



## **Datenfusion**

P45.8

Kurt Holm

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

**2013**

## INHALTSVERZEICHNIS

<b>P45.8 DATEIEN VEREINEN .....</b>	<b>1</b>
P45.8 Datenfusion .....	3
P45.8.1 Schritt 6a: Datenfusion mit dem Allgemeinen Linearen Modell.....	7
P45.8.1.1 Eingabe in Programm Prog45mw zur einseitigen Datenfusion .....	8
P45.8.1.2 Erläuterungen zu den Boxen.....	12
P45.8.1.3 Ausgabe aus Prog45mw.....	20
P45.8.2 Schritt 6b: Datenfusion mit der Logitanalyse .....	24
P45.8.2.1 Eingabe in Prog45my .....	25
P45.8.2.2 Erläuterungen zu den Boxen.....	28
P45.8.2.3 Ausgabe aus Prog45my.....	29
P45.8.3 Schritt 6c: Fusionierte Dateien vereinen.....	33
P45.8.3.1 Eingabe in Prog45mx .....	34
P45.8.3.2 Erläuterung zu den Boxen.....	37
P45.8.3.3 Ausgabe .....	41
P45.8.3.4 Weiterführende Hinweise.....	41

## P45.8 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 (Datenfusion)
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:Pflichtschule ohne Lehre, Pflichtschule mit Lehre, mittlere Schule,  
Gymnasium,Hochschule;  
Name 2=Beruf:Bauern,  
Selbständig,  
Arbeiter,  
Facharbeiter,  
Angestellte/Beamte,  
leitende Angestellte/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:Pflichtschule ohne Lehre, Pflichtschule mit Lehre, mittlere Schule,  
Gymnasium,Hochschule;  
Name 4=Beruf:Bauern,  
Selbständig,  
Arbeiter,  
Facharbeiter,  
Angestellte/Beamte,  
leitende Angestellte/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, deswegen auch nicht „gespendet“ werden sollen. 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. Wir tun so als ob sie nicht vorhanden wären. 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 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 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  
QuantKor.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 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 Handbuch „P45 Data Mining“, 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  
 $\beta_1$  = Regressionskoeffizient für B  
A = Alter  
 $\beta_2$  = 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 Handbuch „P45 Data Mining“, 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 Handbuch „P45 Data Mining“, 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 Handbuch „P45 Data Mining“, 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 Handbuch „P45 Data Mining“, 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.

#### ***P45.8.1.0 Erstellen einer Direktzugriffsdatei***

Die nachfolgenden Programme erfordern es, dass die Daten der zu fusionierenden Dateien in die Form einer Direktzugriffsdatei überführt werden. Das geschieht mit einer der drei Programm-Masken

Prog00mr (empfehlenswert)

Sie finden diese in Almo durch Klick auf den Knopf "Verfahren" in der Knopfleiste unter der Menüleiste; dann „Datei-Operationen"

oder eines der folgenden Data-Mining-Programme

Prog45md

Prog45mh

Diese beiden finden Sie durch Klick auf den Knopf „Data Mining“ in der Knopfleiste. Diese beiden werden im Almo-Handbuch P45 „Data Mining“, Abschnitt P45.1 und P45.2 ausführlich erläutert.

#### ***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 -->

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

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

4     
 **erzeuge zusätzliche Namensfelder**

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

6

**Empfängerdatei**

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

7

**gemeinsame Variable aus Spender- und Empfängerdatei**

U1,2,4,5,6,7, 8 Variablennummern aus Spenderdatei

U3,4,5,6,8,9,10 Variablennummern aus Empfängerdatei

8

**Zu spendende Variable aus der Spenderdatei**

**BEACHTEN:** Erlaubt sind:

1. Beliebig 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

---

2 Zahl der zu spendenden Variablen

9

**Ursächliche Variable**  
für die zu spendende Variable in der Spenderdatei

**ursächliche nominale Variable**

Beruf, Geschlecht

0

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

---

**ursächliche quantitative Variable**

Bildung,Alter,Kirchgang

---

**ursächliche ordinale Variable**


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 Boxen

#### Box 1 und Box 2:

Siehe Anhang P0.1 und P0.2.

#### Box 3: Datei der Variablenamen aus Spenderdatei

Datei der Variablenamen aus Spenderdatei

     zeige = Namensdatei in Output zeigen  
leer = nicht

Geben Sie hier die Datei an, in die Sie die Variablenamen 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 Variablenamen der Empfängerdatei werden nicht angegeben.

Wie Variablenamen geschrieben werden ist ausführlich in Anhang P0.3 dargestellt.

#### Box 4: Freie Namensfelder für Spenderdatei

Zur Funktion dieser Box siehe Anhang P0.3.

Beachte: Es werden keine Variablenamen für die Empfängerdatei angegeben.

#### Box 5: Spenderdatei

#### Box 6: Empfängerdatei

Spenderdatei

     \die Datei umfasst so viele Variable

Empfängerdatei

     \die Datei umfasst so viele Variable

*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.

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

gemeinsame Variable aus Spender- und Empfängerdatei	
↔ [V1,2,4,5,6,7, 8]	\ Variablennummern aus Spenderdatei
↔ [V3,4,5,6,8,9,10]	\ Variablennummern aus Empfängerdatei
↔ [7]	\ Zahl der gemeinsamen Variablen

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 Box zu deklarieren, die in der übernächsten Box 9 als ursächliche Variable eingetragen werden. Anders formuliert: Variable, die in Box 9 nicht als ursächliche Variable verwendet werden, müssen nicht in dieser Box 7 als gemeinsame deklariert werden.

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

**Box 8:** Zu spendende Variable aus der Spenderdatei

Zu spendende Variable aus der Spenderdatei

BEACHTEN: Erlaubt sind:

1. Beliebig viele quantitative und/oder dichotome Variable

oder (exklusiv)

2. Eine nominale Variable mit beliebig vielen Ausprägungen

\

(mehrere) quantitative Variable

\

(nur eine) nominale Variable

In dieser 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 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 Box 12 "5" eingesetzt, dann entsteht für die dichotome "zu spendende" Variable derselbe Wert, egal ob sie im 1. oder 2. Eingabefeld der Box 8 eingetragen wurde. Wird "7" eingesetzt, dann entstehen etwas verschiedene

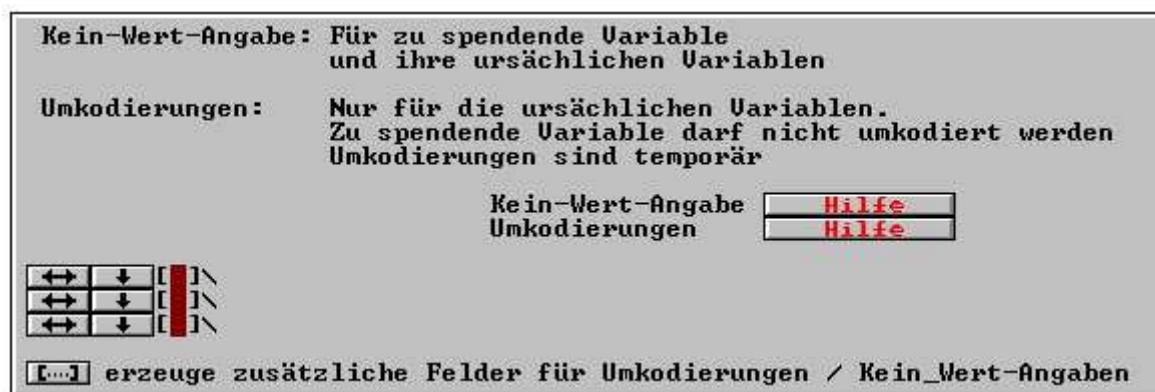


*Eingabefeld 2:* Wenn Sie Interaktionen zwischen den urächlichen nominalen Variablen miteinbeziehen wollen, dann geben Sie hier die Interaktionsordnung an. Siehe dazu die ausführliche Erläuterung für Prog45mf in Handbuch „P45 Data Mining“, Abschnitt P45.15.1.2, Erläuterung zu Box 6 "Ursächliche Variable".

*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.

### **Box 10:** Kein-Wert-Angabe und Umkodierungen



#### **Kein-Wert-Angaben:**

Almo muß selbstverständlich wissen, an welchen Codeziffern es den Kein-Wert-Fall in den in 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 Box "Kein-Wert-Angabe und Umkodierungen" können auch Variable umkodiert werden.

- Es dürfen nur die ursächlichen Variablen umkodiert werden - nicht die "zu spendende Variable"
- 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 Box 14).

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

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

Kein-Wert-Behandlung der ursächlichen Variablen aus Spenderdatei	
<input type="button" value="↓ [3] \"/>	bei Berechnung der Regressionskoeffizienten und Effekte möglich: 1 - 7; empfohlen: 3 =Vollständiges Ausscheiden 1 =Paarweises Ausscheiden
<input type="button" value="↓ [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 Handbuch „P45 Data Mining“, Abschnitt P45.7.3 über die Einsetzung fehlender Wert mit Hilfe des ALM, bei der Erläuterung der

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 Box klickt, dann werden ihm diese Methoden gezeigt.

**Box 12:** Transformation der Prognosewerte für zu spendende Variable

**Box 13:** Startwert für Zufallsgenerator

Diese 2 Boxen entsprechen den bereits in Abschnitt Handbuch „P45 Data Mining“, Abschnitt P45.7.1.3 dargestellten Boxen 10 und 11, so daß wir hier nur eine kurze Ergänzung vornehmen wollen.

Zu 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 Box 8: "Zu spendende Variable" hingewiesen.

**Box 14:** Neue Empfängerdatei

*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 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 anschauliche Datei im freien Format  
mit der Erweiterung `__.fre`  
Im Beispiel: `"C:\Almo6\PROGS\NeuDatenB.fre"`

In der "anschaulichen" Datei können Sie sich nochmals die von Almo errechneten Einsetzungswerte anschauen.

*Eingabefeld 2:* Geben Sie die Nummern für die gespendeten Variablen an, die diese in der Empfängerdatei einnehmen sollen. Sie müssen hinter der letzten Variablen aus der ursprünglichen Empfängerdatei eingesetzt werden. Die Variablennummern müssen in folgender Form geschrieben werden:

V11,12

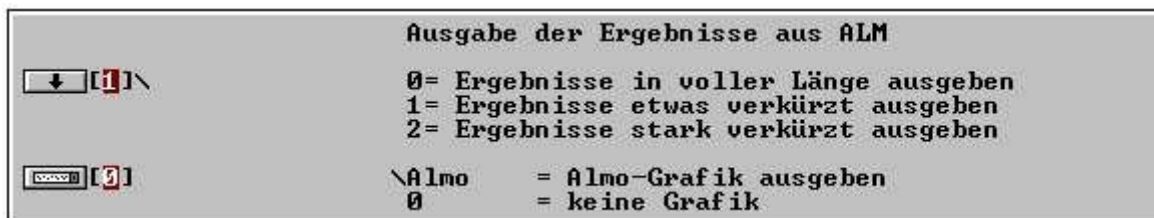
Vorne steht 'V' darauf folgen die Nummern - getrennt durch Beistrich.

*Eingabefeld 3:* Geben Sie die Nummern der Variablen an, die in die neue Empfängerdatei übernommen werden sollen. In der Regel wird man alle Variable aus der Ursprungsdatei plus der gespendeten Variablen übernehmen. Verwenden Sie dabei die Schreibweise

V1:12

Vorne steht 'V' darauf folgt die Nummer 1, dann Doppelpunkt und dann letzte Nummer.

### **Box 15:** Ausgabe der Ergebnisse aus ALM



*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.

*Eingabefeld 2:* Man sollte auf "0" (keine Grafik) einstellen, da man sonst einen sehr umfangreichen Output erhält.

### **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 Box 15 "Ausgabe der Ergebnisse aus ALM" wurde auf 1 (=Ergebnisse etwas verkürzt) eingestellt. Die Ausgabe ist anders wenn in Box 15 auf 0 oder 2 eingestellt wird. Siehe dazu unsere ausführliche Darstellungsstellung in Handbuch „P45 Data Mining“, 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 Lebenszufriedenheit

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 unabhengigen 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	Signifikanz	
	Korrelat.	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	Signifikanz	
	Korrelat.	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	<b>0.535313</b>
F-Wert f. erklarte Streuung	5.980270
Freiheitsgrade Nenner = 9	
Zaehler= 134	
Signifikanz: p	0.000014
Signifikanz: (1-p)*100	<b>99.998561 %</b>
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	<b>0.216601</b>
F-Wert f. erklarte Streuung	0.732913
Freiheitsgrade Nenner = 9	
Zaehler= 134	
Signifikanz: p	0.679626
Signifikanz: (1-p)*100	<b>32.037380 %</b>
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 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	<b>0.5352</b>	6.0225	9	0.0000	<b>99.9986</b>
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 Kirchgang	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 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 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 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 Handbuch „P45 Data Mining“, 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 Handbuch „P45 Data Mining“, 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 ? --> Hilfe  
Bedienung --> Hilfe

- 1 Speicher fuer x Variable Hilfe  
Vereinbare Variable= 22 ;
- 2 ↓ Option: Weitere Vereinbarungen - nur wenn Almo dazu auffordert
- 3 Datei der Variablennamen aus Spenderdatei Hilfe  
↔ 📁 "C:\Almo7\Testdat\DatenB.nam"  
↔ ↓ zeige      zeige = Namensdatei in Output zeigen  
leer = nicht
- 4 Freie Namensfelder für Spenderdatei Hilfe  
↔ ↔ █  
⋮ erzeuge zusätzliche Namensfelder
- 5 Spenderdatei Hilfe  
📁 "C:\Almo7\Testdat\DatenB.dir"
- 6 Empfängerdatei Hilfe  
📁 "C:\Almo7\Testdat\DatenA.dir"
- 7 gemeinsame Variable aus Spender- und Empfängerdatei Hilfe  
↔ 03,4,5,6,8,9,10      Variablennummern aus Spenderdatei  
↔ 01,2,4,5,6,7, 8      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 Boxen

Prog45my entspricht weitgehend dem in Abschnitt P45.8.1.2 bereits erläuterten Prog45mw, so daß wir hier nur 3 Boxen erläutern müssen. Der Benutzer sollte beachten, daß in der 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".

**Box 8:** Zu spendende Variable aus der Spenderdatei

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

**Zu spendende Variable aus der Spenderdatei**

\ (nur eine) nominale Variable

---

**Ursächliche Variable für die zu spendende Variable in der Spenderdatei**

\  
ursächliche nominale Variable

\  
ursächliche quantitative Variable

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 Box 9 in Abschnitt P45.8.1.2 vorgetragen.

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

**Kein-Wert-Behandlung der ursächlichen Variablen aus Spenderdatei**

Hilfe

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

[2] \ 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 Handbuch „P45 Data Mining“, 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

Datensatz 15	
Datensatz 29	
Datensatz 30	
Datensatz 36	<b>Spenderdatei</b>
Datensatz 53	
.	
.	
.	
.	

\*\*\*\*\* **Erläuterung:**

Almo meldet, daß Datensätze in der Spenderdatei ausgeschlossen werden mussten, weil in den Analysevariablen Kein-Wert aufgetreten ist. Wir hatten oben schon darauf hingewiesen, daß in der Logitanalyse ein Datensatz ausgeschieden wird, wenn auch nur in einer Analyse-Variablen der Wert fehlt.

Zahl der eingelesenen Datensätze = 192  
 Zahl der in Analyse einbezogenen Datensätze = 168

Maximum-Likelihood-Schaetzer der Koeffizienten:

Ergebnisse fuer 2. Auspraegung "ÖVP" der abhaengigen Variablen V1 Parteipraefe (als Referenz wird die 1. Auspraegung "SPÖ" verwendet)

unabhaengige Variable	Regress. koeff.β	Risiko epx(β)	relatives Risiko	Signifikanz (1-p)*100	partielle Korrelation
A1 Beruf: Bauern	8.16594	3519.01866	351801.86570	15.96	0.06395
A2 Beruf:Selbstän	-1.71441	0.18007	-81.99301	16.56	-0.06390
A3 Beruf:Arbeiter	-2.09236	0.12340	-87.66048	20.28	-0.06353
A4 Beruf:Facharbe	-2.03013	0.13132	-86.86809	19.69	-0.06359
A5 Beruf:Angest/B	-1.82205	0.16169	-83.83059	17.68	-0.06379
A6 Beruf:leitende	-0.50699	0.60231	-39.76943	5.24	-0.06456
B1 Geschlec: m	-0.04157	0.95928	-4.07166	11.67	-0.06427
B2 Geschlec: w	0.04157	1.04244	4.24448	11.67	0.06427
C1 Gewerksc: ja	-0.41521	0.66020	-33.98006	85.88	-0.01867
C2 Gewerksc: nein	0.41521	1.51469	51.46939	85.88	0.01867
D1 Gastarbe:zu wenig	0.43669	1.54758	54.75821	49.85	0.05685
D2 Gastarbe:ausreich	-0.08969	0.91422	-8.57822	17.24	-0.06384
D3 Gastarbe: zuviel	-0.34701	0.70680	-29.31998	56.02	-0.05411
V5 Alter	-0.01685	0.98330	-1.67043	53.93	-0.05512
V9 Kirchgang	-0.73834	0.47791	-52.20916	100.00	-0.20830

\*\*\*\*\* **Erläuterung:**

Wir zeigen hier nur die Koeffizienten für die ÖVP. Entsprechende Koeffizienten-Tabellen werden auch für die anderen Parteien ausgegeben. Almo teilt die Regressionskoeffizienten (und weitere Koeffizienten) der ursächlichen Variablen mit. Almo verwendet diese Regressionskoeffizienten für 2 Zwecke:

1. Um Prognosewerte für die Personen der Spenderdatei zu rechnen.
2. Um Prognosewerte für die Personen der Empfängerdatei zu rechnen.

Trefferhaeufigkeiten bei Individualdaten fuer abhaengige Variable V1 Parteipraefe

		tatsaechlich					prognostiziert absolut				
		1 SPÖ	2 ÖVP	3 FPÖ	4 Grüne	5 keine	1 SPÖ	2 ÖVP	3 FPÖ	4 Grüne	5 keine
SPÖ	1	53	0	0	0	0	37	8	0	0	8
ÖVP	2	0	49	0	0	0	9	33	0	0	7
FPÖ	3	0	0	20	0	0	7	2	2	0	9
Grüne	4	0	0	0	5	0	1	1	0	3	0
keine	5	0	0	0	0	41	14	5	1	0	21

		prognostiziert relativ					erwartet Zufall				
		1 SPÖ	2 ÖVP	3 FPÖ	4 Grüne	5 keine	1 SPÖ	2 ÖVP	3 FPÖ	4 Grüne	5 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 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 Zufallvariation des Prognosewertes mit der oben angegebenen Standardabweichung hervorgeht.

Ursächliche Variable, die Kein\_Wert waren und durch Schätzwerte ersetzt wurden

V4	Beruf	in	0 Datensätzen
V6	Geschlecht	in	0 Datensätzen
V8	Gewerkschaft	in	5 Datensätzen
V10	Gastarbeiter	in	3 Datensätzen
V5	Alter	in	0 Datensätzen
V9	Kirchgang	in	4 Datensätzen

**\*\*\*\*\* 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 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 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	V3	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 entgeltliche 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	U4	U5	
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	
			46.2	155	
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 -->

Speicher fuer x Variable

1 Vereinbare Variable= 25 ;

2  Option: Weitere Vereinbarungen - nur wenn Almo dazu auffordert

3 Datei A

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

12 die Datei umfasst so viele Variable

4 Datei B

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

12 die Datei umfasst so viele Variable

gemeinsame Variable aus Datei A und Datei B

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

spezifische Variable aus Datei B

U2, 7

Neue zusammengefasste Datei

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 anschauliche Datei im freien Format mit der Erweiterung \_\_.fre

### P45.8.3.2 Erläuterung zu den Boxen

#### Box 1 und Box 2:

Siehe Anhang P0.1 und P0.2.

**Box 3:** Datei A

**Box 4:** Datei B

Datei A	
<input \"="" c:\almo6\fusion\neudatena.dir\"="" type="text" value="\"/>	
<input type="text" value="12"/>	\die Datei umfasst so viele Variable

Datei B	
<input \"="" c:\almo6\fusion\neudatenb.dir\"="" type="text" value="\"/>	
<input type="text" value="12"/>	\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.

#### Box 5: Gemeinsame Variable aus Datei A und Datei B

gemeinsame Variable aus Datei A und Datei B	
<input type="text" value="V1, 2, 3, 4, 5, 6, 7, 8, 9, 12"/>	\Variablennummern aus Datei A
<input type="text" value="V1, 2, 11, 4, 5, 6, 7, 8, 12, 3"/>	\Variablennummern aus Datei B
<input type="text" value="10"/>	\Zahl der gemeinsamen Variablen

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

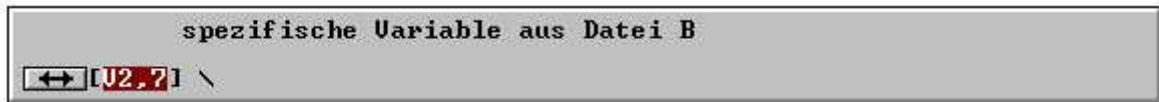
Vorne steht 'V' darauf folgen die Nummern - getrennt durch Beistrich. Zulässig wäre auch die Kurzschreibweise.

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.

**Box 6:** Spezifische Variable aus Datei B



Die spezifischen Variablen aus Datei B in unserem Beispiel sind:

- V2 Parteimitgliedschaft
- V7 ZeitungLesen

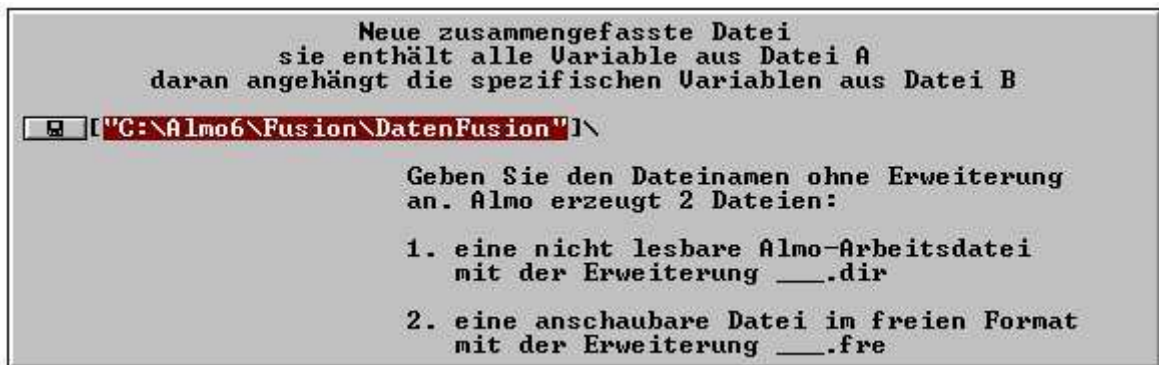
Schreiben Sie nur die Variablennummern. Sie müssen in folgender Form geschrieben werden:

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.

**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 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).