:
in Datei A: U3
in Datei B: U2, U4

Nach der Zusammenfassung ist die neue Datei
zunächst folgende:

	Beruf	Alter		Beruf	Alter		
	U1	U2	U3	U4	U5	U6	U7
Datei A	1	21	1.5	1	kw	21	kw
	3	23	3.3	3	kw	23	kw
	3	27	7.0	3	kw	27	kw
	1	22	2.6	1	kw	22	kw
	1	29	4.1	1	kw	29	kw
	4	31	2.1	4	kw	31	kw
Datei B	2	24	kw	2	46.2	24	155
	2	27	kw	2	23.8	27	268
	3	21	kw	3	10.0	21	888
	1	29	kw	1	9.7	29	342
	1	25	kw	1	97.1	25	98

kw = Kein Wert

Die beiden Dateien A und B sind jetzt die Submatrizen
in der Diagonalen. Die Variablen der Datei B haben
somit die Nummern U4 bis U6. Die Submatrizen in der
Gegendiagonalen werden zunächst auf "KeinWert" gesetzt.
Danach werden die gemeinsamen Variablen nachgetragen.
Damit sind jetzt die gemeinsamen Variablen doppelt
vorhanden. Die Spalte U1 ist identisch mit der
Spalte U4 und Spalte U2 mit Spalte U6.

Es ist deswegen sinnvoll, die Spalte U4 und U6 nicht
in die entgeltliche zusammengefasste Datei zu übernehmen
Also speichert deswegen folgende Variable aus obiger
vorläufigen Datenmatrix: U1,2,3,5,7

Die entgeltige zusammengefasste Datei ist dann folgende:

Beruf		Alter		U4	U5
U1	U2	U3			
1	21	1.5		kw	kw
3	23	3.3		kw	kw
3	27	7.0		kw	kw
1	22	2.6		kw	kw
1	29	4.1		kw	kw
4	31	2.1		kw	kw
2	24	kw	46.2	155	
2	27	kw	23.8	268	
3	21	kw	10.0	888	
1	29	kw	9.7	342	
1	25	kw	97.1	98	

siehe Handbuch "P45 Almo-Data-Mining", Abschnitt P45.8

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

1 Speicher fuer x Variable
 Vereinbare Variable= 25 ;

2 Option: Weitere Vereinbarungen - nur wenn Almo dazu auffordert

3
 "C:\Almo7\Testdat\NeuDatA.dir"
 12 die Datei umfasst so viele Variable

4
 "C:\Almo7\Testdat\NeuDatB.dir"
 12 die Datei umfasst so viele Variable

5
 U1, 2, 3, 4, 5, 6, 7, 8, 9, 12 Variablennummern aus Datei A
 U3, 4, 11, 5, 6, 8, 9, 10, 12, 1 Variablennummern aus Datei B
 10 Zahl der gemeinsamen Variablen

6
 U2, 7

7
 sie enthält alle Variable aus Datei A
 daran angehängt die spezifischen Variablen aus Datei B
 "C:\Almo7\Progs\DatenFusion"
 Geben Sie den Dateinamen ohne Erweiterung an. Almo erzeugt 2 Dateien:
 1. eine nicht lesbare Almo-Arbeitsdatei mit der Erweiterung __.dir
 2. eine anschaulbare Datei im freien Format mit der Erweiterung __.fre

P45.8.3.2 Erläuterung zu den Eingabe-Boxen

Eingabe-Box 1 und Eingabe-Box 2:

Siehe "Arbeiten mit Almo-Datenanalyse-System", Abschnitt P0.1 und P0.2.

Eingabe-Box 3: Datei A

Eingabe-Box 4: Datei B

Two screenshots of software input boxes for file selection. The first box is titled "Datei A" and shows the file path "C:\Almo15\Testdat\NeuDatA.dir" and a variable count of 12. The second box is titled "Datei B" and shows the file path "C:\Almo15\Testdat\NeuDatB.dir" and a variable count of 12. Both boxes include a "Hilfe" button and a description: "die Datei umfasst so viele Variable".

Geben Sie die Namen der beiden Dateien an, sowie die Zahl der Variablen, die in diesen

Dateien enthalten ist. Die Variablen in den Dateien müssen mit V1 beginnend fortlaufend nummeriert sein, d.h. die Dateien müssen ursprünglich mit derart fortlaufenden Variablennummern angelegt worden sein. In der Regel wird das auch so sein. In unserem Fusions-Beispiel ist Datei A um 1 "gespendete" Variable (die Parteipräferenz) vergrößert worden. Sie umfasst also jetzt 12 Variable. Datei B wurde um 2 Variable vergrößert (das Einkommen und die Lebenszufriedenheit), umfasst so nun auch 12 Variable.

Eingabe-Box 5: Gemeinsame Variable aus Datei A und Datei B

Screenshot of the "gemeinsame Variable aus Datei A und Datei B" input box. It shows two rows of variable numbers: "V1, 2, 3, 4, 5, 6, 7, 8, 9, 12" for Datei A and "V3, 4, 11, 5, 6, 8, 9, 10, 12, 1" for Datei B. A third row shows the count of common variables as 10. A "Hilfe" button is also present.

Am besten arbeitet man hier mit Bleistift und Papier. Zuerst schreibt man sich die Variablen aus der Datei A auf, die diese gemeinsam mit der Datei B hat. Dann schreibt man die entsprechenden Variablennummern aus der Datei B darunter. Zu beachten ist nun, daß die gegenseitig "gespendeten" Variablen nunmehr auch gemeinsame Variable sind. So kommt also in der Datei A noch hinzu:

V3 Einkommen

V9 Lebenszufriedenheit

Diese beiden Variablen wurden von Datei A an Datei B gespendet. Sie erhielten in Datei B die Variablennummern V11 und V12. Und es kommt noch hinzu:

V12 Parteipräferenz

Diese Variable wurde von B an A gespendet und in der Datei A als letzte Variable mit der Nummer V12 angehängt.

Die gemeinsamen Variablen sind also folgende

gemeinsame Variable aus Datei A -----	Nummern der gemeinsamen Variablen aus Datei B -----	
V1 Bildungsniveau	V3	
V2 Beruf	V4	
V3 Einkommen	V11	von A an B gespendete Variable
V4 Alter	V5	
V5 Geschlecht	V6	
V6 Gewerkschaftsmitglied	V8	
V7 Kirchengangshäufigkeit	V9	
V8 Gastarbeitern	V10	
V9 Lebenszufriedenheit	V12	von A an B gespendete Variable
V12 Parteipräferenz	V1	von B an A gespendete Variable

Eingabefeld 1: Die ganz links stehenden Variablennummern aus der Datei A werden in das 1. Eingabefeld geschrieben. Sie müssen in folgender Form geschrieben werden:

V1,2,3,4,5,6,7,8,9,12

Vorne steht 'V' darauf folgen die Nummern - getrennt durch Beistrich. Zulässig wäre auch die Kurzschreibweise (mit Verwendung des Doppelpunktes)

V1:9,12

Eingabefeld 2: Die oben rechts stehenden Variablennummern aus der Datei B werden in das 2. Eingabefeld geschrieben. Sie müssen in folgender Form geschrieben werden:

V3,4,11,5,6,8,9,10,12,1

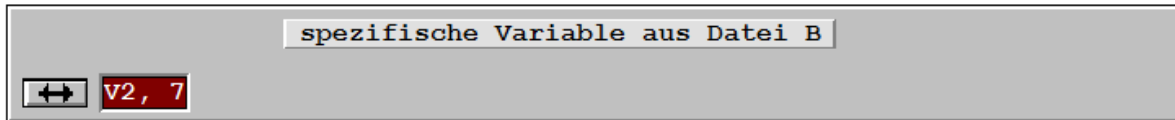
Vorne steht 'V' darauf folgen die Nummern - getrennt durch Beistrich. Zulässig wäre auch die Kurzschreibweise.

Am besten schreibt man die beiden Nummernreihen in den beiden Eingabefeldern exakt untereinander, so daß man optisch kontrollieren kann, welche Variable aus der Datei A welchen Variablen aus der Datei B entsprechen. Das geht nicht, wenn man die Kurzschreibweise verwendet.

Beachte: Wenn Sie gemeinsame Variable vergessen, dann werden diese als spezifische Variable in die neue Datei übernommen und sind dann doppelt in der neuen Datei enthalten.

Eingabefeld 3: Geben Sie an, wie viele Variable Sie im 1. Eingabefeld (bzw. im 2.) geschrieben haben.

Eingabe-Box 6: Spezifische Variable aus Datei B



Die spezifischen Variablen aus Datei B in unserem Beispiel sind:

- V2 Parteimitgliedschaft
- V7 ZeitungLesen

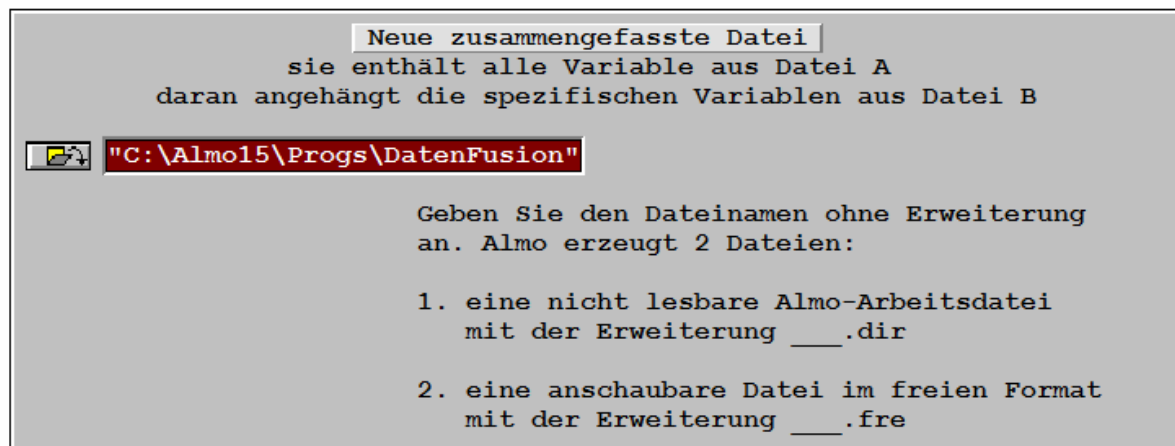
Schreiben Sie nur die Variablennummern. Sie müssen in folgender Form geschrieben werden:

V2,7

Vorne steht 'V' darauf folgen die Nummern - getrennt durch Beistrich. Zulässig wäre auch die Kurzschreibweise (mit Verwendung des Doppelpunktes).

Beachte: Die spezifischen Variablen aus Datei A müssen nicht angegeben werden.

Eingabe-Box 7: Neue zusammengefasste Datei



Die neue Datei enthält alle Variable aus Datei A und daran angehängt die spezifischen Variablen aus Datei B. Für unsere Beispieldaten wird folgende neue Datei erzeugt:

	alle Variable aus Datei A										spezifische Variable aus Datei B			
	V1	V10	V11	V12	V13	V14							
Personen aus der Datei A	4	6	12	65	1	2	5	1	2	2	2	2	kw	kw
	2	5	5	68	1	2	5	3	1	1	2	3	kw	kw
	1	1	kw	66	2	2	4	3	2	1	2	2	kw	kw
	2	6	8	67	1	2	5	2	2	1	1	5	kw	kw

Personen aus der Datei B	2	4	5	64	1	1	4	3	3	kw	kw	2	2	1
	2	3	5	57	1	2	2	3	1	kw	kw	2	2	1
	3	6	8	68	2	2	5	2	1	kw	kw	2	2	1
	2	6	5	59	1	1	6	2	1	kw	kw	1	2	1

Die neue Datei umfasst insgesamt 14 Variable, die von Almo die Nummern V1 bis V14 erhalten. Von V1 bis V12 reichen die Variablen aus der Datei A (und zwar genau in der Reihenfolge in der sie in Datei A stehen). V13 und V14 sind die spezifischen Variablen aus der Datei B, also

V13 Parteimitgliedschaft (ehemals V2)
V14 ZeitungLesen (ehemals V7)

Die Reihenfolge der spezifischen Variablen haben wir in Eingabe-Box 6 "Spezifische Variable aus Datei B" festgelegt.

Der Benutzer sollte nun eine Datei der Variablennamen für die neue Datei erstellen. Da die Variablen V1 bis V12 aus der Datei A sind, kann man die Variablennamen der Datei A übernehmen.

```
Name 1=Bildung:Pflichts ohne Lehre, Pflichts mit Lehre, mittlere Schule,  
Gymnasium,Hochschule;  
Name 2=Beruf:Bauern,  
Selbständig,  
Arbeiter,  
Facharbeiter,  
Angest/Beamte,  
leitende Angest/Beamte;  
Name 3=Einkommen;  
Name 4=Alter;  
Name 5=Geschlecht:m,w;  
Name 6=Gewerkschaft:ja,nein;  
Name 7=Kirchgang:1 mal pro Woche,  
2-3 mal im Monat,  
1 mal im Monat,  
mehrmals im Jahr,  
selten,  
nie;  
Name 8=GastarbeiterRechte:zu wenig,ausreichend,zuviel;  
Name 9=Lebenszufriedenheit:sehr zufrieden,  
ziemlich,  
eher zufrieden,  
eher unzufrieden,  
ziemlich unzufrieden;  
Name 10=FrauNichtBeruf:stimmt,stimmt nicht;  
Name 11=Todesstrafe:dafür,bedingt dafür,dagegen;
```

Datei A erhielt von Datei B die Variable der Parteipräferenz "gespendet". Sie wurde als V12 in Datei A angehängt.

```
Name 12=Parteipraeferenz:SPÖ,ÖVP,FPÖ,Grüne,keine;
```

Es müssen nun nur noch die 2 spezifischen Variable aus der Datei B hinzugefügt werden:

```
Name 13=Parteimitglied:ja,nein;  
Name 14=ZeitungLesen:regelmässig,nicht regelmässig;
```

Wir haben die in der neuen Datei gespeicherte Datenmatrix durch Striche unterteilt. Links des senkrechten Trennstrichs stehen die Variablen aus Datei A und rechts die spezifischen Variablen aus der Datei B. Oberhalb des horizontale Trennstrichs in der Mitte stehen die Personen aus der Datei A. Sie haben selbstverständlich in den beiden letzten Variablen V13 und V14 "KeinWert" (kurz: kw) als Wert. Unterhalb des horizontale Trennstrichs stehen die Personen aus Datei B. Sie haben in V13 und V14 einen Wert. In V10 und V11 haben sie keinen Wert.

Dies sind die spezifischen Variablen der Datei A, und zwar die Variable V10 "FrauNichtBerufstätig" und die Variable V11 "Einstellung zu Todesstrafe"

Von V1 bis V9 und in V12 stehen die gemeinsamen Variablen der Personen aus Datei B, nunmehr in anderer Reihenfolge als in der ursprünglichen Datei B - eben in der Reihenfolge, in der sie in Datei A enthalten sind.

P45.8.3.3 Ausgabe

Die Almo-Ausgabe besteht aus mehreren Mitteilung. Die letzten beiden lauten:

```
***** MITTEILUNG
Lesen oder Schreiben korrekt beendet in Datei
"C:\Almo6\Progs\DatenFusion.fre"

***** MITTEILUNG
Lesen oder Schreiben korrekt beendet in Datei
"C:\Almo6\Progs\DatenFusion.dir"
```

******* Erläuterung:**

Almo teilt mit, daß es die neue Datei "DatenFusion" angelegt hat, einmal im Format FREI (Erweiterung: .fre) und einmal im Format DIREKT (Erweiterung: .dir). Letztere ist eine Almo-Arbeitsdatei, die in allen Data-Mining-Programmen eingesetzt werden kann. Aus diesen beiden Mitteilungen darf aber nicht geschlossen werden, daß die neue Datei auch so angelegt wurde, wie es sich der Benutzer gewünscht hat. Die Möglichkeit von Fehleingaben durch den Benutzer ist in Prog45mx nicht gering. Almo kann nur in beschränktem Maße die Eingaben des Benutzers überprüfen. Der Benutzer sollte also auf jeden Fall die neue Datei kontrollieren. Es genügt eigentlich schon, wenn sich der Benutzer die neue Datei "DatenFusion.fre" in ein Fenster lädt (durch Doppelklick auf den Dateinamen in obiger Mitteilung) und dann bei der 1. Person aus der Datei A und bei der 1. Person aus der Datei B überprüft ob die Variablen-Reihenfolge stimmt.

P45.8.3.4 Weiterführende Hinweise

Damit eine Datenfusion brauchbar ist, sollten mindestens folgende zwei Bedingungen erfüllt sein.

1. Die Spenderdatei und die Empfängerdatei müssen als Stichproben aus derselben Grundgesamtheit entstanden sein.
2. Die Determination der "zu spendenden" Variablen in der Spenderdatei durch die ursächlichen Variablen muß gut sein. Als Maß dafür kann man beim ALM die multiple Korrelation verwenden.

Wie wir bereits erwähnt haben, wird die Datenfusion häufig auch über die Clusteranalyse vorgenommen. Zu einem späteren Zeitpunkt wird Johann Bacher ein derartiges Programm zur Verfügung stellen.

Zur Literatur: Eine Diskussion der Datenfusion aus der Sicht des Statistikers geben S. Räsler/K.H. Fleischer (1997).

Kapitel 5: Mehrere Variable zu einer Messung kombinieren

Betrachten wir ein Beispiel: Die Ausdauerleistung von Sportlern soll gemessen werden. Die Sportler müssen folgende Sportarten absolvieren:

1. 5000 m Lauf
2. Berglauf von 1000 m Höhendifferenz
3. 50 km Zeitfahren mit dem Fahrrad
4. 15 km Ski-Langlauf im Skating-Stil
5. Kugelstoßen
6. Speerwurf

P45.9 Schritt 7a: Aus mehreren Variablenwerten einen Gesamtpunktwert bilden

Die einfachste Methoden besteht nun darin, die sechs Leistungen zu addieren. Da die Leistungen nicht unmittelbar miteinander vergleichbar sind, müssen wir sie standardisieren und erst dann zu einem Gesamtpunktwert summieren. Standardisiert wird nach der Formel

$$x = \frac{y - M}{s}$$

x = standardisierter Wert

y = Rohwert (in der Sportart)

M = Mittelwert (in der Sportart über alle Sportler)

S = Standardabweichung

Der Almo-Benutzer braucht das nicht selbst zu rechnen. Das nachfolgend beschriebene Almo-Programm nimmt ihm diese Arbeit ab.

Wichtig ist nun folgende Feststellung:

Es muß annähernd die Gewißheit bestehen, daß die Variablen, die zu einem Gesamtpunktwert zusammengefasst werden sollen, alle daselbe messen.

Die 6 sportlichen Leistungen sollen die Ausdauerleistung der Sportler messen. Hier sind wohl Zweifel angebracht. Speerwurf und Kugelstoßen messen wohl eher "Kraft" als "Ausdauer". Wir werden im nächsten Abschnitt P45.10 mit der Faktorenanalyse ein Verfahren kennen lernen, daß fähig ist, die Hintergrundfaktoren, die hinter den 6 Sportarten stehen, zu identifizieren.

Eingabe in Prog45mv

Prog45mv.Msk
Gesamtpunktwert bilden

Aus mehreren Variablenwerten einen Gesamtpunktwert bilden
Speichern der Gesamtpunktwerte der Probanden in einer Datei

Der für jeden Probanden ermittelte Gesamtpunktwert wird in
eine neue Variable gegeben. Der so verlängerte Datensatz wird
in eine neue Datei geschrieben

Was ist ein Kurzprogramm ? -->

Bedienung -->

1

Vereinbare Variable = ;

2 Option: Weitere Vereinbarungen - nur wenn Almo dazu auffordert

3

zeige = Namensdatei in Output zeigen
leer = nicht

4

erzeuge zusätzliche Namensfelder

5

6

1 = Variable standardisieren
0 = nicht

7 Option: Ein- und Ausschliessen von Untersuchungseinheiten

Die Variablen müssen gleichgerichtet kodiert sein. Ist dies nicht der Fall, dann müssen die "verkehrt" kodierten Variablen in ihrer Kodierungsrichtung umgedreht werden

8

↓ Loesche wieder diese Box

Umkodierungen und Kein-Wert-Angaben

Umkodierungen
Kein_Wert-Angabe

↔	↓	xLauf5000m	= -	Lauf5000m	;
↔	↓	xBerglauf	= -	Berglauf	;
↔	↓	xRad50km	= -	Rad50km	;
↔	↓	xSkiLanglauf	= -	SkiLanglauf	;

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

9

↓ Option: Untersuchungseinheiten gewichten

10

Gesamtpunktwerte in neue Datei speichern

↔ "C:\Almo7\PROGS\Gespktwert"

Geben Sie einen neuen Dateinamen ohne Erweiterung an
Almo erzeugt 2 Dateien:

- eine nicht lesbare Almo-Arbeitsdatei mit der Erweiterung __.dir
- eine anschauliche Datei im freien Format mit der Erweiterung __.fre

↔ U1:6

aus der "Datei aus der gelesen wurde" sollen diese Variablen in die neue Datei übernommen werden

↔ U7

die Gesamtpunktwerte sollen in der neuen Datei in diese Variable eingetragen werden

BEACHT: Als Nummern für die Gesamtpunktvariable müssen freie, sonst nicht verwendete Variablennummern verwendet werden

↔ *100; Runde 1

die Gesamtpunktvariable transformieren z.B. mit 100 multiplizieren und auf eine Ganzzahl runden - nicht obligatorisch

↑ ↓ ?

Kein-Wert-Behandlung, wenn die Variablen, die zusammengefasst werden, Kein_Wert besitzen
Empfohlen: 5 oder ?

↔ 123457

Startwert für Zufallsgenerator für Kein-Wert-Behandlung 6 und ?

P45.9.1 Erläuterung zu den Eingabe-Boxen

Eingabe-Box 1 bis **Eingabe-Box 5:**

Siehe "Arbeiten mit Almo-Datenanalyse-System", Abschnitt P0.1 bis P0.3.

Eingabe-Box 4: Freie Namensfelder

Freie Namensfelder Hilfe

Leere alle Eingabefelder dieser Sub-Box

Name 10=xLauf5000m
 Name 11=xBerglauf
 Name 12=xRad50km
 Name 13=xSkiLanglauf

erzeuge zusätzliche Namensfelder

4 der Variablen, die zu einem Gesamtpunktwert zusammengefasst werden sollen, müssen in Hilfsvariable umgespeichert werden. Das geschieht in der Eingabe-Box 8 "Kein-Wert-Angabe und Umkodierungen". Wir werden dort auch erklären, warum das notwendig ist. Als Hilfsvariable verwenden wir die freien Variablennummern V10 bis V13. Hier werden nun diesen Hilfsvariablen Namen zugewiesen. Dabei verwenden wir einfach die Namen der Originalvariablen und setzen ein x davor.

Eingabe-Box 5: Datei aus der gelesen wird

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

Eingabe-Box 6: Quantitative oder dichotome Variable, die zu einem Gesamtpunktwert zusammengefasst werden sollen

Quantitative oder dichotome Variable, die zu einem Gesamtpunktwert zusammengefasst werden sollen

xLauf5000m, xBerglauf, xRad50km, xSkiLanglauf, Kugelstossen, Speerwurf

0 1 = Variable standardisieren
0 = nicht

Wir haben hier für den 5000m-Lauf, den Berglauf, das Radfahren und den Skilanglauf nicht die Originalvariablen eingesetzt, sondern Hilfsvariable, für die wir die freien Variablennummern V10 bis V13 verwendet haben und denen wir oben in Eingabe-Box 4 Variablenamen gegeben haben. Wir werden in der Erläuterung zu Eingabe-Box 8 "Kein-Wert-Angabe und Umkodierungen" erklären, warum wir das getan haben.

Zum Messniveau der Variablen gelten folgende Regeln:

1. Die Variablen müssen quantitativ sein. Siehe zum Begriff "quantitativ" P45.12.0.

Oder: Die Variablen müssen dichotom sein. Siehe zum Begriff "dichotom" P45.12.0.

Oder: Die Variablen können quantitativ und dichotom sein. Die Dichotomien müssen dann aber als in 2 Abschnitte geteilte Quantitäten interpretierbar sein. Beispiel: Die Variable des Körpergewichts kann in "leicht" und "schwer" unterteilt werden.

Eingabefeld 1: Geben Sie die Variablen an, die zu einem Gesamtpunktwert zusammengefasst werden sollen. Am einfachsten geht das, wenn Sie auf den Knopf mit den 2 kleinen symbolischen Fenstern klicken. Siehe zu dieser Vorgehensweise "Arbeiten mit Almo-Datenanalyse-System", Abschnitt P0.11.

Eingabefeld 2: Geben Sie "1" ein, wenn der Gesamtpunktwert aus standardisierten Variablen ermittelt werden soll und "0", wenn nicht.

Hier gelten folgende Regeln:

4. Es kann immer standardisiert werden. Das ist nie falsch.
5. Wenn die Variablen in verschiedenen Maßeinheiten gemessen sind, dann muss standardisiert werden. In unserem Beispiel wird etwa der 5000 m Lauf in Sekunden und der Speerwurf in Metern gemessen.
6. Wenn bei gleicher Maßeinheit der Wertebereich der Variablen verschieden ist, dann muß standardisiert werden. In unserem Beispiel wird der 15 km Skilanglauf maximal bis 2 Stunden dauern, der 5000 m Lauf hingegen nur 1/2 Stunde.

Eingabe-Box 7: Option: Ein- und Ausschliessen von Untersuchungseinheiten
Siehe P0.7.

Eingabe-Box 8: Kein_Wert-Angabe und Umkodierungen

Loesche wieder diese Sub-Box

Eingabefelder für Umkodierungen und Kein-Wert-Angaben

	xLauf5000m = - Lauf5000m ;
	xBerglauf = - Berglauf ;
	xRad50km = - Rad50km ;
	xSkiLanglauf = - SkiLanglauf ;

Siehe unsere ausführliche Darstellung in "Arbeiten mit Almo-Datenanalyse-System", Abschnitt P0.5.

Um einen sinnvollen Gesamtpunktwert bilden zu können, müssen die Variablen gleichgerichtet sein. In unserem Beispiel ist die Ausdauerleistung umso besser je kürzer die Zeit für den 5000 m Lauf ist - aber umso besser je länger der Speerwurf ist. Die beiden Variablen sind "gegengerichtet" kodiert. Eine muß "umgedreht" werden.

Die einfachste Methode eine Variable "umzudrehen" ist, ihr ein negatives Vorzeichen zu verleihen. Deswegen haben wir umkodiert

```
xLauf5000m = - Lauf5000m ;  
xBerglauf = - Berglauf ;  
xRad50km = - Rad50km ;  
xSkiLanglauf = - SkiLanglauf ;
```

(Semikolon zum Gleichungsende nicht vergessen !!)

Für die Variablen xLauf5000m etc. haben wir V10 bis V13 verwendet (die sonst nicht benutzt werden). In Eingabe-Box 4 haben wir ihnen die Variablennamen gegeben und in Eingabe-Box 6 haben wir sie als Analysevariable eingesetzt.

Es ist weiterhin sinnvoll, die Variablen so zu kodieren, dass mit wachsenden Variablenwerten die Eigenschaft (die über den Gesamtpunktwert gemessen werden soll) zunimmt. Das ist beim Speerwurf und beim Kugelstoßen der Fall (je weiter je besser) aber nicht bei den anderen 4 Sportarten. Hier gilt: Je größer der Variablenwert umso schlechter die Ausdauerleistung. Auch deswegen haben wir die obigen 4 Variablen "umgedreht" und nicht den Speerwurf und das Kugelstoßen.

Eingabe-Box 9: Gesamtpunktwerte in neue Datei speichern

↔ "C:\Almo15\PROGS\Neudat"

Geben Sie einen neuen Dateinamen ohne Erweiterung an
Almo erzeugt 3 Dateien:

1. eine nicht lesbare Almo-Arbeitsdatei mit der Erweiterung `__.dir`
2. eine anschauliche Datei im freien Format mit der Erweiterung `__.fre`
3. eine Datei der Variablennamen mit der Erweiterung `__.nam`

In den unter 1. und 2. angegebenen neuen Dateien sind nun enthalten:

- die Variablen aus der alten Datei
- die Gesamtpunktvariable, die als letzte Variable hinter die Variablen der alten Datei gestellt wurde

In der unter 3. angegebenen neuen Datei der Variablennamen sind nun enthalten:

- die Variablennamen aus der alten Datei einschliesslich der in der Box "Freie Namensfelder" angegebenen Namen
- der Name "Skala" für die neue angehängte Gesamtpunktvariable

↔ *100; Runde 1

die Gesamtpunktvariable transformieren z.B. mit 100 multiplizieren und auf eine Ganzzahl runden - nicht obligatorisch

↔ 2

Von den zu addierenden Variablen dürfen x Kein_Wert besitzen. Für sie wird nachfolgende Kein-Wert-Behandlung durchgeführt. Sonst wird die Gesamtpunktvariable auf Kein-Wert gesetzt

↑↓ 6

Kein-Wert-Behandlung, wenn die Variablen, die zusammengefasst werden, Kein_Wert besitzen
Empfohlen: 5 oder 7

↔ 123457

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

Wir werden diese Eingabe-Box in Zusammenhang mit dem Rasch-Skalierungsverfahren in Abschnitt P45.11.2 erläutern.

P45.9.2 Ausgabe der Ergebnisse

Die Almo-Ausgabe besteht lediglich aus den Mitteilungen, dass die gewünschten neuen Dateien angelegt worden sind

```
***** MITTEILUNG
Lesen oder Schreiben korrekt beendet in Datei
"C:\Almo6\PROGS\Gespktwert.fre"
```

```
***** MITTEILUNG
Lesen oder Schreiben korrekt beendet in Datei
"C:\Almo6\PROGS\Gespktwert.dir"
```

Besaßen einige der Variablen Kein-Wert, dann werden die von Almo eingesetzten Ersatzwerte mitgeteilt.

P45.10 Schritt 7b: Mit Faktorenanalyse einen gewichteten Gesamtpunktwert (Faktorwert) bilden

Wir verfügen über mehrere Variable (die sechs Sportarten), von denen wir vermuten, daß sie Indikatoren eines oder mehrerer Hintergrundfaktoren sind. Bei den ersten vier Sportarten ist offenkundig, daß sie etwas mit Ausdauerleistung zu tun haben. Ein Sportler mit guter Ausdauerleistung wird in allen diesen vier Sportarten sicherlich gute Leistungen – allerdings wohl unterschiedlich gute Leistungen – erbringen. Kugelstoßen und Speerwurf sind sicherlich auch anstrengend, aber vermutlich werden sie keine gute Indikatoren für die Ausdauerleistung sein.

Auch für die Faktorenanalyse müssen die Variablen quantitativ sein. Die Faktorenanalyse leistet folgendes: Sie überprüft zuerst, ob die Variablen (die sportlichen Leistungen) auf einem, oder zwei oder mehreren Hintergrundfaktoren messen. Wir werden dabei erkennen, daß die ersten vier Sportarten überwiegend auf einem gemeinsamen Hintergrundfaktor messen (der Ausdauerleistung). Kugelstoßen und Sperrwurf sind Indikatoren für einen zweiten Hintergrundfaktor, den man „Kraft“ nennen könnte.

Wir werden die Variablen Kugelstoßen und Sperrwurf ausschließen und die Faktorenanalyse fortsetzen. Die Faktorenanalyse errechnet nun das Gewicht mit dem die verbliebenen vier Sportarten jeweils auf dem Hintergrundfaktor „Ausdauerleistung“ liegen und ermittelt mit Hilfe dieser Gewichtszahlen einen „gewichteten“ Gesamtpunktwert für die Ausdauerleistung. In der Sprache der Faktorenanalyse heißen die Gewichtszahlen „Faktorladungen“ und die gewichteten Gesamtpunktwerte „Faktorwerte“.

Die Faktorenanalyse würde auch folgende Vorgehensweise erlauben: Kugelstoßen und Sperrwurf werden nicht aus der Analyse ausgeschlossen. Die Faktorenanalyse ermittelt dann 2 Hintergrundfaktoren, die wir „Ausdauerleistung“ und „Kraft“ nennen und errechnet für jede der sechs Sportarten je zwei Gewichtszahlen, eine für die Ausdauerleistung und eine für die Kraft. Dabei wird beispielsweise der 5000 m Lauf auf dem Faktor „Ausdauer“ eine hohe Gewichtszahl und auf dem Faktor „Kraft“ eine sehr niedrige Gewichtszahl besitzen. Und umgekehrt wird z.B. Kugelstoßen auf dem Faktor „Ausdauer“ eine sehr niedrige Gewichtszahl und auf dem Faktor „Kraft“ eine sehr hohe Gewichtszahl besitzen. Dann errechnet die Faktorenanalyse aus allen sechs Sportarten zwei „gewichtete“ Gesamtpunktwerte, einen für die Ausdauer und einen für die Kraft.

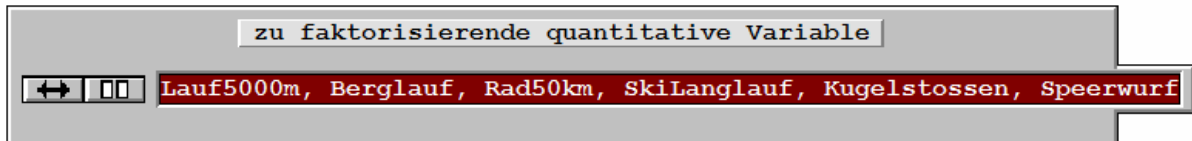
Eingabe in das Almo-Programm Prog45ms

P45.10.1 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: Zu faktorisierte quantitative Variable



zu faktorisierte quantitative Variable

↔ [] Lauf5000m, Berglauf, Rad50km, SkiLanglauf, Kugelstossen, Speerwurf

Geben Sie hier die Variable an, die faktorisiert werden sollen. Die Variablen müssen quantitativ sein. Siehe dazu P45.12.0.

Eingabe-Box 7: Option: Ein- und Ausschließen von Untersuchungseinheiten.

Wenn Sie diese OptionsEingabe-Box öffnen, dann wird es Ihnen ermöglicht, bestimmte Datensätze aus der Analyse auszuschließen bzw. in die Analyse einzuschließen. In "Arbeiten mit Almo-Datenanalyse-System", Abschnitt P0.7 haben wir das ausführlich dargestellt.

Nun ist jedoch folgendes zu berücksichtigen. Wir wollen das an einem Beispiel zeigen: Sie wollen nur Männer in die Faktorenanalyse einschließen. Die Faktorladungen und alle anderen Ergebnisse, die Almo liefert, beziehen sich also nur auf Männer. Wenn Sie nun Faktorwerte berechnen wollen (siehe dazu weiter unten) und diese als zusätzliche Variable an die Datensätze anhängen wollen und in eine neue Datei speichern wollen, dann geschieht dies nur für die "männlichen" Datensätze. Die neue Datei besteht dann nur aus den Männern.

Eingabe-Box 8: Kein-Wert-Angabe und Umkodierungen

In P0.5 haben wir ausführlich dargestellt, wie umkodiert wird. Wenn Sie das Programm Prog45ms auch dazu verwenden wollen, Faktorwerte zu bilden und in eine neue Datei zu speichern, dann ist zu berücksichtigen, dass umkodierte Variable in ihrer umkodierten Form in die neue Datei eingehen.

Eingabe-Box 9: Option: Faktorwerte ermitteln und speichern



↓ [] Option: Faktorwerte ermitteln und speichern

Sie können für jeden Sportler "Faktorwerte" (wie wir sie oben beschrieben haben) bilden und als zusätzliche Variable an seinen Datensatz anhängen. Man wird dies aber erst dann tun, wenn man durch (eventuell mehrere) vorausgehende Analysen Klarheit über die Faktorenstruktur gewonnen hat. Vor allem muss man wissen, wieviel Faktoren sinnvollerweise extrahiert werden sollen.

Wenn Sie die Optionsbox öffnen dann sehen Sie folgendes

Loesche wieder diese Box (dann Voreinstellungen wieder gueltig)

Option: Faktorwerte ermitteln und speichern

Zahl der Faktorwertvariablen,
die gebildet werden sollen

die Faktorwert-Variablen transformieren
z.B. auf 3 Dezimalstellen runden
das ist nicht obligatorisch

Maximal x % der faktorisierten Variablen,
dürfen Kein_Wert besitzen. Für sie wird
der Mittelwert eingesetzt.
Sonst werden die Faktorwert-Variable
auf Kein_Wert gesetzt

Geben Sie einen neuen Dateinamen
ohne Erweiterung an
Almo erzeugt 3 Dateien:

1. eine nicht lesbare Almo-Arbeitsdatei
mit der Erweiterung ____.dir
2. eine anschaulbare Datei im freien Format
mit der Erweiterung ____.fre
3. eine Datei der Variablennamen
mit der Erweiterung ____.nam

In den unter 1. und 2. angegebenen neuen Dateien sind nun enthalten:

- die unveränderten Variablen aus der alten Datei
- die Faktorwertvariable, die als letzte Variablen
hinter die Variablen der alten Datei gestellt wurden
was wird gespeichert ? ----->

In der unter 3. angegebenen neuen Datei der Variablennamen sind nun
enthalten:

- die Variablennamen aus der alten Datei einschliesslich der
in der Box "Freie Namensfelder" angegebenen Namen
- die Namen "Fakwert.." für die neuen angehängten Faktorwert-
variablen. Anstelle der 2 Punkte schreibt Almo die Variablen-
nummern der neuen Variablen, also z.B. "Fakwert7", "Fakwert8"
bitte lesen ----->

Eingabefeld 1: Geben Sie die Zahl der "Faktorwert-Variablen" an, die gebildet werden sollen. In unserem Beispiel sollen 2 "Faktorwert-Variable" gebildet werden, eine für "Kraft" und eine für "Ausdauer".

Eingabefeld 2: Die Faktorwert-Variablen besitzen Werte von ca. -4 bis +4. Sie können mehrere signifikante Kommastellen haben. Es ist nun durchaus zulässig die Faktorwert-Variablen linear zu transformieren. In unserem Beispiel haben wir die Werte auf die 3 Kommastelle gerundet. Es könnte aber auch +4 addiert werden (damit keine negativen Werte auftreten), dann mit 10 multipliziert und dann auf

Ganzzahl gerundet werden (damit keine Dezimalwerte auftreten). Das kann man machen, muss es aber nicht.

Eingabefeld 4: Geben Sie einen Namen für die neue Datei an. In diese werden die seitherigen Variablen geschrieben, sowie die neu gebildeten "Faktorwert-Variablen". Geben Sie den Dateinamen ohne Erweiterung an. Almo erzeugt dann 3 Dateien:

1. eine nicht lesbare Almo-Arbeitsdatei mit der Erweiterung `__.dir`
2. eine anschauliche Datei im freien Format mit der Erweiterung `__.fre`
3. und eine Datei der Variablennamen.

Klicken Sie auf die Hilfe-Knöpfe in dieser Sektion der Eingabe-Box und lesen Sie die ausführlichen Erläuterungen

Lesen Sie nochmals unsere obigen Erläuterungen zu den Eingabe-Boxen 7 und 8. Werden Datensätze ein- oder ausgeschlossen oder werden Variable umkodiert, dann wirkt sich das auch auf die Faktorwerte und deren Abspeichern in einer neuen Datei aus.

P45.10.2 Ausgabe aus Prog45ms

Almo liefert eine sehr lange Ergebnisliste. Wir werden im folgenden nur die wirklich bedeutsamen Ergebnisteile erläutern. Eine vollständige Besprechung der Ergebnisliste ist im Handbuch "P30 Faktorenanalyse, nomiale Faktorenanalyse, multiple Korrespondenzanalyse" enthalten.

Almo berechnet zuerst die Interkorrelationsmatrix der Variablen. Auf diese wird dann der Kalkül der Faktorenanalyse angewendet. Almo gibt deswegen zuerst die Korrelationsmatrix (und einige zusätzlichen Ergebnisse) aus.

Fuer Analyse ausgewaehlte Variable

```
V1    Lauf5000m
V2    Berglauf
V3    Rad50km
V4    SkiLanglauf
V5    Kugelstossen
V6    Speerwurf
```

Zahl der insgesamt eingelesenen Einheiten 500

Zahl der in die Analyse einbezogenen Einheiten 500

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

		Lauf5000	Berglauf	Rad50km	SkiLangl	Kugelsto	Speerwur
		V1	V2	V3	V4	V5	V6
Lauf5000	V1	500	500	500	500	500	500
Berglauf	V2	500	500	500	500	500	500
Rad50km	V3	500	500	500	500	500	500
SkiLangl	V4	500	500	500	500	500	500
Kugelsto	V5	500	500	500	500	500	500
Speerwur	V6	500	500	500	500	500	500

harmonisches Mittel der in die Analyse einbezogenen Einheiten
(ohne Diagonale) - wird fuer Signifikanztest verwendet 500

Standardabweichungen
 (Standardabweichung ist mit n
 nicht mit n-1 dividiert)

Lauf5000	V1	4.1282
Berglauf	V2	4.7040
Rad50km	V3	7.5887
SkiLangl	V4	4.5187
Kugelsto	V5	1.4583
Speerwur	V6	5.1090

Mittelwerte

Lauf5000	V1	47.5080
Berglauf	V2	44.4000
Rad50km	V3	79.4040
SkiLangl	V4	35.6880
Kugelsto	V5	13.8680
Speerwur	V6	47.3740

Korrelations-Matrix

		Lauf5000	Berglauf	Rad50km	SkiLangl	Kugelsto	Speerwur
		V1	V2	V3	V4	V5	V6
Lauf5000	V1	1.0000	0.6172	0.4018	0.4600	-0.1025	-0.2347
Berglauf	V2	0.6172	1.0000	0.4305	0.4558	0.0071	-0.0796
Rad50km	V3	0.4018	0.4305	1.0000	0.5379	-0.0617	-0.1701
SkiLangl	V4	0.4600	0.4558	0.5379	1.0000	-0.1371	-0.0815
Kugelsto	V5	-0.1025	0.0071	-0.0617	-0.1371	1.0000	0.4096
Speerwur	V6	-0.2347	-0.0796	-0.1701	-0.0815	0.4096	1.0000

******* Erläuterung:**

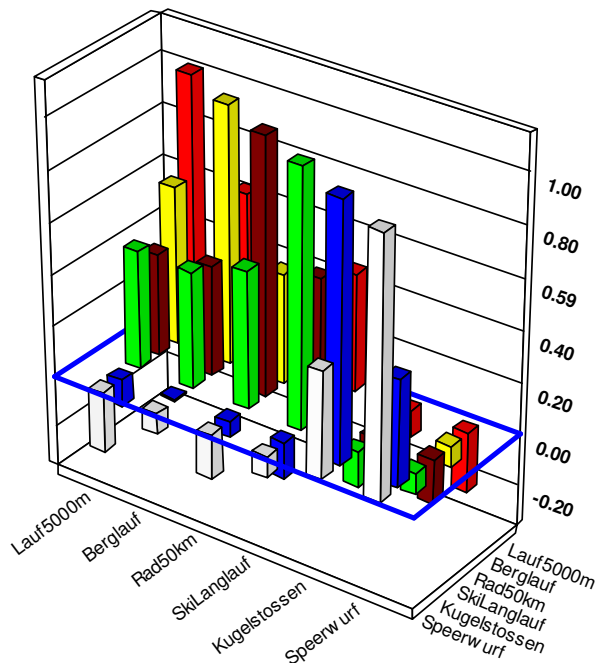
Man erkennt schon an den Korrelationskoeffizienten, dass die 4 Ausdauer-Sportarten und die 2 Kraft-Sportarten untereinander gut korrelieren aber gegeneinander schwach korrelieren.

Mindestgroesse des Produkt-Moment-Korrelationskoeffizienten r	bei Signifikanz (1-p)*100	fuer df=n-2=500-2=498
0.0574	80	
0.0645	85	
0.0736	90	
0.0796	92.5	
0.0876	95	
0.1002	97.5	
0.1149	99	
0.1483	99.9	

***** MITTEILUNG
 Fuer Analyse-Variable mit "Kein_Wert"
 wurde zur Berechnung der Streuungsmatrix
 folgende Kein-Wert-Behandlung durchgefuehrt

Kein-Wert-Behandlung=1: "Paarweises Ausscheiden"

Almo zeichnet die Korrelationsmatrix als Balkendiagramm.
Korrelations-Matrix



Ergebnisse aus Faktorenanalyse

Eingesetzte Werte fuer Kommunalitaeten
(Multiple Bestimmtheitsmasse)

Lauf5000	V1	0.4545
Berglauf	V2	0.4443
Rad50km	V3	0.3513
SkiLangl	V4	0.3931
Kugelsto	V5	0.1910
Speerwur	V6	0.2261

*******Erläuterung:**

Die Faktorenanalyse im Rahmen des Data Mining verwendet die „multiplen Bestimmtheitsmaße“ R^2 als Kommunalitätenschätzungen. Der Vorteil dieser Schätzmethode ist, daß eine minimale Faktorenzahl entsteht. Siehe dazu die ausführliche Darstellung im Handbuch „P30 Faktorenanalyse“.

Zahl der Kommunalitaeten-Iterationen: 0

Koeffizienten fuer Faktoren

Eigenwerte (Varianz je Faktor)
1.9414 0.5728 -0.2862

Prozent der Varianz
32.3560 9.5470

Zu erklaerende Gesamtvarianz= 6.0000
Durch 2 Faktoren erklarte Varianz= 2.5142
Prozentsatz der erklarten Varianz= 41.9030

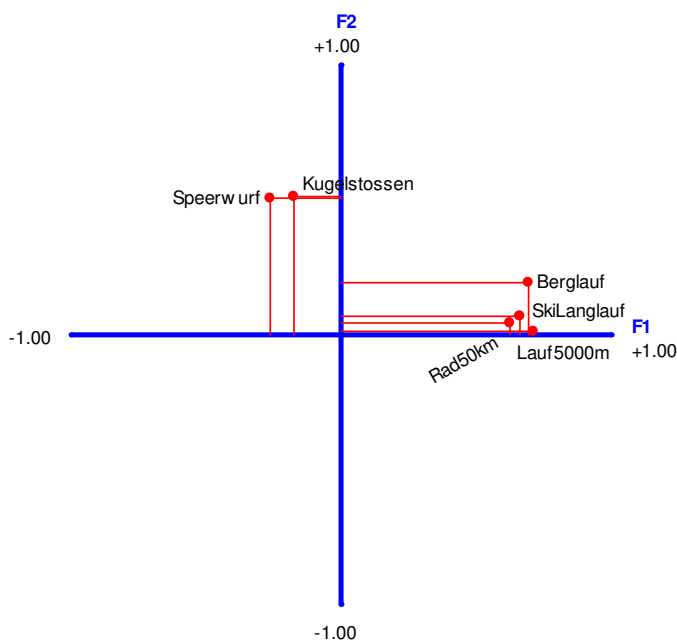
Matrix der Faktorladungen

		Faktor 1	Faktor 2
Lauf5000	V1	0.7187	0.0161
Berglauf	V2	0.6938	0.1992
Rad50km	V3	0.6303	0.0491
SkiLangl	V4	0.6664	0.0693
Kugelsto	V5	-0.1761	0.5145
Speerwur	V6	-0.2667	0.5108

******* Erläuterung:**

Dies sind die unrotierten Faktorladungen. Sie sind, sofern 2 oder mehr Faktoren extrahiert wurden, eher schlecht zu interpretieren. Die grafische Darstellung dieser Matrix liefert jedoch ein eindrucksvoll klares Bild der ermittelten Faktorenstruktur. Also liefert folgende Grafik:

Faktorladungen



******* Erläuterung:**

Die Grafik zeigt 2 "Punktewolken", die der beiden Kraft-Sportarten "Kugelstoßen" und "Speerwurf" und die der 4 Ausdauer-Sportarten.

Kommunalitaeten je Variable

Lauf5000	V1	0.5168
Berglauf	V2	0.5210
Rad50km	V3	0.3997
SkiLangl	V4	0.4489
Kugelsto	V5	0.2957
Speerwur	V6	0.3321

Multiple Faktorbestimmtheit

(Quadrierte multiple Korrelation zwischen Variablen und Faktor)

Faktor 1 0.7664
 Faktor 2 0.4275

Unrotierte Faktor-Betaladungen (Faktorwert-Koeffizienten)
 (Regressions-Loesung)

		Faktor 1	Faktor 2
Lauf5000	V1	0.3079	-0.0159
Berglauf	V2	0.2814	0.2016
Rad50km	V3	0.2278	0.0295
SkiLangl	V4	0.2584	0.0498
Kugelsto	V5	-0.0613	0.3630
Speerwur	V6	-0.0872	0.3836

Schiefwinklige Rotation mit 2 Faktoren

Quartimin-Kriterium 0.0453
 letzte Iterationsdifferenz bei Quartimin-Rotation 0.0000

Variablengruppe 1: V1 V2 V3 V4

Variablengruppe 2: V5 V6

******* Erläuterung:**

In obiger Grafik der (unrotierten) Faktorladungsmatrix wird ersichtlich, dass die 6 Variablen in 2 "Punktewolken", anders formuliert: in 2 Variablengruppen zerfallen. Also identifiziert diese beiden Variablengruppen.

Wie diese beiden Punktewolken, bzw. Variablengruppen zu bezeichnen sind, überläßt Also dem Forscher. Wir haben uns entschieden die 1. Variablengruppe als "Ausdauer" und die 2. als "Kraft" zu bezeichnen.

Also legt nun je eine Achse mittern durch diese beiden Punktewolken. Diesen Vorgang nennt man "schiefwinklige Rotation". Wir werden dies weiter unten grafisch darstellen.

Durch zugehoerige Achse erklarte Varianz je Variablengruppe
 1.8675 0.6242

Matrix der Korrelationen zwischen den schiefwinkligen Achsen

		Faktor 1	Faktor 2
Faktor 1		1.0000	-0.2820

Faktor 2	-0.2820	1.0000
----------	---------	--------

Matrix der Winkel zwischen den schiefwinkligen Achsen

	Faktor 1	Faktor 2
Faktor 1	0	-73.6185
Faktor 2	-73.6185	0

******* Erläuterung:**

Die beiden schiefwinkligen Koordinatenachsen (siehe Grafik weiter unten) stehen in einem Winkel von -73 Grad, d.h. in einem stumpfen Winkel von $90 + (90-73) = 107$ Grad aufeinander. Der Cosinus dieses Winkels kann als Korrelation (von -0.2820) zwischen den beiden Achsen, d.h. zwischen der "Ausdauer" und der "Kraft" interpretiert werden.

Transformationsmatrix
von orthogonaler Faktorladungsmatrix zu nachfolgender Strukturmatrix

	Faktor 1	Faktor 2
Faktor 1	0.9923	-0.3985
Faktor 2	0.1237	0.9172

Matrix der auf die schiefwinkligen Achsen
rechtwinklig projizierten Faktorladungen
(Strukturmatrix)

		Faktor 1	Faktor 2
Lauf5000	V1	0.7152	-0.2717
Berglauf	V2	0.7131	-0.0938
Rad50km	V3	0.6315	-0.2062
SkiLangl	V4	0.6698	-0.2020
Kugelsto	V5	-0.1111	0.5420
Speerwur	V6	-0.2015	0.5748

Transformationsmatrix
von orthogonaler Faktorladungsmatrix zu nachfolgender Ladungsmatrix

	Faktor 1	Faktor 2
Faktor 1	0.9560	-0.1289
Faktor 2	0.4154	1.0343

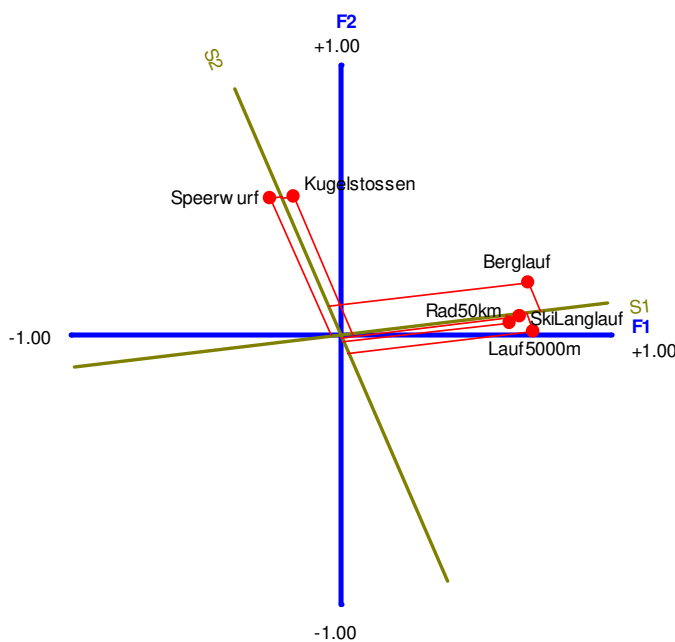
Matrix der auf die schiefwinkligen Achsen
achsparell projizierten Faktorladungen
(Ladungsmatrix)

		Faktor 1	Faktor 2
Lauf5000	V1	0.6938	-0.0760
Berglauf	V2	0.7460	0.1166
Rad50km	V3	0.6229	-0.0305
SkiLangl	V4	0.6658	-0.0143
Kugelsto	V5	0.0454	0.5548
Speerwur	V6	-0.0428	0.5628

***** **Erläuterung:**

Wir haben die maximalen Ladungen je Zeile fett gedruckt. In dieser Matrix ist nun deutlich zu erkennen, dass V1 bis V4 mit hohen Ladungen auf dem 1. (schiefen) Faktor und V5 sowie V6 auf dem 2. (schiefen) Faktor liegen. Diese Ladungsmatrix wird von Almo grafisch dargestellt:

Faktorladungen im recht- und schiefwinkligen Koordinatensystem (achsparelle Projektion)



***** **Erläuterung:**

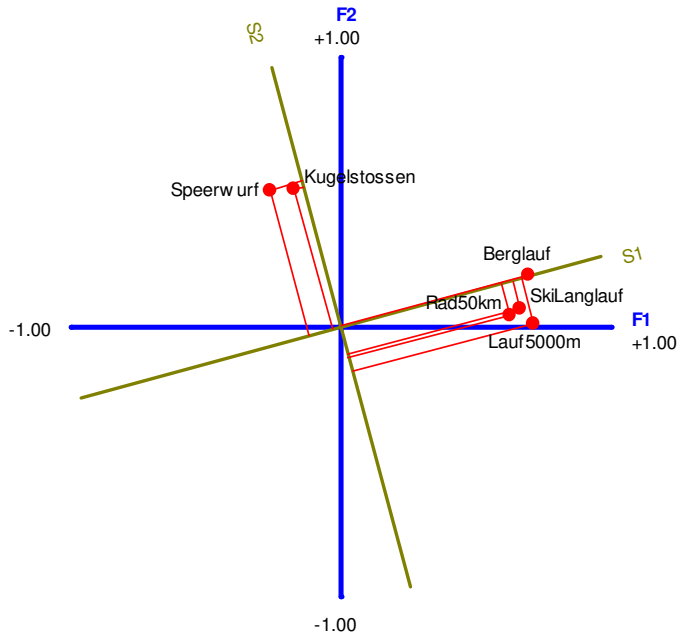
Almo legt je eine Achse mitten (genauer: varianzmaximierend) durch die beiden Punktwolken. In obiger Grafik werden sie mit S_1 und S_2 bezeichnet. Diesen Vorgang nennt man "schiefwinklige Rotation" - "Rotation" deswegen, weil das Koordinatensystem gedreht wird und "schiefwinklig" weil der rechte Winkel zwischen den Koordinatenachsen nicht beibehalten wird. In unserem Beispiel stehen die Achsen in einem stumpfen Winkel von 107 Grad aufeinander.

Die Variablenpunkte werden "achsparell" auf die schiefen Achsen projiziert. Weiter oben in der Ausgabeliste wurde die sogenannte "Strukturmatrix" angegeben. Bei ihr werden die Variablenpunkte rechtwinklig auf die schiefen Achsen projiziert. Das dabei entstehende grafisch Abbild ist nicht so anschaulich, wie das bei der achsparellen Projektion.

Almo rotiert standardmäßig "schiefwinklig". Natürlich ist es auch möglich, "rechtwinklig" zu rotieren. Dabei verlaufen aber die Achsen in aller Regel nicht mehr

mitten durch die Punktwolken. In unserem Beispiel würde sich folgende Grafik ergeben:

Faktorladungen im orthogonalen und im Varimax-Koordinatensystem



***** **Erläuterung:**

Die beiden rechtwinklig rotierten Achsen S₁ und S₂ laufen jetzt nicht mehr „mitten durch“ die beiden Punktwolken.

Multiple Faktorbestimmtheit

(Quadrierte multiple Korrelation zwischen Variablen und Faktor)

Faktor 1 0.7646

Faktor 2 0.4711

Rotierte Faktor-Betaladungen (Faktorwert-Koeffizienten)

(Regressions-Loesung)

		Faktor 1	Faktor 2
Lauf5000	V1	0.3036	-0.1372
Berglauf	V2	0.3042	0.0727
Rad50km	V3	0.2297	-0.0637
SkiLangl	V4	0.2626	-0.0573
Kugelsto	V5	-0.0160	0.3574
Speerwur	V6	-0.0390	0.3865

***** **Erläuterung:**

Diese Matrix der "Faktorwert-Koeffizienten" wird nur ausgegeben, wenn der Benutzer, die Optionsbox "Faktorwerte ermitteln und speichern" aktiviert hat, d.h. wenn er Faktorwerte je Untersuchungseinheit ermitteln und speichern möchte. Wir haben das getan und angegeben, dass wir 2 Faktorwertvariable bilden möchten, eine für "Ausdauer" und eine für "Kraft". Also bildet nun aus den rotierten Faktorladungen aus obiger "Ladungsmatrix" durch eine einfache lineare Transformation die Matrix der "Faktorwert-Koeffizienten". Diese werden dazu verwendet Faktorwerte zu bilden, gemäß folgender Formel:

$$FW1 = 0.3036*V1 + 0.3042*V2 + 0.2297*V3 + 0.2626*V4 - 0.0160*V5 - 0.0390*V6$$

$$FW2 = -0.1372*V1 + 0.0727*V2 - 0.0637*V3 - 0.0573*V4 + 0.3574*V5 + 0.3865*V6$$

FW1 =Faktorwertvariable 1 (=Ausdauer)

FW2 =Faktorwertvariable 2 (=Kraft)

Es wird also ein gewichteter Gesamtpunktwert gebildet, bei dem die Faktorwert-Koeffizienten als Gewichtungszahlen verwendet werden. Die Variablenwerte von V1 bis V6 müssen standardisiert werden. Von ihrem Variablenwert wird der Variablen-Mittelwert subtrahiert, dann wird mit der Standardabweichung der Variablen dividiert.

Diese Faktorwertvariable haben wir als V7 und V8 an das Ende jedes Datensatzes gestellt und die so verlängerten Datensätze in die neue Datei "Fakwert.dir" und "Fakwert.fre" gespeichert. Zur Illustration geben wir die ersten 5 Datensätze aus:

Datensatz	V1	V2	V3	V4	V5	V6	V7	V8
1	44	46	82	34	17	44	-0.1825925	0.6532379
2	55	51	79	42	16	47	1.3118200	0.2704992
3	46	41	75	33	15	48	-0.6374510	0.3934108
4	46	45	79	37	13	47	0.0042991	-0.1948428
5	44	44	89	35	12	43	0.0205175	-0.7501247

Die Faktorwert-Variablen besitzen Werte von ca. -4 bis +4. Sie können mehrere signifikante Kommastellen haben. Es ist nun durchaus zulässig die Faktorwert-Variablen linear zu transformieren. In unserem Beispiel haben wir in der Optionsbox "Faktorwerte ermitteln und speichern" +4 addiert (damit keine negativen Werte auftreten), dann mit 10 multipliziert und dann auf Ganzzahl gerundet (damit keine Dezimalwerte auftreten). Das kann man machen, muss es aber nicht. Für V7 und V8 haben wir dabei erhalten

V7	V8
38	47
53	43
34	44
40	38
40	32

Betrachten wir die V8 (=Kraft) beim ersten und zweiten Datensatz. Person 1 ist bei den unveränderten Faktorwerten in "Kraft" mehr als doppelt so stark wie Person 2. Durch das Hinzufügen von +4 wird diese Relation auf 47 zu 43 verringert. Das Hinzuaddieren einer Zahl (damit negative Vorzeichen verschwinden) ist also eher nicht günstig (aber nicht falsch).

Matrix der räumlichen Entfernungen der Variablen von den schiefwinkligen Achsen

		Faktor 1	Faktor 2
Lauf5000	V1	0.0729	0.6656
Berglauf	V2	0.1119	0.7157
Rad50km	V3	0.0293	0.5976
SkiLangl	V4	0.0137	0.6388
Kugelsto	V5	0.5323	0.0436
Speerwur	V6	0.5399	0.0411

Räumliche Geschlossenheit der Variablengruppen

(mittlere Entfernung der Variablen einer Variablen-Gruppe von "ihrer" Achse)

0.0569 0.0423

```
***** MITTEILUNG
Lesen oder Schreiben korrekt beendet in Datei
"C:\Almo6\PROGS\Fakwert.fre"
```

```
***** MITTEILUNG
Lesen oder Schreiben korrekt beendet in Datei
"C:\Almo6\PROGS\Fakwert.dir"
```

******* Erläuterung:**

Almo teilt zum Schluss noch mit, dass es 2 neue Dateien (mit gleichem Inhalt) angelegt hat.

P45.10.3 Beispiel für eine Faktorenanalyse aus der Umfragesforschung

Betrachten wir ein sozialwissenschaftliches Beispiel. Die politische Aktionsbereitschaft von Studenten soll durch eine Befragung ermittelt werden. Folgende Fragebatterie, bestehend aus 5 Fragen, wird den Studenten vorgelegt:

7	6	5	4	3	2	1
stimme	stimme	stimme	bin	lehne	lehne	lehne
sehr stark	stark	zu	unent-	ab	stark	sehr stark
zu	zu		schieden		ab	ab
-----	-----	-----	-----	-----	-----	-----

1. Man sollte sich an einer Unterschriftenaktion gegen xxx beteiligen
2. Man sollte wegen xxx einen Brief an die Politiker schreiben
3. Man sollte gegen xxx an einer Demonstration teilnehmen
4. Man sollte gegen xxx an einer Bürgerinitiative mitarbeiten
5. Man sollte wegen xxx an der Besetzung eines öffentlichen Gebäudes teilnehmen

In diesem Falle möchte man, dass alle 5 Fragen auf nur einem gemeinsamen Faktor messen. Die Faktorenanalyse wird dazu eingesetzt, die gewünschte Eindimensionalität, nachzuweisen.

Eingabe in Prog45ms

Wir zeigen nur die Eingabe in die Eingabe-Boxen 3 bis 6

Variablenamen

Datei der Variablenamen

"C:\Almo15\TESTDAT\DM_FBatt.nam"

zeige zeige = Namensdatei in Output zeigen
 leer = nicht zeigen

Datei aus der gelesen wird

"C:\Almo15\TESTDAT\DM_FBatt.dir"

zu faktorisierte quantitative Variable

V1:5

BEACHTEN: In der Eingabe-Box "Zu faktorisierte quantitative Variable" haben wir nicht die Variablenamen der 5 Variablen geschrieben, sondern kurz die Variablennummern "V1:5". Es ist prinzipiell möglich, Variablenamen und Variablennummern auszutauschen, sogar zu mischen.

Da wir auch Faktorwerte ermitteln wollen und diese als zusätzliche (Faktorwert-) Variable an jeden Datensatz abhängen wollen, wird die Eingabe-Box "Option: Faktorwerte ermitteln und speichern" geöffnet und in folgender Weise ausgefüllt:

Loesche wieder diese Box (dann Voreinstellungen wieder gueltig)

Option: Faktorwerte ermitteln und speichern

Zahl der Faktorwertvariablen,
die gebildet werden sollen

die Faktorwert-Variablen transformieren
z.B. auf 3 Dezimalstellen runden
das ist nicht obligatorisch

Maximal x % der faktorisierten Variablen,
dürfen Kein_Wert besitzen. Für sie wird
der Mittelwert eingesetzt.
Sonst werden die Faktorwert-Variable
auf Kein_Wert gesetzt

Geben Sie einen neuen Dateinamen
ohne Erweiterung an
Almo erzeugt 3 Dateien:

1. eine nicht lesbare Almo-Arbeitsdatei
mit der Erweiterung ____.dir
2. eine anschaubare Datei im freien Format
mit der Erweiterung ____.fre
3. eine Datei der Variablennamen
mit der Erweiterung ____.nam

In den unter 1. und 2. angegebenen neuen Dateien sind nun enthalten:

- die unveränderten Variablen aus der alten Datei
- die Faktorwertvariable, die als letzte Variablen
hinter die Variablen der alten Datei gestellt wurden
was wird gespeichert ? ----->

In der unter 3. angegebenen neuen Datei der Variablennamen sind nun
enthalten:

- die Variablennamen aus der alten Datei einschliesslich der
in der Box "Freie Namensfelder" angegebenen Namen
- die Namen "Fakwert.." für die neuen angehängten Faktorwert-
variablen. Anstelle der 2 Punkte schreibt Almo die Variablen-
nummern der neuen Variablen, also z.B. "Fakwert7", "Fakwert8"
bitte lesen ----->

Ausgabe

Almo liefert folgendes Ergebnis, das wir hier nur stark gekürzt wiedergeben.

Korrelations-Matrix

		Untersch	BriefanP	Demonstr	Buergeri	Besetzun
		V1	V2	V3	V4	V5
Untersch	V1	1.0000	0.3606	0.5616	0.4515	0.7230
BriefanP	V2	0.3606	1.0000	0.3013	0.1828	0.3954
Demonstr	V3	0.5616	0.3013	1.0000	0.4713	0.5117
Buergeri	V4	0.4515	0.1828	0.4713	1.0000	0.4739
Besetzun	V5	0.7230	0.3954	0.5117	0.4739	1.0000

Matrix der Faktorladungen

Untersch	V1	0.8097
BriefanP	V2	0.4367
Demonstr	V3	0.6722
Buergeri	V4	0.5782
Besetzun	V5	0.8058

Wie gewünscht, entsteht nur ein Faktor. Natürlich hätten auch mehrere Faktoren entstehen können. Dann hätte man solange Variable eliminiert, bis eine einfaktorielle Lösung entstanden wäre.

Unrotierte Faktor-Betaladungen (Faktorwert-Koeffizienten)
(Regressions-Loesung)

Untersch	V1	0.3469
BriefanP	V2	0.0858
Demonstr	V3	0.2059
Buergeri	V4	0.1443
Besetzun	V5	0.3473

Dies sind die Faktorwert-Koeffizienten, die als "Gewichtszahlen" verwendet werden, um die Faktorwertvariable V6 als "gewichteter Gesamtpunktwert" zu bilden. Die Faktorwert-Koeffizienten entstehen durch eine lineare Transformation der Faktorladungen.

Almo schreibt dann die Faktorwertvariable als V6 in die neue Datei ein. Almo teilt dann noch mit:

```
***** MITTEILUNG
Lesen oder Schreiben korrekt beendet in Datei
"C:\Almo6\PROGS\SozFakwert.fre"

***** MITTEILUNG
Lesen oder Schreiben korrekt beendet in Datei
"C:\Almo6\PROGS\SozFakwert.dir"
```

Durch Doppelklick auf "C:\Almo6\PROGS\SozFakwert.fre" in obiger Mitteilung kann man dann die neue Datei in ein Fenster laden und anschauen.

P45.10.4 Weiterführende Hinweise

Die Faktorenanalyse ist als Programm 30 in Almo mit vielfältigen Optionen enthalten. Im Handbuch „P30 Faktorenanalyse, nominale Faktorenanalyse,

multiple Korrespondenzanalyse“ wird die Almo-Faktorenanalyse ausführlich dargestellt.

Einführungen und Gesamtdarstellungen der Faktorenanalyse sind zu finden bei Arminger (1979), Holm (1976) und Überla (1968).

P45.11 Schritt 7c: Rasch-Skalierungsverfahren

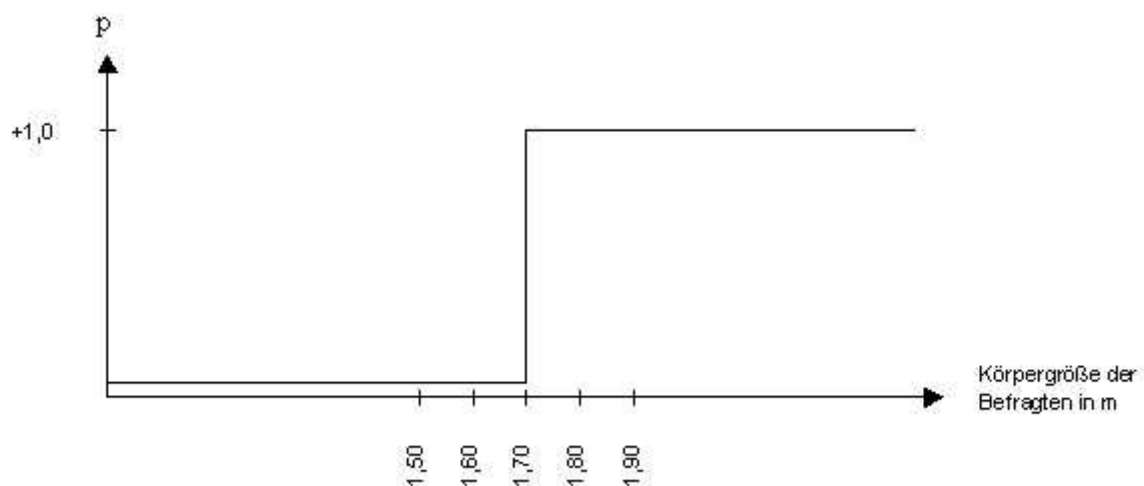
Das Rasch-Skalierungsverfahren ist in Almo in doppelter Form vorhanden. Einmal als "dichotomes Rasch-Modell" (für dichotome Items) und einmal als "allgemeines ordinale Rasch-Modell" (für dichotome und polytome Items). Das letztere inkludiert das "dichotome Rasch-Modell". Es bietet dem (erfahrenen) Benutzer auch vielfältige Optionen an und liefert umfassende Ergebnisse. Es wird in dem Almo-Handbuch "Das allgemeine ordinale Rasch-Modell" ausführlich dargestellt. In Almo wird es gefunden durch Klick auf den Knopf "Vefahren/Rasch-Modell", dann Prog14m3 oder Prog14m4.

Wir werden im folgenden das "dichotome Rasch-Modell" behandeln. Es kann als eine Weiterführung des Guttman-Skalierungsverfahren betrachtet werden. Wir wollen deswegen das Guttman-Verfahren zuerst kurz darstellen. Siehe dazu auch die ausführliche Darstellung von Joachim Gerich im Almo-Handbuch „Sozialwissenschaftliche Skalierungsverfahren“, Abschnitt P16.

Eine Skala, die nach dem von L. Guttman entwickelten Verfahren konstruiert werden soll, muß aus sogenannten "monotonen Fragen" zusammengesetzt sein. Eine monotone Frage besitzt die Eigenschaft, die Untersuchungseinheiten, die ein quantitatives Merkmal in verschiedenen Ausprägungen besitzen an einer bestimmten Trennstelle in die zwei Untergruppen der "Ja"-Sager und "Nein"-Sager aufzuteilen. Ein Beispiel wäre die Frage:

"Sind Sie größer als 1,70 m?"

Befragte bis zur Größe von 1,70 m antworten mit Nein (Wahrscheinlichkeit einer bejahenden Antwort $p = 0$), alle anderen mit Ja ($p = 1$). Graphisch dargestellt:



Die p-Kurve, auch "trace line" genannt, steigt an einem bestimmten Punkt von 0 auf 1 und bleibt dann auf gleicher Höhe. Die Kurve "steigt monoton". Die

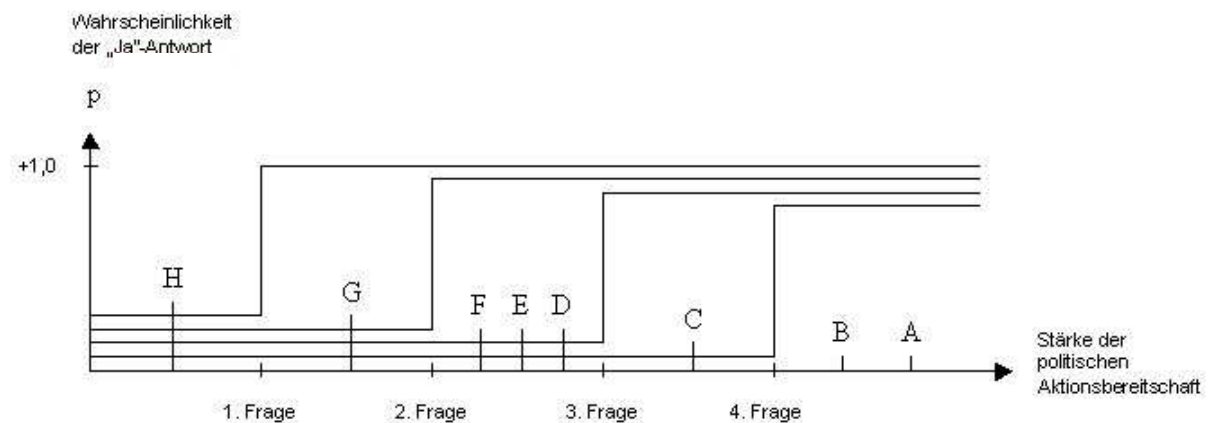
Trennstelle (hier 1,70 m) wird als "Ort" der Frage auf dem Merkmalskontinuum begriffen.

Bei der Guttman'schen Skalierungstechnik wird ein "gemeinsames Kontinuum" von Befragten und Fragen – eine "joint scale" – angenommen. Über das Kontinuum sind sowohl die Befragten verteilt – nach der Menge an jener Eigenschaft, die es zu messen gilt – als auch die Fragen – nach dem Ort, an dem sie die Befragten in Ja- und Nein-Sager trennen.

Betrachten wir folgendes sozialwissenschaftliches Beispiel:

Es soll die politische Aktivitätsbereitschaft von Studenten ermittelt werden. Dazu werden folgende 4 Fragen gestellt:

1. An Volksbegehren teilnehmen
2. An Demonstrationen teilnehmen
3. Bei Bürgerinitiative aktiv mitmachen
4. In einer politischen Partei aktiv mitarbeiten



Wir haben die Kurven so gezeichnet, daß jede einzelne von ihnen sichtbar wird. Tatsächlich beginnen alle Kurven bei $p = 0$ und verlaufen dann nach dem "Treppensprung" mit $p = 1$.

Greifen wir die 3. Frage heraus. Sie hat einen Skalenwert von x auf der Meßdimension „Stärke der politischen Aktionsbereitschaft“. Alle Personen, die in ihrer „politischen Aktionsbereitschaft“ kleiner sind als x , kürzer formuliert, die einen Skalenwert kleiner x haben, müssen die 3. Frage verneinen. Dies sind die Personen H, G, F, E, D. Alle Personen die in ihrer „politischen Aktionsbereitschaft“ größergleich x sind, müssen sie bejahen. Dies sind die Personen C, B, A.

Betrachten wir die 3 Personen F, E, D. Sie bejahen die 1. und 2. Frage, die 3. und 4. Frage ist ihnen jedoch zu "radikal". Sie verneinen sie. Die Personen A und B hingegen weisen eine höhere politische Aktivitätsbereitschaft auf: Sie bejahen alle vier Items.

Der Skalenwert einer Person auf der Meßdimension, und diesen suchen wir ja, ergibt sich nun sehr einfach als die Zahl der Ja-Antworten. Nun kann ein Problem auftreten. Betrachten wir Person C. Die Frage 4 wird von ihr verneint und die Frage 1, 2 und 3 bejaht. Bei empirischen Daten kann es aber nun geschehen, daß die Person C, obwohl sie höher liegt als x , die 3. Frage verneint. Damit ist ein „Fehler“ entstanden. Wenn mehrere Personen, die in ihrem Skalenwert höher als x liegen,

fälschlicherweise die 3. Frage verneinen, dann müssen wir die 3. Frage eliminieren. Möglicherweise mißt die 3. Frage auf einer andern Dimension als die übrigen Fragen. Das Guttman-Skalierungsverfahren besteht nun darin, daß man die Fragen und die Personen solange „hin und her schiebt“ und Fragen eliminiert - bis die Zahl der „Fehler“ ein Minimum ist. Joachim Gerich hat, in Prog16, das Guttman-Verfahren für Almo programmiert.

Das Rasch-Verfahren kann nun als eine Weiterentwicklung des Guttman-Verfahrens betrachtet werden. (Das gilt auch für das Lazarsfeld'sche Ogiven-Modell, das wir im Almo-Dokument Nr. 20 "Latent Structure Analysis" und im Almo-Handbuch „Sozialwissenschaftliche Skalierungsverfahren“, Abschnitt P15 behandeln).

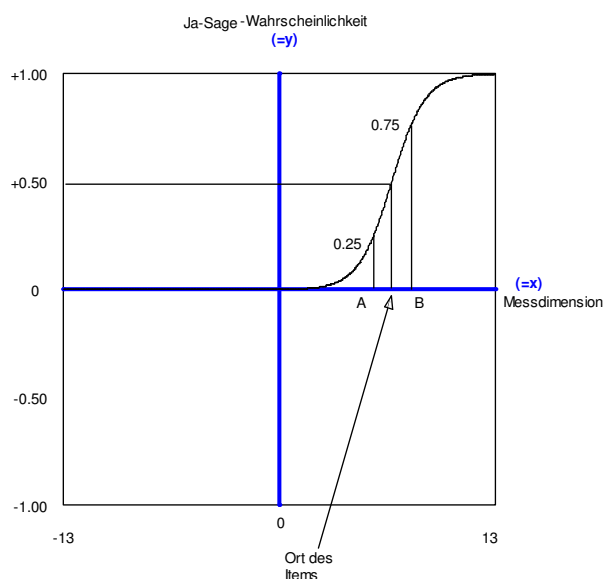
Bei Rasch wird als Kurve der "Ja-Sage-Wahrscheinlichkeit" ("trace line") anstatt einer Treppenfunktion die logistische Funktion angenommen.

$$p = \frac{1}{1 + e^{-b \cdot x}}$$

Beachte: Der Exponent lautet b minus x.

Wir werden später diese Funktion detaillierter darstellen.

Wir erhalten z.B. folgende Kurve für Frage 2 (an Demonstrationen teilnehmen)



Die logistische Kurve hat die Form eines "S". Der Ort der Frage 2 auf der Meßdimension ist da, wo die Kurve eine Ja-Sage-Wahrscheinlichkeit von 0.5 hat.

Die Person A in obiger Grafik hat hinsichtlich der Frage 2 eine Ja-Sage-Wahrscheinlichkeit von 0.25, die Person B von 0.75.

Der Unterschied zum Guttman-Verfahren ist offenkundig. Die Ja-Sage-Wahrscheinlichkeit kann dort nur 0 oder 1 sein.

Das Rasch-Skalierungsverfahren besteht nun darin, aus den Ja- und Nein-Antworten der Befragten auf die verschiedenen Fragen

- (1) den „Ort der Fragen auf der Meßdimension“, also den Skalenwert der Fragen und
- (2) den „Ort der Personen“, also den Skalenwert der Personen

unter der Annahme der logistischen Kurve der Ja-Sage-Wahrscheinlichkeit rückzurechnen.

P45.11.1 Eingabe mit Prog45mr

Für die zu skalierenden Variablen gelten folgende 2 Bedingungen:

1. Die Variablen müssen 0-1 kodiert sein
Ist dies nicht der Fall, dann müssen die Variablen entsprechend umkodiert werden
2. Die Variablen müssen in Richtung der Messdimension kodiert sein. 1 bedeutet also: "liegt höher auf der Messdimension" als 0
Ist dies nicht der Fall, dann müssen die Variablen in ihrer Kodierungsrichtung umgedreht werden

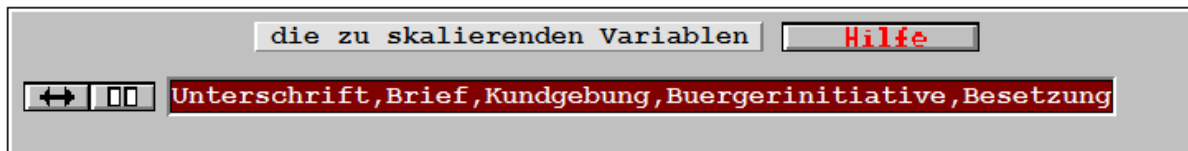
9		Option: Umkodierungen und Kein-Wert-Angaben
10		Option: Skalenwerte ermitteln und speichern
11		Grafik-Optionen
12		Programmende

P45.11.2 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: Schätzmethode

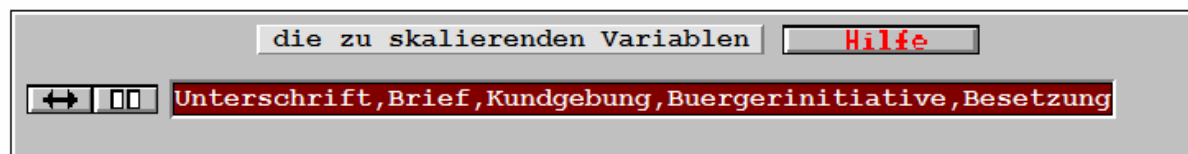


Almo bietet 2 Methoden der Schätzung der Skalenwerte der Variablen (=Schwierigkeitsparameter) und der Skalenwerte der Probanden (=Fähigkeitsparameter) an:

- 1 = unbedingte Maximun-Likelihood-Schätzung
- 2 = bedingte Maximun-Likelihood-Schätzung

Die beiden Methoden liefern unterschiedliche Ergebnisse. In der Regel sind sie nur wenig verschieden. Die bedingte Maximun-Likelihood-Schätzung gilt eher als überlegen.

Eingabe-Box 7: Die zu skalierenden Variablen



Geben Sie hier die Variable an, die für das Rasch-Verfahren verwendet werden sollen. Die Variablen müssen dichotom sein. Sind sie das nicht, dann müssen sie in der übernächsten Eingabe-Box entsprechend umkodiert werden.

Eingabe-Box 8: Option: Ein- und Ausschliessen von Untersuchungseinheiten.

Wenn Sie diese Optionsbox öffnen, dann wird es Ihnen ermöglicht, bestimmte Datensätze aus der Analyse auszuschließen bzw. in die Analyse einzuschließen. In P0.7 haben wir das ausführlich dargestellt.

Nun ist jedoch folgendes zu berücksichtigen. Wir wollen das an einem Beispiel zeigen: Sie wollen nur Männer für das Rasch-Verfahren verwenden. Die Skalenwerte und alle anderen Ergebnisse, die Almo liefert, beziehen sich also nur auf Männer.

Wenn Sie nun die Skalenwerte als zusätzliche Variable an die Datensätze anhängen wollen und in eine neue Datei speichern wollen, dann geschieht dies nur für die "männlichen" Datensätze. Die neue Datei besteht dann nur aus den Männern.

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

In P0.5 haben wir ausführlich dargestellt, wie umkodiert wird.

Für die zu skalierenden Variablen gelten folgende 2 Bedingungen:

1. Die Variablen müssen dichotom mit 0-1 kodiert sein.

Beispiel: Die Antwortvorgabe zu einer Frage V5 lautet:

	Codeziffer
stimme stark zu	1
stimme zu	2
unentschieden	3
lehne ab	4
lehne stark ab	5

Dann muß V5 umkodiert werden

V5 (1,2=1; 3,4,5=0)

Die Zustimmungseite haben wir mit 1 und die Ablehnungsseite mit 0 kodiert, wobei wir "unentschieden" auf die Ablehnungsseite genommen haben.

2. Die Variablen müssen 0-1 kodiert sein. Ist dies nicht der Fall, dann müssen die Variablen entsprechend umkodiert werden.

Beispiel: Variable V3 sei 1 - 2 kodiert. Es muß dann die Anweisung

V3 (1=0; 2=1)

geschrieben werden.

3. Die Variablen müssen in Richtung der Messdimension kodiert sein. 1 bedeutet also: "liegt höher auf der Messdimension" als 0. Ist dies nicht der Fall, dann müssen die Variablen in ihrer Kodierungsrichtung umgedreht werden. Beispiel: Wir wollen annehmen, dass Frage 1 folgendermaßen formuliert worden sei:

"Würden Sie sich weigern, an einer Unterschriftenaktionen teilzunehmen"

Nein	(0)
Ja	(1)

Die Antwort "Ja" mit Codeziffer 1 drückt also aus, dass man eher gegen Aktionen ist. Es muss als umkodiert werden:

V1 (1=0; 0=1) # d.h. aus 1 soll 0 werden und aus 0 soll 1 werden #

Ein weiteres Beispiel: Alle Fragen sind so kodiert: nein = 0 ja = 1

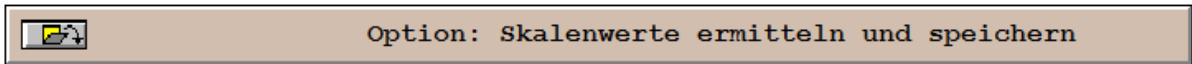
Aus irgend einem Grund wurde bei Frage V2 umgekehrt kodiert. Wir müssen also umdrehen:

V2 (0=1; 1=0)

Beachte:

Wenn Sie das Programm Prog45mr auch dazu verwenden wollen, die Skalenwerte der Personen an den Datensatz als zusätzliche Variable anzuhängen und in eine neue Datei zu speichern, dann ist zu berücksichtigen, dass umkodierte Variable in ihrer unkodierten Form in die neue Datei eingehen.

Eingabe-Box 10: Option: Skalenwerte ermitteln und speichern



Sie können für jede einzelne Person Skalenwerte bilden und als zusätzliche Variable an seinen Datensatz anhängen. Man wird dies aber erst dann tun, wenn man durch (eventuell mehrere) vorausgehende Analysen Klarheit über die Brauchbarkeit der Fragen gewonnen hat. In der Regel wird es notwendig sein, eine oder mehrere Fragen zu eliminieren. Wir werden darauf zurückkommen, wenn wir die Ausgabe des Programms erläutern.

Wenn Sie die Optionsbox öffnen, dann sehen Sie folgendes

X Loesche wieder diese Box (dann Voreinstellungen wieder gueltig)

Option: Skalenwerte ermitteln und speichern

"C:\Almo15\PROGS\Skalwert"

Geben Sie einen neuen Dateinamen ohne Erweiterung an
Almo erzeugt 3 Dateien:

1. eine nicht lesbare Almo-Arbeitsdatei mit der Erweiterung __.dir
2. eine anschaulbare Datei im freien Format mit der Erweiterung __.fre
3. eine Datei der Variablennamen mit der Erweiterung __.nam

In den unter 1. und 2. angegebenen neuen Dateien sind nun enthalten:

- die Variablen aus der alten Datei
- die Skalenwertvariable, die als letzte Variable hinter die Variablen der alten Datei gestellt wurde

In der unter 3. angegebenen neuen Datei der Variablennamen sind nun enthalten:

- die Variablennamen aus der alten Datei einschliesslich der in der Box "Freie Namensfelder" angegebenen Namen
- der Name "Skala" für die neue angehängte Skalenwertvariable

die Skalenwert-Variable transformieren
z.B. auf 3 Dezimalstellen runden
das ist nicht obligatorisch

Von den zu skalierenden Variablen dürfen x
Kein Wert besitzen. Für sie wird nachfolgende
Kein-Wert-Behandlung durchgeführt.
Sonst wird die Skalenwert-Variable auf
Kein-Wert gesetzt

Kein-Wert-Behandlung, wenn
zu skalierende Variable Kein Wert besitzen
Möglich 4 - 7; empfohlen: 6

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

Eingabefeld 1: Geben Sie einen Namen für die neue Datei an. In diese werden die seitherigen Variablen geschrieben, sowie die neu gebildete "Skalenwert-Variable". Geben Sie den Dateinamen ohne Erweiterung an. Almo erzeugt dann 2 Dateien:

1. eine nicht lesbare Almo-Arbeitsdatei mit der Erweiterung __.dir
2. eine anschaulbare Datei im freien Format mit der Erweiterung __.fre
3. eine Datei der Variablennamen

In den unter 1. und 2. angegebenen neuen Dateien sind nun enthalten:

- die Variablen aus der alten Datei
- die Skalenwertvariable, die als letzte Variable hinter die Variablen der alten Datei gestellt wurde

In der unter 3. angegebenen neuen Datei der Variablennamen sind nun enthalten:

- die Variablennamen aus der alten Datei einschliesslich der in der Eingabe-Box "Freie Namensfelder" angegebenen Namen
- der Name "Skala" für die neue angehängte Skalenwertvariable

Eingabefeld 2: Die Skalenwert-Variable besitzt Dezimalwerte. Sie kann mehrere signifikante Kommastellen haben. Es ist nun durchaus zulässig die Skalenwert-Variable zu transformieren. In unserem Beispiel haben wir auf die 3. Kommastelle gerundet. Das kann man machen, muss es aber nicht.

Eingabefeld 3 und 4: Es ist nahezu normal, dass einzelne Personen in einigen Variablen keinen Wert besitzen (etwa weil sie die Antwort verweigert haben). In diesem Fall muss folgendes geschehen:

a. Von den zu skalierenden Variablen dürfen x (in unserem Beispiel: 2) Kein_Wert besitzen. Sonst wird die Skalenwert-Variable auf Kein-Wert gesetzt

b. Für sie wird dann eine zu wählende Kein-Wert-Behandlung durchgeführt. Es wird ein Schätzwert für "Kein-Wert" eingesetzt, damit der Skalenwert berechenbar ist. Dazu gibt es im Prog45mr folgende Möglichkeiten:

Kein-Wert-Behandlung 4 und 5

Als Kein-Wert-Einsetzungswert wird die Codeziffer der häufigsten Ausprägung verwendet.

Beispiel:

Antwort	Code	Häufigkeit
Nein	0	200
Ja	1	300

Als Kein-Wert-Einsetzungswert wird 1 verwendet

Kein-Wert-Behandlung 6 und 7

Es wird der wahrscheinlichste Ausprägungswert eingesetzt.

Beispiel:

Antwort	Code	Häufigkeit	in %	in % kumuliert
Nein	0	200	40	40
Ja	1	300	60	100

Dann wird eine gleichverteilte Zufallszahl zwischen 0 und 100 erzeugt.

Liegt sie zwischen

0 und 40, dann wird für den fehlenden Wert 0 eingesetzt
40 100 1

Eingabefeld 5: Startwert für Zufallsgenerator

Für die Kein-Wert-Behandlung 6 und 7 muss ein Zufallswert generiert werden. Also besitzt einen eingebauten Zufallsgenerator. Die Folge der Zufallszahlen, die

Almo erzeugt, ist bei jeder Wiederholung von Prog45mr dieselbe - es sei den der Startwert für den Zufallsgenerator wird verändert. Verwenden Sie eine ungerade 6-stellige Zahl.

Lesen Sie nochmals unsere obigen Erläuterungen zu den Eingabe-Boxen 7 und 8. Werden Datensätze ein- oder ausgeschlossen oder werden Variable umkodiert, dann wirkt sich das auch auf die Faktorwerte und deren Abspeichern in einer neuen Datei aus.

P45.11.3 Ausgabe aus Programm P45mr

Wir werden nur die wichtigsten Teile der Ausgabe, die Almo liefert erläutern. Die übrigen Ausgabeteile sind eher für jene Benutzer gedacht, die sich mit den Details des Rasch-Verfahrens beschäftigen wollen.

```
Fuer Analyse aus Datenvektor ausgewaehlte Variable
 1 Unterschrift
 2 Brief
 3 Kundgebung
 4 Buergerinitiative
 5 Besetzung
```

```
Verwendete Schätzmethode: Unbedingte Maximum-Likelihood-Schätzung
Zahl der eingelesenen Datensätze = 61
Zahl der verarbeiteten Datensätze = 61
```

Variable	ja-absolut mit Extremgruppen	ja-relativ mit Extremgruppen
1 Unterschrift	37	0.607
2 Brief	33	0.541
3 Kundgebung	29	0.475
4 Buergerinitiativ	26	0.426
5 Besetzung	13	0.213

******* Erläuterung:**

Beispiel: Die Frage V2 "Brief" wird von 33 Personen (=54.1) % bejaht. Allgemein gilt, dass eine Frage umso "radikaler" ist, d.h. auf der Meßdimension weiter rechts liegt, je weniger Ja-Antworten sie erhält

Der Begriff "**Extremgruppen**" in der 2. Spalte, meint folgendes: Es gibt die Extremgruppe jener Personen, die alle Fragen verneinen, also 0 Gesamtpunkte haben. Und es gibt die Extremgruppe jener Personen, die alle Fragen bejahen also 5 Gesamtpunkte haben.

Gesamtpunkte	absolut	relativ
Summe der Einsen		
0	15	0.246
1	10	0.164
2	6	0.098
3	12	0.197
4	10	0.164
5	8	0.131

******* Erläuterung:**

Beispiel: 3 der 5 Fragen werden von 12 Personen (=19.7 %) bejaht.

Variable	ja-absolut ohne Extremgruppen*	ja-relativ ohne Extremgruppen*
1 Unterschrift	29	0.475
2 Brief	25	0.410
3 Kundgebung	21	0.344
4 Buergerinitiativ	18	0.295
5 Besetzung	5	0.082

*Probanden, die eine Gesamtpunktzahl von 0 oder 5 haben sind nicht beruecksichtigt

Relative Haeufigkeit von Ja-Antworten fuer jeden Gesamtpunktwert ohne Extremgruppen

Ges. pkte	Untersc V1	Brief V2	Kundgeb V3	Buerger V4	Besetzu V5
0	0.000	0.000	0.000	0.000	0.000
1	0.700	0.200	0.100	0.000	0.000
2	0.667	0.667	0.333	0.333	0.000
3	0.750	0.833	0.667	0.583	0.167
4	0.900	0.900	1.000	0.900	0.300
5	1.000	1.000	1.000	1.000	1.000

******* Erläuterung:**

Betrachten wir die letzte Spalte. Die Zahlen in dieser Spalte von oben nach unten sagen aus, dass die radikalste Frage V5 "Besetzen öffentlicher Gebäude".

von 0 % der Personen mit 1 Gesamtpunkt bejaht wurden
von 0 % der Personen mit 2 Gesamtpunkten bejaht wurden
von 16.7 % der Personen mit 3 Gesamtpunkten bejaht wurden
von 30 % der Personen mit 4 Gesamtpunkten bejaht wurden

Selbstverständlich gilt, daß bei einem Gesamtpunktwert von 0 die relative Häufigkeit durchgehend 0 ist und bei einem Gesamtpunktwert von 5 (wenn alle Fragen bejaht wurden) durchgehend 1.0 ist.

Startwerte fuer die Schwierigkeitsparameter
8 Iterationen, Konvergenzbedingung = 0

Startwerte fuer die Faehigkeitsparameter
27 Iterationen, Konvergenzbedingung = 0

6 Iterationen
Maximaler Fehler = 0.00007
Konvergenzbedingung = 0

Variable	Skalenwert S der Variablen (Schwierigkeitsparameter)	exp(S)
1 Unterschrift	-1.6143	0.1990
2 Brief	-0.9020	0.4058

3 Kundgebung	-0.2601	0.7709
4 Buergerinitiativ	0.2035	1.2256
5 Besetzung	2.5730	13.1048

***** **Erläuterung:**

Dies ist eines der besonders wichtigen Ergebnisse. Der Skalenwert der Variablen wird mitgeteilt. Dieser wird in der Literatur zum Rasch-Verfahren auch "Schwierigkeitsparameter" genannt. Wir erkennen, dass die jeweils nachfolgende Frage "radikaler" bezüglich der politischen Aktionsbereitschaft ist. Besonders groß ist der Sprung von "Bürgerinitiative bilden" zu "Besetzung öffentlicher Gebäude".

In der letzten Spalte wird e^S ausgegeben, weil in der Literatur gelegentlich die Schwierigkeitswerte in dieser Form ausgegeben werden und weil das Almo-Rasch-Programm in früheren Versionen auch nur e^S ausgegeben hat.

Gesamtpunkte	Skalenwert F der Probanden (Faehigkeitsparameter)	exp(F)
0	-3.0652	0.0466 (linear extrapoliert)
1	-1.8723	0.1538
2	-0.6794	0.5069
3	0.4253	1.5300
4	1.9124	6.7691
5	3.3995	29.9480 (linear extrapoliert)

***** **Erläuterung:**

Dies ist das wichtigste Ergebnis. Der Skalenwert der Probanden wird mitgeteilt. Dieser wird in der Literatur zum Rasch-Verfahren auch "Fähigkeitsparameter" genannt. Beispiel: Personen, die 3 der 5 Fragen bejaht haben, also einen Gesamtpunktwert von 3 haben, besitzen einen Skalenwert von 0.4253 in der Messdimension der "politischen Aktionsbereitschaft".

Zu beachten ist dabei, dass nicht alle diese Personen notwendigerweise die Fragen V1, V2 und V3 bejaht haben. Einige von ihnen können auch „Fehler“ gemacht haben und Fragen bejaht haben, die Ihnen eigentlich zu radikal sein müssten.

Zu beachten ist weiters dass die Skalenwerte für die "Extremgruppen" mit 0 Ja-Antworten und mit 5 (also allen) Ja-Antworten im Kalkül des Rasch-Verfahrens nicht berechenbar sind. Theoretisch sind sie minus unendlich bzw. plus unendlich. Bei der unbedingten Maximum-Likelihood-Schätzung werden sie linear extrapoliert. Bei der bedingten ML-Schätzung ergeben sie sich im Rahmen des Kalküls. Siehe dazu Handbuch zu P14.

In der letzten Spalte wird e^F ausgegeben, weil in der Literatur gelegentlich die Fähigkeitswerte in dieser Form ausgegeben werden und weil das Almo-Rasch-Programm in früheren Versionen auch nur e^F ausgegeben hat.

Empirische und theoretische Haeufigkeiten
je Variable und je Gesamtpunktwert
ohne Extremwerte 0 und 5

Variable	Gesamtpunkte	empirische Haeufigkeit	theoretische Haeufigkeit	Stand.abweichg. der theoret. Haeufigkeit

V1 Unterschrift				
	1	7	4.3586	1.5681
	2	4	4.3084	1.1021
	3	9	10.6187	1.1056
	4	9	9.7144	0.5268

V2 Brief				
	1	2	2.7483	1.4117
	2	4	3.3326	1.2172

	3	10	9.4847	1.4100
	4	9	9.4345	0.7304

V3 Kundgebung				
	1	1	1.6629	1.1775
	2	2	2.3802	1.1983
	3	8	7.9793	1.6351
	4	10	8.9775	0.9581

V4 Buergerinitiativ				
	1	0	1.1148	0.9952
	2	2	1.7556	1.1144
	3	7	6.6627	1.7215
	4	9	8.4669	1.1393

V5 Besetzung				
	1	0	0.1160	0.3386
	2	0	0.2235	0.4638
	3	2	1.2545	1.0599
	4	3	3.4060	1.4986

******* Erläuterung:**

Betrachten wir V1 "Unterschrift" und die 2. Zeile. Dort stehen die Personen, die insgesamt 2 Gesamtpunktwerte erzielt haben, also zu 2 von den 5 Fragen "Ja" gesagt haben. In der Spalte "empirische Häufigkeit" erfahren wir, dass 4 Personen, die 2 Gesamtpunktwerte haben, zu V1 "Unterschrift" auch "Ja" gesagt haben. In der 4. Spalte "theoretische Häufigkeit" steht die Zahl der Personen, die entsprechend dem Rasch-Kalkül "Ja" sagen müssten. Dies sind 4.3084 Personen.

Chi-Quadrat-verteilte Pruefgrößen

Variable	Testwert	Freiheitsgrade	Signifikanz (1-p)*100
1 Unterschrift	6.8985	3	92.61
2 Brief	1.0690	3	21.27
3 Kundgebung	1.5567	3	32.63
4 Buergerinitiativ	1.5600	3	32.71
5 Besetzung	0.9175	3	17.73
Gesamtskala	12.0018	12	55.39

******* Erläuterung:**

Dies ist ebenfalls eines der wichtigen Ergebnisse. Es teilt uns mit, ob die Fragen "brauchbar" sind.

Für jede Variable wird durch einen Chi-Quadrat-Test überprüft, ob sie die Modellannahmen des Rasch-Verfahrens erfüllt. Dabei ist "umgekehrt" zu denken. Ist die Signifikanz (1-p)*100 hoch, dann ist die Variable schlecht.

Der Chi-Quadrat-Test überprüft für jede Variable, ob die empirische und die theoretische Häufigkeit verschieden sind. Betrachten wir in der obigen Tabelle "Empirische und theoretische Häufigkeiten je Variable und je Gesamtpunktwert" Variable V1 Unterschrift. Für Personen mit

1 Gesamtpunktwert erhalten wir eine Differenz von	7 - 4.3586
2	4 - 4.3084
3	9 - 10.6187
4	9 - 9.7144

Der Chi-Quadrat-Beitrag ergibt sich aus

$$\text{Chi-Quadrat-Beitrag} = ((\text{empHäuf} - \text{theoHäuf}) / \text{Stdabwg}) ** 2$$

empHäuf = empirische Häufigkeit

theoHäuf= theoretische Häufigkeit

Stdabwg = Standardabweichung der theoretischen Häufigkeit

Für 1 Gesamtpunktwert ergibt sich so $(7 - 4.3586) / 1.5681 = 1.6845$. Dieser Wert ist noch zu quadrieren, so daß sich ergibt: 2.83758.

Insgesamt für alle 4 Gesamtpunktwerte erhalten wir einen Chi-Quadrat-Wert von =6.8985, der mit 3 Freiheitsgraden eine Signifikanz $(1-p)*100$ von 92.61 % besitzt. Wenn wir die konventionelle Schranke von 95% anwenden, dann heißt dies, daß sich empirische und theoretische Häufigkeit nicht signifikant unterscheiden. Für uns bedeutet dies, daß V1 Unterschrift eine brauchbare Variable ist.

Der Chi-Quadrat-Test macht eine Aussage darüber, wie gut bzw. schlecht die logistische Kurve an die empirischen Häufigkeiten je Gesamtpunktwert angepasst ist. Er ist ein „Kurvenanpassungstest“. Wir werden das später, zum Ende dieses Abschnitts, mit einem Kurvendiagramm ersichtlich machen. Variable mit einer Signifikanz $(1-p)*100$ über 95% sollten eliminiert werden. Danach sollte eine neue Analyse gerechnet werden. Das wird solange wiederholt, bis keine Variable mehr über 95 % liegt.

Wir erkennen, dass die Frage V1 "Unterschrift" mit einer Signifikanz $(1-p)*100$ von 92.61 knapp an die 95% Schwelle herankommt, also keine gute Frage aber immerhin noch akzeptabel ist.

Der Testwert und die Signifikanz der Gesamtskala (12.0018 und 55.39) sollten möglichst klein sein. Die Signifikanz sollte nicht über 95% liegen.

Eine Warnung muß jedoch ausgesprochen werden: Der Chi-Quadrat-Test ist kein guter Test. Bei großen Stichproben sind schon sehr kleine Differenzen zwischen empirischen und theoretischen Häufigkeiten signifikant, so daß sehr häufig Variable zurückgewiesen werden, die eigentlich noch akzeptabel erscheinen.

Im oben erwähnten "allgemeine ordinalen Rasch-Modell" (Almo-Programm-Maske Prog14m3 oder Prog14m4) wird noch ein weiterer Test angeboten, der in der Literatur besser beurteilt wird. Es ist der Outfit- und Infit-Test. Er liefert für die 5 Items folgendes Ergebnis:

Itemfit: Outfit und Infit

Variable	Chi-Quad.	outfit	t *)	infit	t *)
V1 Unterschrift	51.28997	1.30522	0.73477	1.20652	1.09879
V2 Brief	34.25486	1.02595	0.19773	0.98384	-0.02766
V3 Kundgebung	26.56771	0.89993	-0.15844	0.91480	-0.43661
V4 Buergerinitiativ	27.05111	0.90786	-0.12526	0.94460	-0.26696
V5 Besetzung	24.19530	0.86104	0.06142	0.95023	-0.15439

*) ein t-Wert zwischen -1.96 und +1.96 bedeutet: Die Variable ist "Rasch-konform" ein nicht-Rasch-konformer Item wird durch ein Fragezeichen ? markiert

Almo liefert nun für alle 5 Fragen Detailanalysen und je 2 Grafiken. Wir wollen hier nur Frage V4 "Buergerinitiativ" betrachten. Betrachten wir die 3. Zeile:

Detailanalyse fuer Variable Buergerinitiativ

"trace line"
Ja-Sage-Wahrscheinlichkeit
bei jeweiligem Gesamtpunktwert

Gesamtpunktwert	Faehigkeitswert	empirisch	theoretisch
-----------------	-----------------	-----------	-------------

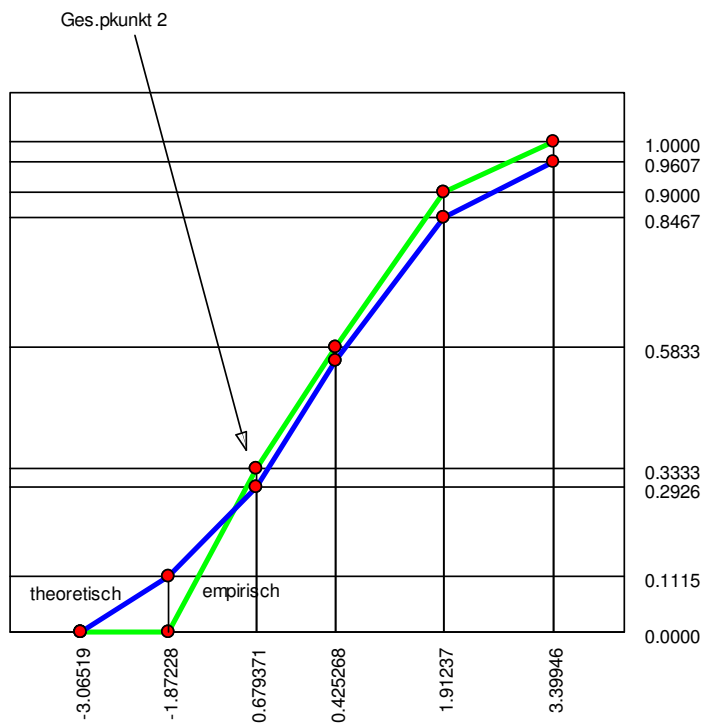
0	-3.0652 *)	0.0000	0.0367
1	-1.8723	0.0000	0.1115
2	-0.6794	0.3333	0.2926
3	0.4253	0.5833	0.5552
4	1.9124	0.9000	0.8467
5	3.3995 *)	1.0000	0.9607

*) ist extrapoliert

Sie enthält Personen mit einem Gesamtpunktwert von 2. Diese haben also zu 2 von 5 Fragen "Ja" gesagt. Sie haben einen Skalenwert (=Fähigkeitsparameter) von -0.6794. Siehe auch oben in der Tabelle "Skalenwert F der Probanden".

Die Frage V4 "Buergerinitiativ" wurde von 33.33 % dieser Personen bejaht. Eigentlich müßten es 29.26 % sein - wenn die Variable V4 exakt die Voraussetzung der logistischen "trace line" erfüllen würde. Diese beiden Werte stehen in den mit "empirisch" und "theoretisch" überschriebenen Spalten.

Empirische und theoretische Trace-Line fuer Variable Buergerinitiativ



Die empirische Kurve ist grün, die theoretische ist blau

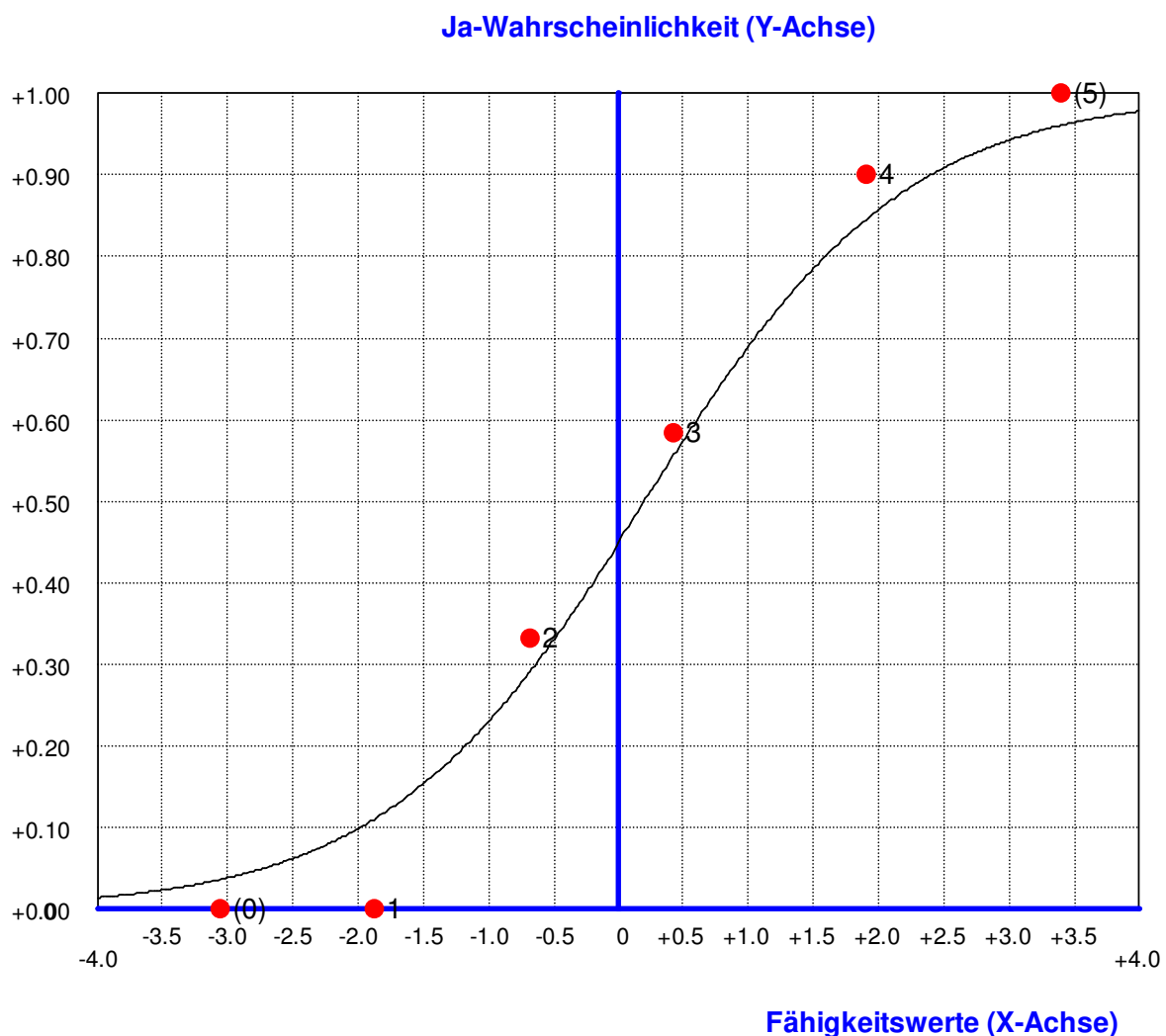
Die grüne Kurve (im gedruckten Handbuch: hell) ist die empirische "trace line". Der durch den Pfeil markierte Punkt repräsentiert die Personen mit 2 Gesamtpunktwerten. Dieser Punkt hat die Koordinaten $x = -0.6794$ (Fähigkeitswert) und $y = 0.3333$ (empirischer Ja-Sage Prozentsatz)

Die blaue Kurve (im gedruckten Handbuch: dunkel) ist die theoretische "trace line". Der durch den Pfeil markierte Punkt repräsentiert die Personen mit 2 Gesamtpunktwerten. Dieser Punkt hat die Koordinaten $x = -0.6794$ (Fähigkeitswert) und $y = 0.2926$ (theoretische Ja-Sage Wahrscheinlichkeit)

Die beiden Kurven liegen sehr dicht beieinander. Dies ist ein Hinweis darauf, dass die Frage V4 "Buergerinitiativ" eine gute Frage ist - gut in dem Sinne, dass sie die Bedingung der logistischen trace line erfüllt.

Die theoretische trace line in obiger Zeichnung entstand sehr einfach dadurch, daß die Punkte der theoretischen "Ja-Wahrscheinlichkeiten" linear miteinander verbunden wurden. Nun wird beim Rasch-Verfahren unterstellt, daß die trace line eine logistische Kurve ist. Also zeichnet deswegen noch folgende Grafik:

Logistische Funktion $Y = 1/(1+e^{-(0.203-X)})$ für Variable V4 Buergerinitiativ
 rote Punkte=empirische Ja-Wahrscheinlichkeit bei einem Gesamtpunktwert von x
 Die Ges.punktswerte stehen bei den Punkten. Ges.punktswerte in Klammern geschätzt



Punkt 2, 3 und 4 (also Gesamtpunktwert 2, 3 und 4) liegen fast genau auf der logistischen Kurve. Punkt 1 liegt etwas unterhalb der Kurve. Punkt 0 und 5 (in Klammern gesetzt) wurden bezüglich ihrer Fähigkeitswerte bei der unbedingten ML-Schätzung linear interpoliert. Bei der bedingten ML-Schätzung ergeben sich ihre Fähigkeitswerte im Rahmen des Kalküls. Theoretisch sind sie minus unendlich bzw. plus unendlich.

Bei der Erläuterung zu den „Chi-Quadrat-verteilten Prüfungsgrößen“ je Variable haben wir ausgeführt, daß der Chi-Quadrat-Test als ein Kurvenanpassungstest betrachtet werden kann. Durch den Kalkül des Rasch-Verfahrens wird die horizontale Lage der Logitkurve bestimmt (durch den Schwierigkeitsparameter), ebenso die x-Werte der empirischen Punkte (die Fähigkeitswerte). Die y-Werte der empirischen Punkte sind die empirischen relativen Häufigkeiten. Die Steilheit der logistischen Kurve ist immer dieselbe. Der entsprechende Parameter der Logitfunktion wird beim Rasch-Verfahren auf 1 gesetzt. Der Chi-Quadrat-Test sagt dann aus, wie gut oder schlecht die Logitkurve durch die empirischen Punkte verläuft. Dabei werden die beiden Punkte 0 (keine Frage bejaht) und 5 (alle Fragen bejaht) nicht berücksichtigt.

Bei dieser Grafik ist auch der Schwierigkeitswert von V4 gut zu erkennen. Dieser Wert liegt an der Stelle der x-Achse, wo die logistische trace line den Wert 0.5 auf der y-Achse durchläuft. Der Wert ist (genau) 0.2035. Dies ist der Skalenwert (= Schwierigkeitsparameter) der Variablen. Siehe auch oben in der Tabelle „Skalenwert S der Variablen“.

P45.11.4 Einschreiben der Skalenwerte in den Datensatz

Der Zweck des Rasch-Skalierungsverfahrens ist es, einen Skalenwert je Versuchsperson aus einer Batterie von Items zu erhalten. Im Programm-Output werden diese Skalenwerte als "Fähigkeitsparameter" bezeichnet. Betrachten wir noch einmal unser Beispiel mit 5 Variablen:

Ist der Gesamtpunktwert gleich 5, d.h. "bejaht" die Versuchsperson alle 5 Variable, dann ist der Skalenwert (=Fähigkeitsparameter), der im Output aufscheint, 3.3995. Ist der Gesamtpunktwert gleich 4, d.h. bejaht die Versuchsperson 4 Variable - gleichgültig welche - dann ist der Skalenwert (=Fähigkeitsparameter) 1.9124 etc.

Wir müssen nun diese Skalenwerte (=Fähigkeitsparameter) den Versuchspersonen als neue Variable anfügen.

Almo gibt zum Schluss der Ausgabe folgende Mitteilung aus:

```
***** MITTEILUNG
Lesen oder Schreiben korrekt beendet in Datei
"C:\Almo6\PROGS\Skalwert.fre"

***** MITTEILUNG
Lesen oder Schreiben korrekt beendet in Datei
"C:\Almo6\PROGS\Skalwert.dir"
```

Durch Doppelklick auf den Dateinamen "C:\Almo6\PROGS\Skalwert.fre" kann man die neue Datei laden und anschauen. Wir zeigen hier die ersten 5 Datensätze.

Person	V1	V2	V3	V4	V5	V6
1	1	1	1	1	1	340
2	1	1	1	1	0	191
3	1	1	1	0	1	191
4	1	1	1	1	0	191
5	1	1	1	1	1	340
.
.
.

V1 bis V5 wurden aus der alten Datei übernommen. V6 ist die neue Skalenwert-Variable. In der Eingabe-Box 9 haben wir sie mit 100 multipliziert und auf eine Ganzzahl gerundet. Betrachten wir Person 2. Sie hat 4 Fragen bejaht, V5 hat sie verneint. Aus obiger Tabelle der "Skalenwert F der Probanden" entnehmen wir für 4

Gesamtpunkte eine Skalenwert (Fähigkeitswert) von 1.9124. Wird dieser Wert mit 100 multipliziert und auf eine Ganzzahl gerundet, dann entsteht für die Skalenwert-Variable $V_6=191$.

P45.11.5 Weitere in Almo vorhandene Skalierungsverfahren

Neben dem dichotomen Rasch-Skalierungsverfahren sind in Almo noch folgende Skalierungsverfahren vorhanden:

- das allgemeine ordinale Rasch-Modell für dichotome und polytome Items programmiert von Heinrich Potuschak in Prog14m3 und Prog14m4
- dichotome und polytome Guttman-Skalierung in Prog 16, programmiert von Joachim Gerich
- dichotome und polytome Mokken-Skalierung in Prog 16, programmiert von Joachim Gerich
- Lazarsfelds "latent structure analysis" in Prog 15, programmiert von Hermann Denz

Siehe zu diesen Verfahren auch das Almo-Handbuch „Sozialwissenschaftliche Skalierungsverfahren“.

P45.11.6 Faktorenanalyse oder Rasch-Skalierung

Das Rasch-Verfahren sollte verwendet werden, wenn die Variablen dichotom sind. Die Faktorenanalyse kann bei dichotomen Variablen Artefakte liefern. Siehe dazu etwa Denz (1982, Abschnitt 1.1 und 1.2).

Sind die Variablen polytom und können sie als quantitativ eingeschätzt werden, dann sind Faktorenanalyse und Rasch-Verfahren Konkurrenten. Das Rasch-Verfahren setzt allerdings voraus, dass die Items eindimensional sind und dieselbe Trennschärfe besitzen. Wie oben ausgeführt, kann das annähernd durch „outfit“ bzw. „infit“ überprüft werden. Siehe hierzu die ausführliche Darstellung des Rasch-Modells im Almo-Dokumen Nr. 7 „Allgemeines ordinale Rasch-Modell“.

Schlagwortverzeichnis

Allgemeines Lineares Modell	67	Erwartungswert	58, 62
ALM	67	Excel	19, 23
Almo-Arbeitsdatei	17	Fähigkeitsparameter	186, 191
Balkendiagramm	162	Faktorenanalyse	156
Cramers V	107	Faktorenzahl	162
Data Mining	8	Faktorladung	156, 163, 172
Datei der Variablennamen	25, 26	Faktorwert	156, 158, 167
Dateien vereinen	109, 140	Faktorwert-Koeffizient	164, 172
Datenfusion	109	Faktorwertvariable	159, 172
Daten-Imputation	57	fehlende Werte	57
Dezimalzeichen	32	Fehlerstreuung	80
dichotome Variable	73	Feld	34
direktes Format	25	Format	31
diverse Werte	41, 45	freies Format	25, 26
Effekte	69, 113	gemeinsame Variable	110, 119, 144
Eigenwert	162	Gesamtpunktwert	149
Eindimensionalität	169	gespendete Variable	110, 120
Empfängerdatei	118, 125	gewichteter Gesamtpunktwert	156

gewichtetes ALM 113
 Grafik 51
 Gruppierungsvariable 50
 Guttman-Skalierung 173, 193
 Häufigkeitsverteilung 51
 Hintergrundfaktor 156
 Imputation 57
 Ja-Sage-Wahrscheinlichkeit 175
 joint scale 174
 Kein-Wert 41, 57
 Kein-Wert-Angabe 122
 Kein-Wert-Behandlung 77
 Kein-Wert-Code 75
 Kommunalitäten 162
 Korrelationsmatrix 160
 Kurvenanpassungstest 188, 191
 Ladungsmatrix 166
 latent structure analysis 193
 logistische Funktion 175
 logistische Kurve 190
 Logitanalyse 68, 86, 113, 131
 Median 55, 58
 Messung 149
 Mittelwert 41, 55
 Mittelwert für fehlende Werte 58
 Mokken-Skalierung 193
 monotone Frage 173
 multiple Faktorbestimmtheit 164
 multiple Imputation 95
 multiple Korrelation 82, 114, 129
 nominal-polytome Variable 86
 Obergrenze 45
 ordinale Variable 42
 paarweises Ausscheiden 77, 124
 Prognosewert 83, 113, 124, 129
 Prognosewert-Behandlung 78
 Prognosewerte für fehlende Werte 67
 Punktwolke 163
 Quadratsummenmatrix 77
 quantitativ 153
 Quartil 55
 Quartilsabstand 55, 63
 Rasch-Skalierungsverfahren 173
 Regressionskoeffizient 69, 113
 Residuen 79, 138
 Schiefwinkliger Rotation 164
 Schwierigkeitsparameter 186
 Skalenwert 175, 181, 186
 Skalierungsverfahren 173
 Spenderdatei 118
 spezifische Variable 110
 Standardabweichung 41, 55
 Standardauswertung 8
 standardisieren 149
 Strukturmatrix 165
 Survey 104
 Tabs getrennt 19
 trace line 173, 189, 190
 Trefferhäufigkeit 92, 138
 Trennzeichen 33
 Untergrenze 45
 ursächliche Variable 74, 121
 vollständiges Ausscheiden 78, 124
 zählen 47
 Zahlenvariable 24
 Zeichenvariable 24, 32
 Zeilenumbruch 33
 Zufallsgenerator 64, 124
 Zufallsüberlagerung 62, 64, 138
 Zufallswert 62
 zusammengefasste Datei 146

Literatur

Siehe Teil II