Ausreisser erkennen

Daten von Ausreissern bereinigen

Kurt Holm

Almo Statistik-System
www.almo-statistik.de
holm@almo-statistik.de
kurt.holm@jku.at

2014

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 kostenlos von der Handbuchseite in

www.almo-statistik.de

heruntergeladen werden

0. Arbeiten_mit_Almo.PDF (1 MB)
1. 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 (0,8 MB)
23. Ausreisser entdecken (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 Mehrfach-antworten (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)

Was sind Ausreisser ?

Ausreisser sind Werte, die ausserhalb "valider Grenzen" liegen. Die "validen Grenzen" definiert der Forscher. Anders formuliert: Es gibt keine "objektive", eindeutige Definition, was ein Ausreisser ist. Der Forscher legt fest, was für ihn ein Ausreisser ist. Werden Ausreisser vom Forscher aus der Analyse ausgeschlossen, dann tut er dies, weil er unterstellt, dass diese Daten - obwohl empirisch gewonnen - falsch sind oder er tut dies, weil sie ihm einen Variablen-Zusammenhang seiner Meinung nach verfälschen.

In Almo werden 2 Typen von Ausreissern unterschieden:

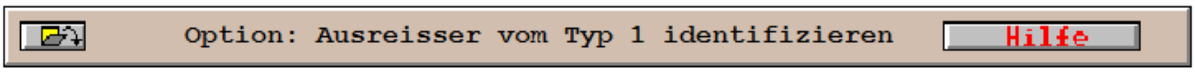
Ausreisser vom Typ 1:
Ein Variablenwert liegt ausserhalb des "validen Wertebereichs"
der Variablen. Hier können nochmals 2 Untertypen unterschieden werden
a. Schreibfehler
b. Extremwerte

Ausreisser vom Typ 2:
Ein Variablenwert liegt ausserhalb der "validen Punktewolke"
eines mehrdimensionalen Variablen-Zusammenhangs.

Zu Typ 1a: **Schreibfehler oder Messfehler als Ausreisser**
Ausreisser entstehen sehr oft dadurch, dass beim Schreiben der Daten Fehler gemacht werden. Beispiel: Anstelle 9 wird versehentlich der Wert 99 geschrieben. Diese Art der Ausreisser sollte der Benutzer durch das Programm Prog03m versuchen zu entdecken. Dieses Programm untersucht, ob Variablenwerte auftreten, die ausserhalb der zulässigen Unter- und Obergrenzen liegen. Diese muss der Benutzer in das Programm eingeben. Wird beispielweise Geschlecht (männlich, weiblich) mit 1 und 2 kodiert, dann liegt der Wert 3 ausserhalb der zulässigen Unter- und Obergrenzen und beruht auf einem Schreibfehler. Prog03m findet man durch Klick auf den Knopf "Verfahren", dann Eintrag "Fehlersuche". Diese Schreibfehler sollte man, bevor man mit der Datenanalyse überhaupt beginnt, bereinigen, d.h. in den Daten selbst korrigieren.
Messfehler können, müssen aber nicht, ausserhalb der zulässigen Wertegrenzen liegen. Liegen Sie innerhalb der Wertegrenzen werden sie in der Regel nicht entdeckt. Am ehesten gelingt es noch, sie als Ausreisser vom Typ 2 zu entdecken, d.h. innerhalb eines mehrdimensionalen Variablen-Zusammenhangs.

Zu Typ 1b: **Extremwert als Ausreisser**

Natürlich gibt es auch "echte" Ausreisser, die nicht durch Schreibfehler entstanden sind. Beispiel: Für eine Stichprobe von 1000 Personen wird das Einkommen erhoben. Dabei sind einige wenige Milliardäre in die Stichprobe gelangt. Deren Einkommen liegt ausserhalb des "validen Wertebereichs". Wird nun die Korrelation zwischen Einkommen und beispiesweise Schulbildung ermittelt, so kann der Korrelationskoeffizient durch die Milliardäre dramatisch verändert werden. Hier ist es sinnvoll, die Milliardäre als Ausreisser zu identifizieren und aus der Analyse auszuschliessen. Für diesen Zweck ist die Optionsbox

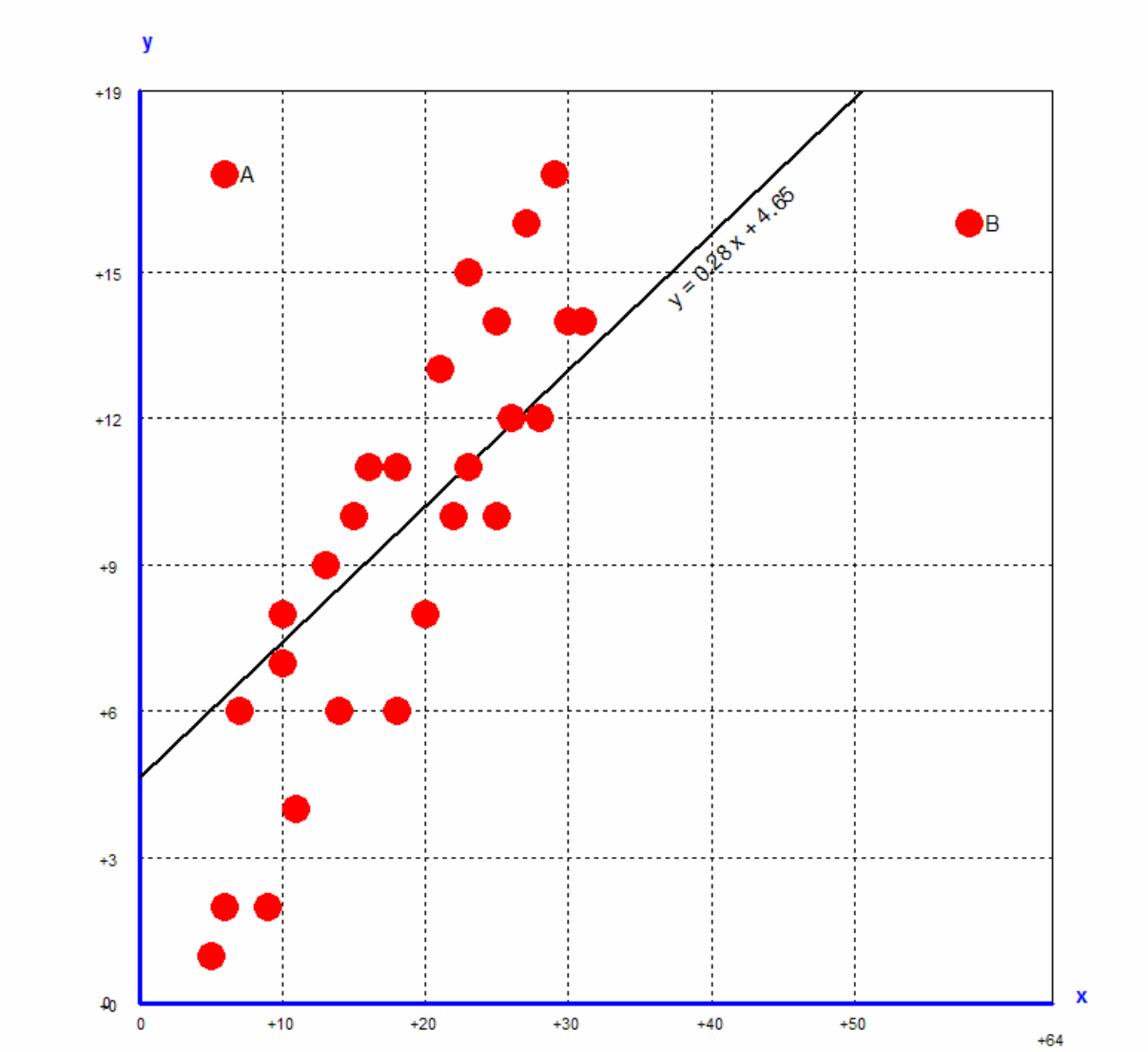


gedacht, die in vielen Almo-Maskenprogrammen angeboten wird. Natürlich kann Sie auch verwendet werden, um Ausreisser vom Typ 1a, also Schreibfehler und Messfehler, zu finden. Der Benutzer kann auch **Prog05m3** einsetzen. Dies ist ein spezielles Programm, dessen Zweck folgender ist: Die Ausreisser vom Typ 1 sollen identifiziert werden und es soll eine neue "Ausreisser-bereinigte" Datei erstellt werden. Diese kann dann für weitere Analysen verwendet werden. Man findet dieses Programm durch Klick auf den Knopf "Verfahren", dann "Ausreisser".

Zu Typ 2: **Ausreisser liegt ausserhalb der "validen Punktewolke"**

Betrachten wir ein Beispiel:

`Grafik 1`



Der Zusammenhang zwischen dnr Variablen x und y wird durch ein Streudiagramm grafisch dargestellt. Die kleinen roten Punkte sind Messpunkte. Die durchgezogene Linie ist die Regressionsgerade.

Der Messpunkt B ist ein Ausreisser vom Typ I. Sein x-Wert liegt weit ausserhalb des validen Wertebereichs von x. Ausreisser vom Typ 1 sind also in der Regel auch Ausreisser vom Typ 2. Anders formuliert, mit der in Almo angebotenen Methode zur Identifizierung von Ausreissern vom Typ 1 (mit der Optionsbox "Ausreisser vom Typ 1 identifizieren") wird auch ein Teil der Ausreisser vom Typ 2 ermittelt - aber eben nur ein Teil.

Der Messpunkt A ist ein Ausreisser vom Typ II. Sein x-Wert und sein y-Wert liegt zwar innerhalb des validen Wertebereichs von x und y. In Bezug auf den Zusammenhang von x und y ist er jedoch ein Ausreisser. Er liegt ausserhalb der "validen Punktewolke xy".

Um Ausreisser vom Typ 2 zu identifizieren muss der Benutzer das Programm Prog20bm verwenden. Man findet dieses Programm durch Klick auf den Knopf "Verfahren", dann "Ausreisser".

Ausreisser vom Typ 1 erkennen

Almo bietet folgende Möglichkeiten an, Ausreisser vom Typ 1 in den Analysevariablen zu identifizieren:

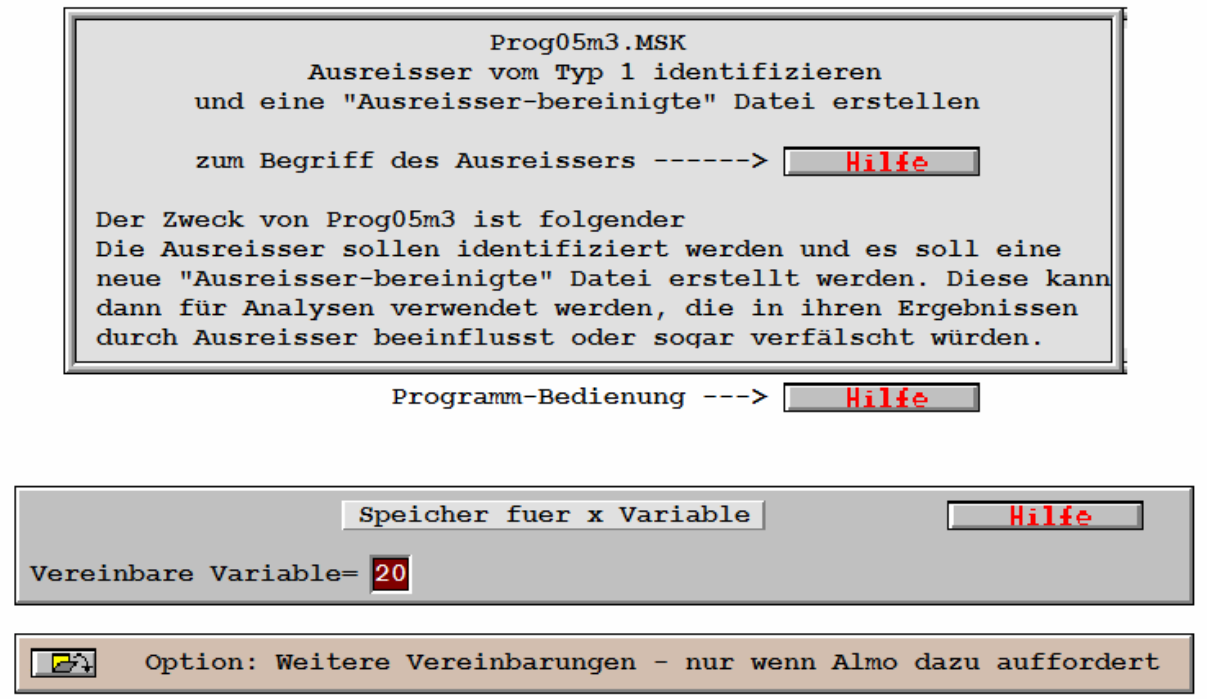
(1) Programm Prog03m ist speziell für das Auffinden von Schreibfehlern gedacht. Es untersucht, ob Variablenwerte auftreten, die ausserhalb der zulässigen Unter- und Obergrenzen liegen. Prog03m findet man durch Klick auf den Knopf "Verfahren", dann Eintrag "Fehlersuche".

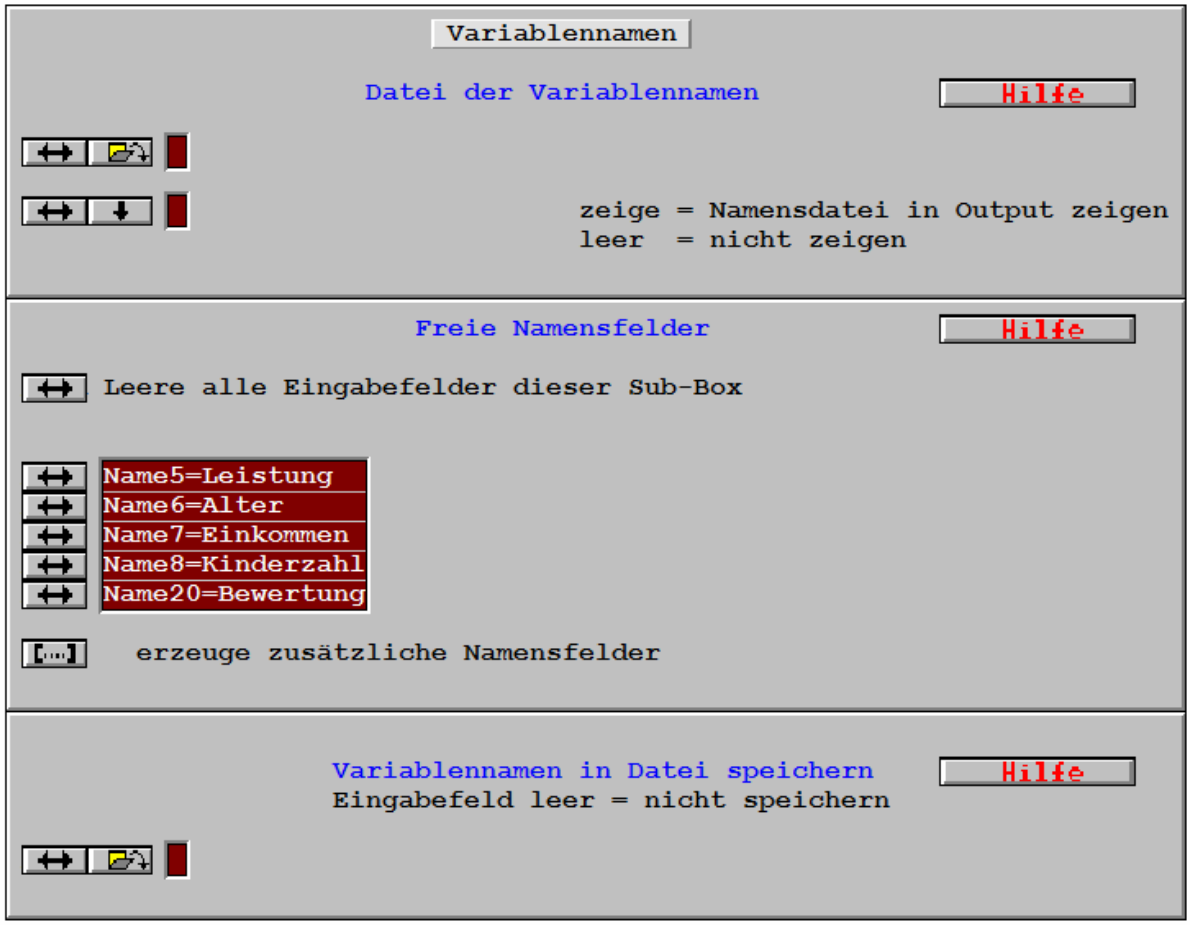
(2) In vielen Programmen, z.B. dem Korrelationsprogramm Prog19bm, wird eine Option angeboten, die es erlaubt, Ausreisser vom Typ 1 zu identifizieren und zu bereinigen - so dass z.B. Korrelationen nicht durch Ausreisser vom Typ 1 beeinflusst werden. Nominale Variable können über diese Option nicht auf Ausreisser untersucht werden.

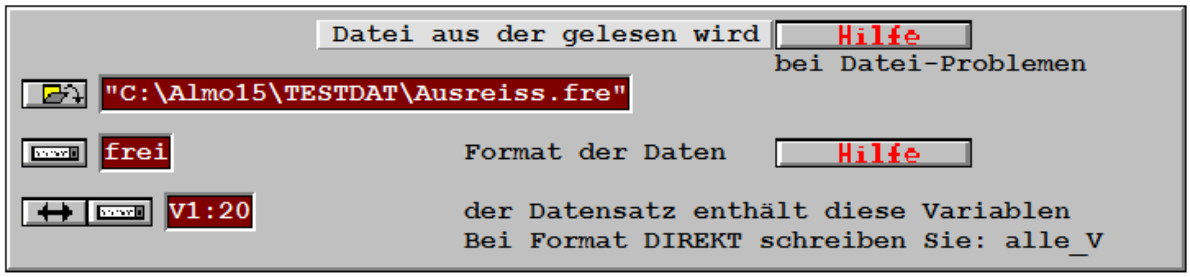
(3) Prog05m3 ist ein spezielles Programm, dessen Zweck folgender ist: Die Ausreisser vom Typ 1 sollen identifiziert werden und es soll eine neue "Ausreisser-bereinigte" Datei erstellt werden. Diese kann dann für Analysen verwendet werden, die in ihren Ergebnissen durch Ausreisser vom Typ 1 beeinflusst oder sogar verfälscht würden. Sie finden Prog05m3 durch Klick auf den Knopf "Verfahren", dann "Ausreisser". Nominale Variable können in Prog05m3 nicht auf Ausreisser untersucht werden.

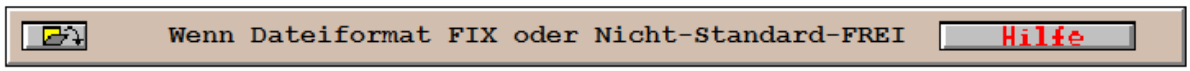
In den unter (2) erwähnten Almo-Programm-Masken und auch im speziellen in (3) genannten Prog05m3 ist dieselbe "Ausreisser-Optionsbox" enthalten. Sie wird im folgenden ausführlich erläutert.

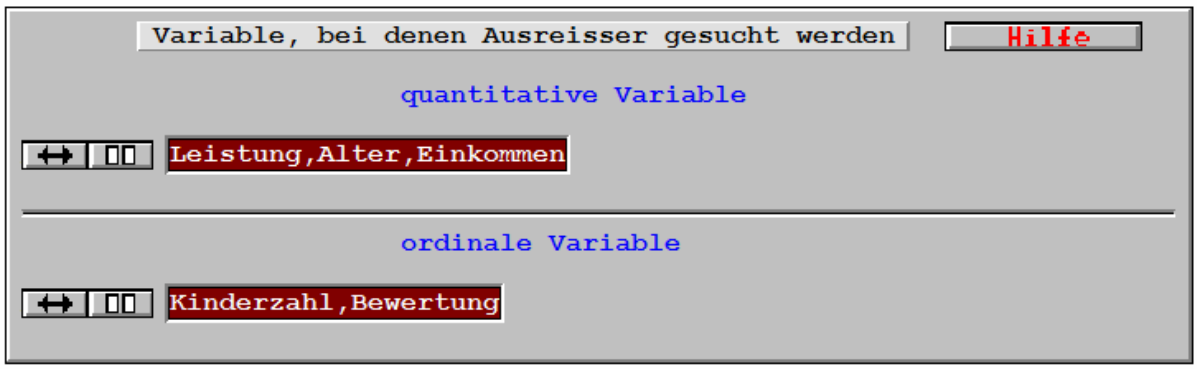
Programm-Maske Prog05m3



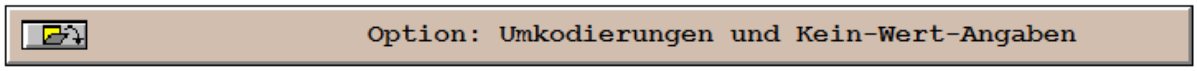

Variablennamen
Datei der Variablennamen
Hilfe
zeige = Namensdatei in Output zeigen
leer = nicht zeigen
Freie Namensfelder
Hilfe
Leere alle Eingabefelder dieser Sub-Box
Name5=Leistung
Name6=Alter
Name7=Einkommen
Name8=Kinderzahl
Name20=Bewertung
erzeuge zusätzliche Namensfelder
Variablennamen in Datei speichern
Hilfe
Eingabefeld leer = nicht speichern

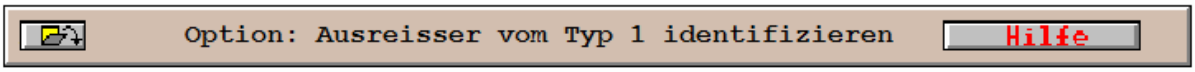

Datei aus der gelesen wird
Hilfe
bei Datei-Problemen
"C:\Almo15\TESTDAT\Ausreiss.fre"
frei
Format der Daten
Hilfe
V1:20
der Datensatz enthält diese Variablen
Bei Format DIREKT schreiben Sie: alle_V


Wenn Dateiformat FIX oder Nicht-Standard-FREI
Hilfe


Variable, bei denen Ausreisser gesucht werden
Hilfe
quantitative Variable
Leistung,Alter,Einkommen
ordinale Variable
Kinderzahl,Bewertung

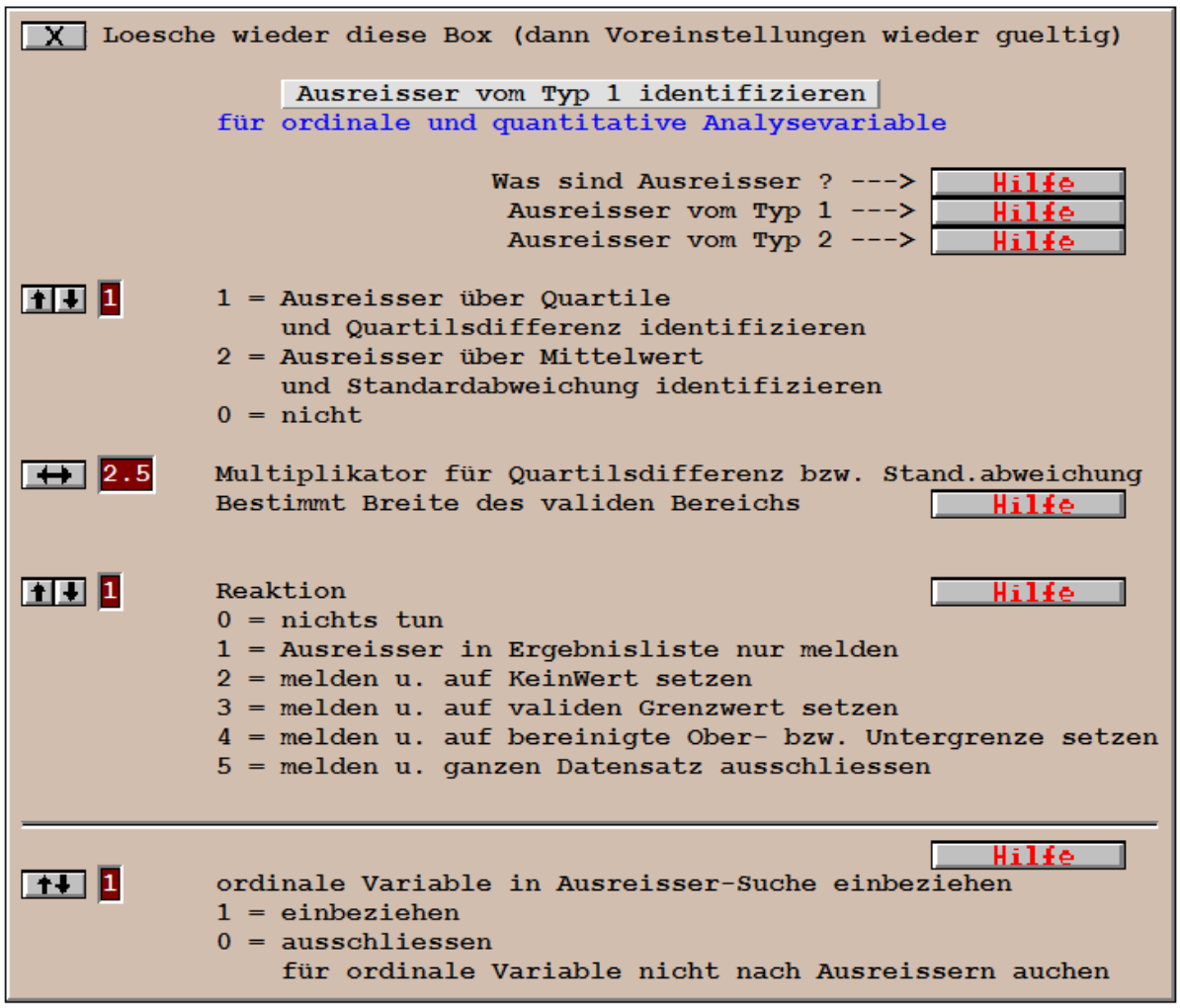

Option: Ein- und Ausschliessen von Untersuchungseinheiten


Option: Umkodierungen und Kein-Wert-Angaben

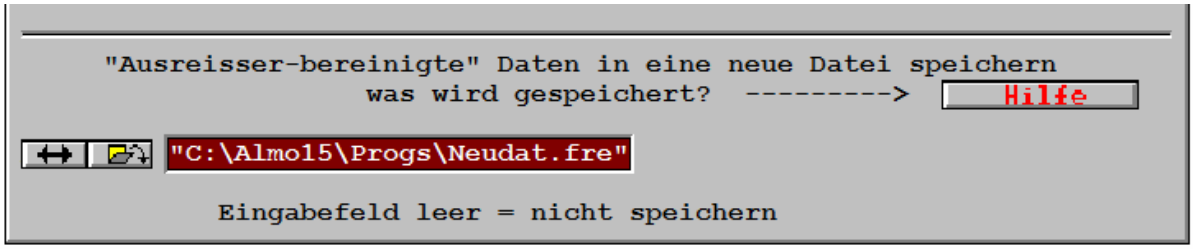


Diese Eingabe-Box ist nicht nur inProg05m3 enthalten, sondern auch in vielen Almo-Programm-Masken, bei denen es sinnvoll ist, Ausreisser zu identifizieren und in einer bestimmten Weise zu behandeln.

Wird die Optionsbox geöffnet, dann sieht man folgendes. Im speziellen Ausreisser-Programm Progo5m3 ist die "Ausreisser-Box" bereits in geöffneter Form enthalten



Im speziellen Ausreisser-Programm Prog05m3 wird diese Optionsbox noch durch folgenden Teil ergänzt



Wir werden nur die "Ausreisser-Box" betrachten. Die übrigen Eingabe-Boxen sind alle

ausführlich im Almo-Dokument 0 "Arbeiten mit Almo" ausführlich erläutert.

Im *1. Eingabefeld* der Ausreisser-Optionsbox bietet Almo 2 Methoden an, einen Variablenwert als Ausreisser vom Typ 1 zu identifizieren.

Methode 1: Ausreisser-Identifizierung über Quartile und Quartilsabstand

Für jede quantitative und ordinale Analysevariable wird das 1. und 3. Quartil und der Quartilsabstand ermittelt.

Wenn ein Variablenwert ausserhalb des "validen Bereichs" von

```
1. Quartil-x*Quartilsabstand
und
3. Quartil+x*Quartilsabstand
```

liegt, dann betrachtet Almo ihn als Ausreisser.

Der Benutzer kann x, den Multiplikator für den Quartilsabstand im *2. Eingabefeld* beliebig verändern. Er bestimmt damit die Breite des "validen Bereichs", in dem ein Variablenwert als "valide", also nicht als Ausreisser betrachtet wird.

Der Multiplikator wird im 2. Eingabefeld eingetragen.
Für x sollte ein Wert zwischen ca. 1.5 und 3.5 eingesetzt werden.

Generell gilt (auch für die folgende Methode 2), dass man mit verschiedenen x-Werten experimentieren sollte.

Methode 1 kann nur verwendet werden, wenn die Zahl der diversen Werte, die eine Analysevariable annimmt, kleiner 500 ist.

Anmerkung: Werden Residuen auf Ausreisser untersucht (wie in Prog20bm), dann sollte Methode 1 nur verwendet werden, wenn die Zahl der Datensätze kleiner 500 ist, da die Werte der Residuen alle voneinander verschieden sein können (so dass die Zahl der diversen Werte gleich der Zahl der Datensätze sein kann).

Methode 2: Ausreisser-Identifizierung über Mittelwert und Standardabweichung

Für jede quantitative und ordinale Analysevariable wird das arithmetische Mittel und die Standardabweichung ermittelt. Beachte: Auch für die ordinalen Variablen wird für die Ausreisser-Identifizierung das arithmetische Mittel und die Standardabweichung berechnet. Die ordinalen Variablen werden also so behandelt, wie wenn sie quantitativ wären.

Wenn ein Variablenwert ausserhalb des "validen Bereichs" von

```
Mittelwert-x*Standardabweichung
und
Mittelwert+x*Standardabweichung
```

liegt, dann betrachtet Almo ihn als Ausreisser.

Der Benutzer kann x, den Multiplikator der Standardabweichung im *2. Eingabefeld* beliebig

verändern. Er bestimmt damit die Breite des "validen Bereichs", in dem ein Variablenwert als "valide", also nicht als Ausreisser betrachtet wird.

Für x sollte ein Wert zwischen ca. 2.5 und 5 eingesetzt werden.

Grubbs-Test

Wird Methode 2 gewählt, dann errechnet Almo den Test nach Grubbs. Dieser setzt voraus, dass die Daten normalverteilt sind - was zuvor durch einen Test auf Normalverteilung überprüft werden muss (z.B. mit Prog04m2 durch einen Chi-Quadrat-Test oder den Kolmogorov-Smirnov-Test).

Beim Grubbs-Test wird zuerst der maximale Datenwert Xmax ermittelt. Dieser wir dann mit den von Grubbs entwickelten Formeln daraufhin untersucht, ob er sich in die normalverteilte Datenmenge X1....Xn einfügt oder ob er ein Ausreisser ist. Der entdeckte Ausreisser wird eliminiert und dann das Verfahren mit dem nächsten Maximum wiederholt usw. Dieses Verfahren wurde durch einen von Rossner entwickelten Test ergänzt, der versucht, die Zahl der signifikanten Ausreisser zu ermitteln. Siehe dazu in der Literatur-Angabe "NIST-Agency", Abschnitt 1.3.5.17.3.

In Almo werden in einem Datendurchlauf alle ausserhalb des (vom Benutzers definierten) validen Bereichs liegenden Werte durch den Grubbs-Test überprüft. Es wird in folgender Weise verfahren:

Der ausserhalb des validen Bereichs liegender Wert wird standardisiert nach der Formel

```
(1) Gx = (X-M)/s
```

```
Gx=standardisierter Wert
X =Rohwert
M =Mittelwert
s =Standardabweichung
```

X ist ein Ausreisser, wenn Gx grösser ist als Gz, wobei

```
(2) Gz =  [(n-1)/sqrt(n)] * [sqrt( tq / (n-2+tq) )]
```

```
sqrt = Wurzel aus(...)
n    = Zahl der Untersuchungsobjekte
tq   = ist der quadrierte t-Wert für die Signifikanz alpha/2n
       mit n-2 Freiheitsgraden
alpha= ist die vom Benutzer festgelegt Signifikanz
```

Almo unterstellt, dass der Benutzer mit dem von ihm eingegebenen Multiplikator auch das Signifikanzniveau alpha für den Grubbs-Test festgelegt hat. Hat der Benutzer den Multiplikator z.B. auf 2 (genauer 1.96) gesetzt, dann ist alpha=0.05. Bei einem Multiplikator von 2.5 (genauer 2.58) ist alpha=0.01 Almo ermittelt den (quadrierten) t-Wert für alpha/2n und errechnet gemäß (2) den Gz-Wert.

Liegt Gx über Gz, dann ist der Wert ein Ausreisser aus der normalverteilten Datenmenge X1...Xn

BEACHTE:
Der Grubbs-Test wird von Almo nur zur Information ausgegeben. Er entscheidet nicht darüber, ob ein Variablenwert als Ausreisser deklariert wird oder nicht. Wenn also beispielsweise ein Variablenwert als Ausreisser identifiziert wurde (weil er ausserhalb des validen Bereichs liegt), der Grubbs-Test jedoch diesen nicht als Ausreisser anerkennt, dann bleibt er trotzdem ein Ausreisser und wird entsprechend der vom Benutzer gewählten Reaktion behandelt.

Reaktion

Im *3. Eingabefeld* kann der Benutzer festlegen, wie Almo auf identifizierte Ausreisser reagieren soll. Folgende Möglichkeiten gibt es dabei:

Wird in der Optionsbox im 3. Eingabefeld eine 0 eingesetzt, dann reagiert Almo nicht. Es erfolgt auch keine Warnung. Der Ausreisser wird wie ein valider Wert behandelt.

Wird eine 1 eingesetzt,
dann bringt Almo die Warnung, dass dieser Wert ein Ausreisser ist.

Wird eine 2 eingesetzt,
dann bringt Almo die Warnung, dass dieser Wert ein Ausreisser ist und setzt den Variablenwert auf KeinWert. Die Folge davon ist in der Regel, dass dieser Wert für die Analyse nicht berücksichtigt wird.

Wird eine 3 eingesetzt,
dann bringt Almo die Warnung, dass dieser Wert ein Ausreisser ist und setzt den Variablenwert auf den nächst liegenden validen Grenzwert. Beispiel: Die Grenzwerte des validen Bereich sind -3 und +10. Hat der Ausreisser z.B. den Wert 15, dann wird er auf +10 gesetzt. Hat der Ausreisser z.B. den Wert -5, dann wird er auf -3 gesetzt. Siehe nachfolgende Erläuterung zum Begriff "valider Grenzwert". Diese Reaktion ist nicht möglich, wenn eine Residuen-Variable auf Ausreisser untersucht wird (wie in Prog20bm).

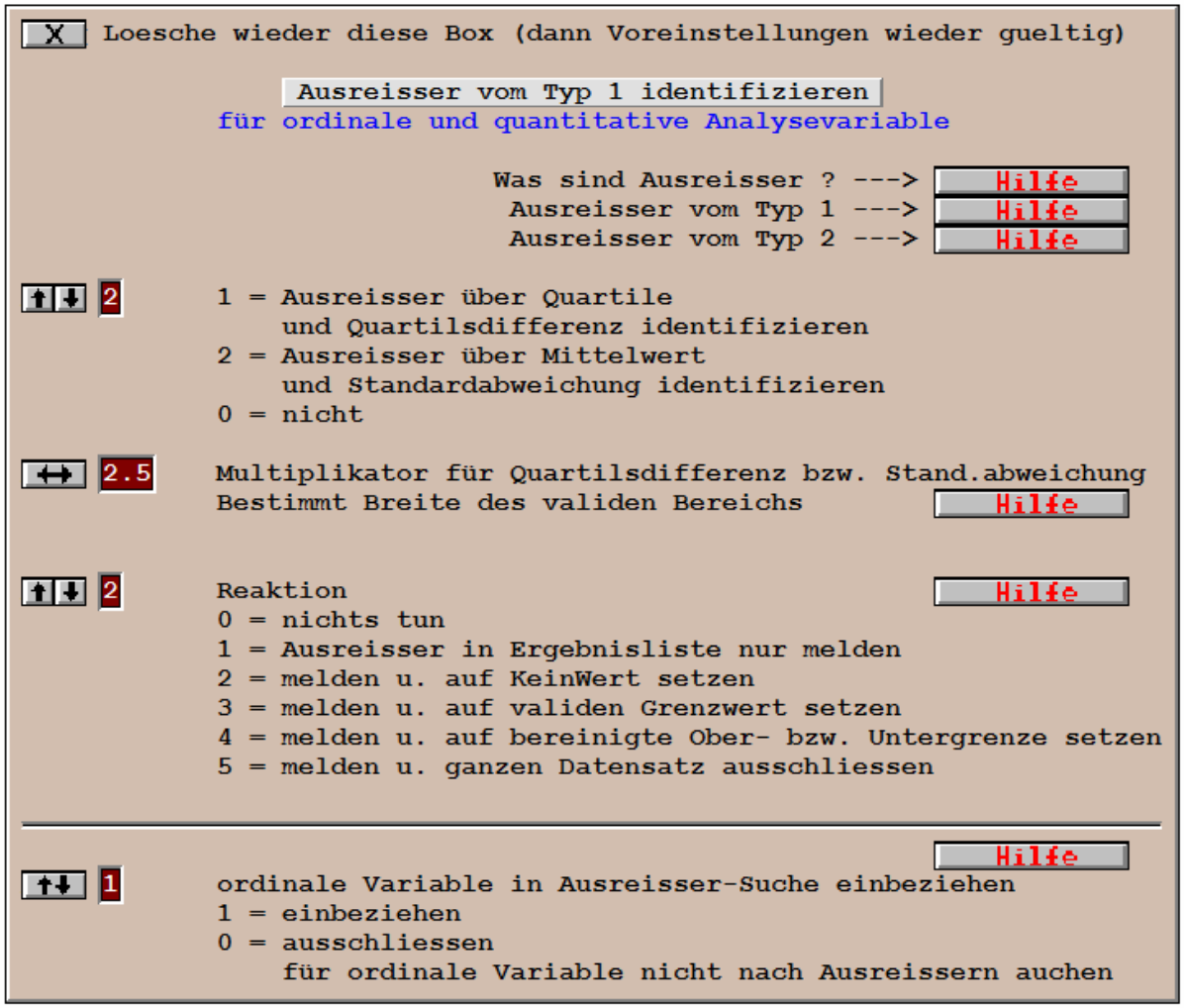
Wird eine 4 eingesetzt,
dann bringt Almo die Warnung, dass dieser Wert ein Ausreisser ist und setzt den Variablenwert auf die "Ausreisser-bereinigte Unter- bzw. Obergrenze" Beispiel: Die Ausreisser-bereinigte Untergrenze ist 0 und die Ausreisser-bereinigte Obergrenze ist 13.
Hat der Ausreisser z.B. den Wert 15 und liegt damit oberhalb des "oberen validen Grenzwerts", dann wird er auf die "bereinigte Obergrenze", also auf 13 gesetzt. Hat der Ausreisser z.B. den Wert -5 und liegt damit unterhalb des, "unteren validen Grenzwerts", dann wird er auf die"bereinigte Untergrenze", also auf 0 gesetzt. Siehe nachfolgende Erläuterung zum Begriff"Ausreisser-bereinigte Unter- bzw. Obergrenze" Diese Reaktion ist nicht möglich, wenn Residuen auf Ausreisser untersucht werden (wie in Prog20bm).

Wird eine 5 eingesetzt,
dann bringt Almo die Warnung, dass dieser Wert ein Ausreisser ist und schliesst den gesamten Datensatz aus der Analyse aus.

Empfehlung: Wir empfehlen, zunächst eine 1 einzusetzen, sich also die Ausreisser zunächst nur melden zu lassen. Dabei kann der Benutzer verschiedene Werte für den Multiplikator ausprobieren. Erst dann sollte er eine 2 oder 3 oder 4 oder 5 einsetzen. Die klarste Lösung des Ausreisser-Problems entsteht sicherlich durch Reaktion 5 "gesamten Datensatz überspringen"

Ausgabe aus Prog05m3

Das Prog wurde mit folgenden Einstellungen gerechnet.



Mit diesen Einstellungen entstand folgende Ausgabe

```
          Ein Wert wird als Ausreisser identifiziert, wenn er ausserhalb
des validen Bereichs von Mittelwert plus  2.5 * Standardabweichung   und
                         Mittelwert minus 2.5 * Standardabweichung liegt

                                        ******************** MITTEILUNG
                        Fuer die Identifikation der Ausreisser wurden
                  folgende Mittelwerte und Standardabweichungen ermittelt

Variable              arithmetisches Mittel     Standardabweichung
   8 Kinderzahl              4.65574                  2.52778
  20 Bewertung               4.63934                  3.19853
   5 Leistung                5.86885                  8.87892
   6 Alter                   4.01639                   10.803
   7 Einkommen               4.19672                  4.47147

                                        ******************** MITTEILUNG
       Fuer die Identifizierung von Ausreissern wurden folgende kleinste
und groesste Werte als gerade noch valide angenommen (valide Grenzwerte)
       Werte, die ausserhalb liegen, werden als Ausreisser identifiziert


                                    valide Grenzwerte
```

```
Variable              valider kleinster Wert   valider groesster Wert
   8 Kinderzahl              -1.66372                 10.9752
  20 Bewertung               -3.35698                 12.6357
   5 Leistung                -16.3285                 28.0662
   6 Alter                   -22.9911                 31.0239
   7 Einkommen               -6.98197                 15.3754


                                          ******************** WARNUNG
Datensatz 5:  V5 Leistung      = 44 ist ein Ausreisser    wird auf KeinWert gesetzt
                              Gx=4.295 Gz=3.624 p=0.015 *)

Datensatz 6:  V6 Alter         = -55 ist ein Ausreisser    wird auf KeinWert gesetzt
                              Gx=5.463 Gz=3.624 p=0.015 *)

Datensatz 7:  V6 Alter         = 61 ist ein Ausreisser    wird auf KeinWert gesetzt
                              Gx=5.275 Gz=3.624 p=0.015 *)

Datensatz 7:  V7 Einkommen     = 33 ist ein Ausreisser    wird auf KeinWert gesetzt
                              Gx=6.442 Gz=3.624 p=0.015 *)

Datensatz 8:  V20 Bewertung    = 19 ist ein Ausreisser    wird auf KeinWert gesetzt
                              Gx=4.490 Gz=3.624 p=0.015 *)

Datensatz 8:  V5 Leistung      = 44 ist ein Ausreisser    wird auf KeinWert gesetzt
                              Gx=4.295 Gz=3.624 p=0.015 *)

Datensatz 9:  V5 Leistung      = 44 ist ein Ausreisser    wird auf KeinWert gesetzt
                              Gx=4.295 Gz=3.624 p=0.015 *)

*) Gx=standardisierter Datenwert
   Gz=Grenzwert
   p=Signifikanz
   Datenwert ist Ausreisser wenn Gx groesser Gz
   Voraussetzung: Daten muessen normalverteilt sein

Zahl der eingelesenen Datensaetze:      61

                                          ******************** WARNUNG
                                  Es wurden 7 Ausreisser identifiziert
                                        Prozentanteil der Ausreisser
                             an Zahl der eingelesenen Datensaetze: 11.48 %
```

Begriffe "valider Grenzwert" und "Ausreisser-bereinigte Unter- bzw. Obergrenze"

Rechnen Sie Prog05m3. Sie finden dieses Prog unter "Verfahren / Ausreisser". Die Variable V6 nimmt folgende diversen Werte an:

```
-55  0  1  2  3  4  5  6  7  8  9  13  61
```

Almo errechnet für diese Variable mit Methode 1 (Quartilsdifferenz) einen "validen Bereich" mit dem "unteren validen Grenzwert" = -8
und dem "oberen validen Grenzwert" = 16

Würden die Ausreisser, die diesen 'validen Bereich' verlassen aus der Datei herausgenommen, dann würde der kleinste vorkommende Wert 0 und der grösste vorkommende Wert 13 sein. Diese bezeichnen wir als "Ausreisser-bereinigte Untergrenze" und "Ausreisser-bereinigte Obergrenze" oder auch kurz als "bereinigte Unter- bzw. Obergrenze"

Abfolge

1. Schritt: Almo berechnet in einem 1. Daten-Durchlauf (bei Methode 1) die Quartile und den

Quartilsabstand bzw. bei Methode 2 den Mittelwert und die Standardabweichung. Dabei werden die Ausreisser mitgerechnet.

2. Schritt: In einem 2. Daten-Durchlauf werden - bevor die Ausreisser identifiziert werden - vom Benutzer eventuell geschriebene Ein- bzw. Ausschluss-Optionen sowie Umkodierungen und KeinWert-Angaben ausgeführt. Dann erst erfolg die Ausreisser-Identifizierung.

Beispiel: Der valide Bereich der Variablen V55 reicht von 1 bis 10. Ein Wert von 30 ist dann ein Ausreisser. Der Benutzer schreibt in das Eingabefeld für das Umkodieren folgende Anweisung:

```
V55(1:5=1; 6:100=2)
```

Dadurch wird auch der Ausreisser mit V55=30 auf 2 umkodiert und wird dadurch nicht mehr identifiziert.

Gewichtungen haben keinen Einfluß auf die Ausreisser-Identifizierung. Bei ihnen werden Datensätze gewichtet (z.B. dupliziert) und nicht Variablenwerte.

Problem

Wie oben ausgeführt, berechnet Almo in einem 1. Daten-Durchlauf (bei Methode 1) die Quartile und den Quartilsabstand bzw. bei Methode 2 den Mittelwert und die Standardabweichung. Dabei werden die - bis jetzt noch nicht entdeckten Ausreisser - mitgerechnet. Die Ausreisser-Werte beeinflussen also die Berechnung dieser Größen.

Wenn sehr extreme Ausreisser-Werte auftreten, dann kann bei der 2. Methode der Mittelwert und die Standardabweichung sehr stark beeinflusst werden. Bei Methode 1 ist die Auswirkung extremer Ausreisser-Werte sehr viel geringer,so dass diese Methode vorzuziehen ist. Erst in einem 2. Daten-Durchlauf werden dann, wie oben beschrieben, die Variablenwerte daraufhin überprüft, ob sie Ausreisser sind oder nicht.

Eine 3. Methode

Im 2. Daten-Durchlauf werden - bevor die Ausreisser identifiziert werden - vom Benutzer geschriebene Ein- bzw. Ausschluss-Optionen sowie Umkodierungen und KeinWert-Angaben ausgeführt.

Beispiel: Der Benutzer schreibt in der Optionsbox "Ein- und Ausschliessen von Untersuchungseinheiten" in das Eingabefeld für das Ausschliessen folgende Anweisung:

```
V55 groesser 9
```

Hat ein Ausreisser z.B. den Wert 10 dann wird er bereits durch diese Anweisung abgefangen (und eliminiert) - so dass er als Ausreisser durch die oben beschriebenen Methoden 1 oder 2 nicht mehr identifiziert wird - nicht mehr identifiziert werden muss.

Ein weiteres Breispiel: Der Benutzer schreibt in der Optionsbox "Umkodierungen und Kein-Wert-Angaben" die Anweisung:

```
V55 ( 9.001 bis 1000 = KeinWert )
```

Auch hier wird der Ausreisser mit dem Wert 10 abgefangen und eliminiert. Oder er schreibt:

```
V55 ( 9.001 bis 1000 = 9 )
```

Der Ausreisser mit dem Wert 10 wird auf den noch validen Wert 9 gesetzt.

Eine sehr "elegante" Umkodierung ist folgende

```
V55 ( 1:9 = V55;  Sonst = KeinWert )
```

Wenn der Wert der Variablen V55 im validen Bereich von 1 bis 9 liegt, dann wird er auf sich selbst gesetzt - sonst (also wenn er ausserhalb des validen Bereichs liegt) wird er auf KeinWert gesetzt.

Der Benutzer kann als, indem er die Optionsbox für das Ein- und Ausschliessen oder die Optionsbox für das Umkodieren in der beschriebenen Art und Weise benutzt, sehr zielgerichtete und flexibel Ausreisser aus der Analyse herausnehmen.

Ausreisser vom Typ 1 bei nominalen Variablen

Bei nominalen Variablen dient der Variablenwert, der einer Ausprägung zugewiesen wird, nur der Klassifikation. Die nominale Variable "Geschlecht" kann mit 1 (=männlich) und 2 (=weiblich) kodiert werden, aber auch mit 1 (=männlich) und 99 (=weiblich). Diese Kodierung ist zwar ungewöhnlich, aber korrekt. Der Wert 99 ist dabei kein Ausreisser. Wenn allerdings ein 3. Wert auftritt, dann beruht dieser auf einem Fehler beim Schreiben der Daten. Wir haben oben bereits darauf hingewiesen, dass man diese Schreibfehler mit Prog03m identifizieren sollte.

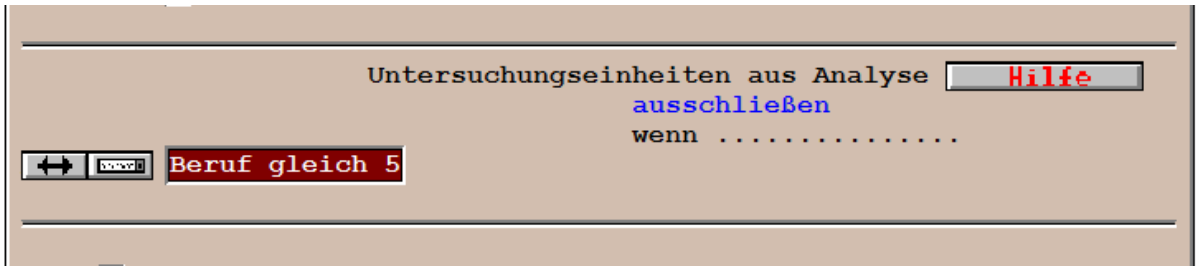
In Almo werden deswegen die nominalen Variablen standardmäßig nicht in die Ausreisser-Identifikation für den Typ 1 miteinbezogen.

Eine Möglichkeit besteht jedoch darin, mit Prog04m1 oder Prog04m2 die Häufigkeiten der Ausprägungen der betreffenden nominalen Variablen auszuzählen.
Betrachten wir ein Beispiel. Bei dem Versuch, das Einkommen von Akademikern durch die Variable Beruf (durch eine Varianzanalyse) zu erklären, wurden folgende Berufshäufigkeiten gefunden:

Code		Häufigkeit
1	praktischer Arzt	80
2	Facharzt	33
3	Architekt	21
4	Rechtsanwalt	110
5	Theaterintendant	1
	..	
	..	
	..	

In den Untersuchungsdaten befindet sich nur 1 Theaterintendant. Soll er in die Varianzanalyse miteinbezogen werden oder soll er ausgeschlossen werden? Der Forscher muss das entscheiden. Er könnte so vorgehen: Mit Prog04m1 oder Prog04m2 wird die Häufigkeit der akademischen Berufe ermittelt. Im Varianzanalyse-Programm Prog20md wird dann in der Optionsbox

folgende Eingabe geschrieben



Prog04m1 bzw. Prog04m2 wird gefunden über Klick auf den Knopf "Verfahren", dann "Häufigkeitsverteilung", Prog20md über "Verfahren/Varianzanalyse"

Dem Benutzer bleibt weiterhin die Möglichkeit, die nominalen Variablen in Prog05m2 (über Knopf Verfahren / Basis-Statistiken) als quantitative zu deklarieren und somit auf Ausreisser zu untersuchen. Allerdings funktioniert das nur, wenn die nominale Variable genügend viele Ausprägungen besitzt.

Ausreisser von Typ 1 bei ordinalen Variablen

Im Prinzip gilt für die ordinalen Variablen ähnliches wie für die nominalen. Der einer Ausprägung zugewiesene Wert drückt nur die Rangfolge aus.
Beispiel:

```
Schulbildung
                    Code
                    ----
                      1 Volksschulabschluß
                      2 Hauptschulsabschluß
                      3 Gymnasium
                      3 Fachschule
                      4 Universitätsabschluß
                      5 Habilitation
```

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

Korrekt (aber wohl ungewöhnlich) wäre es aber auch so zu kodieren:

```
                    Code
                    ----
                      4 Volksschulabschluß
                     10 Hauptschulsabschluß
                     14 Gymnasium
                     14 Fachschule
                     16 Universitätsabschluß
```

An der Rangfolge der Werte ändert sich dadurch nichts.

Möglicherweise soll aber durch die abweichend hohe Kodierung der Habilitation ausgedrückt werden, dass diese sehr viel mehr ist als ein Universitätsabschluß. Erfahrungsgemäß wird auch bei Auswertungen ein und dieselbe Variable ein Mal als quantitativ und in einer anderen Analyse als ordinal behandelt. Als quantitative Variable würde die Habilitation mit 99 als Ausreisser identifiziert werden. In Almo werden deswegen die ordinalen Variablen in die Ausreisser-Identifikation miteinbezogen. Der Benutzer kann jedoch im *4. Eingabefeld* der Optionsbox entscheiden, ob er die ordinalen Variablen miteinbeziehen will oder nicht.

In Almo wird bei verschiedenen Verfahren (z.B. dem Allgemeinen Linearen Modell) obige ungewöhnliche Kodierung intern zwangsweise umkodiert auf 1,2,3,4,5. Almo beginnt mit 1 und geht dann mit Schrittweite 1 weiter. Wenn der Benutzer die Ausreisser-Identifikation auch auf ordinale Variable ausdehnt, dann wird (bei obiger ungewöhnlicher Kodierung) die Habilitation zuerst als Ausreisser erkannt und die gewählte Reaktion (z.B. "auf KeinWert setzen") durchgeführt. Erst danach wird die interne zwangsweise Umkodierung vorgenommen.

Hat der Benutzer "auf KeinWert setzen" als Reaktion gewählt, dann wird der Wert 99 auf KeinWert gesetzt und danach die Variable intern zwangsweise auf 1,2,3,4 umkodiert. Die Ausprägung "Habilitation" existiert dann nicht mehr - wird also nicht etwa auf 5 umkodiert.

Hat der Benutzer "auf validen Grenzwert setzen" als Reaktion gewählt, dann wird der Wert 99 auf etwa 20 gesetzt und danach die Variable intern zwangsweise auf 1,2,3,4,5 umkodiert. Die Ausprägung "Habilitation" existiert also weiterhin - wird aber automatisch auf 5 umkodiert.

Ausreisser vom Typ 2 mit Prog20bm identifizieren

Betrachten wir nochmals die Grafik 1

Der Zusammenhang zwischen der Variablen x und y wird durch ein Streudiagramm grafisch dargestellt. Die kleinen runden Kreise sind Messpunkte. Die gepunktete Linie ist die Regressionsgerade.

Der Messpunkt B ist ein Ausreisser vom Typ I. Sein x-Wert liegt weit ausserhalb des validen Wertebereichs von x.

Der Messpunkt A ist ein Ausreisser vom Typ II. Sein x-Wert und sein y-Wert liegt zwar innerhalb des validen Wertebereichs von x und y. In Bezug auf den Zusammenhang von x und y ist er jedoch ein Ausreisser. Er liegt ausserhalb der "validen Punktewolke xy".

Dieses Streudiagramm kann mit dem Programm "Ausreiss.Alm" erzeugt werden. Sie finden dieses Programm durch Klick auf das Menü "Almo" und den Eintrag "Liste aller Almo-Programme".

Almo bietet das Programm Prog20bm an, das Ausreisser vom Typ 2 in den Analysevariablen identifiziert und (optional) die "Ausreisser-bereinigten" Daten in eine neue Datei schreibt.

In Prog20bm wird das Allgemeine Lineare Modell verwendet um zu ermitteln, ob ein Messpunkt (so wie der Punkt A in obigem Streudiagramm) ausserhalb der "validen

Punktewolke" liegt. Danach wird eine "Ausreisser-bereinigte" Datei erstellt. Der Benutzer findet dieses Programm durch Klick auf den Knopf "Verfahren", dann "Ausreisser".

Prog20bm verläuft in folgenden Schritten:

1. Wir betrachten der Einfachheit halber zunächst den Fall der bivariaten Regressionsanalyse: x ist die unabhängige quantitative Variable und y ist die abhängige quantitative Variable, für die wir Ausreisser identifizieren wollen. Wir bezeichnen y auch kurz als "Zielvariable".

2. Es wird ein Allgemeines Lineares Modell (ALM) vom Typ der Regressionsanalyse gerechnet. Dabei sind eventuelle Ausreisser-Werte noch in den Daten enthalten. Die Regressionskoeffizienten, die Prog20bm als Ergebnis aus dem ALM ausgibt, sind also durch die Ausreisser beeinflusst.

3. Als Ergebnis der ALM entsteht ein Regressionskoeffizienten der Variablen x hinsichtlich der Variablen y. Mit diesem werden für die vorhandenen Daten für y "Prognosewerte" errechnet. Aus der Differenz von tatsächlichem Wert von y und Prognosewert für y ergeben sich die "Residuen" für y.

4. Die Orginaldaten werden zusammen mit den Residuen in eine interne Zwischendatei gespeichert. Die Residuen werden als letzte Variable an die vorhandenen Variablen angehängt.

5. Die Daten werden aus der Zwischendatei wieder gelesen. Die Residuen-Variable wird nun wie eine normale Variable behandelt und auf Ausreisser vom Typ 1 untersucht. Liegt in einem Datensatz, das Residuum von y weit ausserhalb des Wertebereichs der übrigen Residuen, dann wird dieses Residuum als Ausreisser betrachtet. Die Vorgehensweise ist dabei folgende: Betrachten wir nochmals das oben abgebildete Streudiagramm und dabei den Ausreisser-Punkt A. Der x-Wert von A ist =6, der y-Wert =17. Der Prognosewert von A ist der y-Wert auf der Regressionsgeraden über dem x-Wert 6. Das Residuum des Punktes A ist die vertikale Distanz des Punktes A von der Regressionsgeraden. Wie man sieht ist diese sehr groß. Sie beträgt rund 11 y-Einheiten. Werden für alle 28 Datensätze die Residuen ermittelt, so kann man deren Standardabweichung berechnen. Sie beträgt 3,4. Der Punkt A ist also etwas mehr als 3 Standardabweichungen von seinem Prognosewert entfernt. Wenn wir festlegen, dass Punkte, die mehr als 3 Standardabweichungen von ihrem Prognosewert entfernt sind, Ausreisser sind, dann ist der Punkt A als Ausreisser identifiziert.

6. Ist in der beschriebenen Weise ein Residuum als Ausreisser identifiziert, dann kann man den betreffenden Datensatz ausschliessen oder die Variable y auf Kein-Wert setzen. Man beachte: Nicht das Residuum von y sondern die Variable y selbst wird auf Kein-Wert gesetzt.
Prog20bm erzeugt nun eine neue Datei, in der - je nach Reaktion des Benutzers - Datensätze mit Ausreissern nicht enthalten sind oder die y-Variable auf Kein-Wert gesetzt ist. Diese Datei wird ausgegeben. Sie umfasst die ursprünglichen Variablen mit den "Ausreisser-Bereinigungen" - nicht mehr jedoch die Residuen.

Kurz zusammengefasst: Prog20bm ermittelt die Residuen der Zielvariablen y und untersucht die Residuen auf Ausreisser vom Typ 1. Die "Ausreisser-bereinigten" Daten werden dann in eine neue Datei ausgegeben.

Eine grafische Methode

In Prog02mb und Prog02mc wird eine 2- bzw. 3-dimensionale Regressionsanalyse gerechnet. Der Benutzer findet diese Programme durch Klick auf den Knopf "Verfahren", dann

"Regressionsanalyse". Das Ergebnis dieser Programme ist ein 2- bzw. 3-dimensionales Streudiagramm. Datenpunkte, die deutlich "ausreissen" können dabei vom Benutzer optisch identifiziert werden. Rechnen Sie beispielsweise das Programm "Ausreiss.Alm". Sie finden dieses Programm durch Klick auf das Menü "Almo" und den Eintrag "Liste aller Almo-Programme".

Mehrere unabhängige und abhängige Variable bei Ausreisser vom Typ 2

Sie können auch gleichzeitig mehrere Zielvariable, d.h. abhängige quantitative Variable auf Ausreisser vom Typ 2 untersuchen. In der Box

```
Analyse-Variable: Zielvariable
     (Abhängige Variable)
```

geben Sie einfach mehrere quantitative Variable an.

Als unabhängige Variable für eine (oder mehrere) Zielvariable können mehrere quantitative und / oder mehrere nominale Variable (einschliesslich Interaktionen) eingesetzt werden. Dies gilt auch für den Fall, dass die Zielvariable nominal ist. Siehe dazu nachfolgenden Abschnitt.

Ausreisser vom Typ 2 bei nominalen Zielvariablen

Das Allgemeine Lineare Modell ist auch anwendbar, wenn die abhängige Variable nominal ist. Zur Problematik dieser Art von Analyse siehe Handbuch P45 "Data Mining", Abschnitt P45.15.1.0. Betrachten wir ein Beispiel: Die nominale Zielvariable ist die Wahl einer Studienrichtung durch Studienanfänger. Die Variable besitze 3 Ausprägungen:

1. Naturwissenschaft
2. Rechtswissenschaft
3. Geisteswissenschaft

Im ALM werden 3 Dummy-Variable gebildet, die diesen 3 Ausprägungen entsprechen. Für jede Dummy-Variable als Zielvariable wird eine Analyse gerechnet. Das geschieht in einem Rechengang. Als Prognosewerte werden dann für jeden Studienbeginner 3 Wahrscheinlichkeiten ermittelt: die Wahrscheinlichkeit, dass er die Naturwissenschaft wählt, dass er die Rechtswissenschaft wählt und dass er die Geisteswissenschaft wählt. Da sich der Studienbeginner schon entschieden hat, sind die tatsächlichen Werte bekannt. Beispielweise hat er sich für die Naturwissenschaft entschieden, dann ist die tatsächliche Wahrscheinlichkeit für diese 1.0 und für die beiden anderen 0.0. Das Resduum ist dann die Differenz zwischen Prognosewert und tatsächlichem Wert. Und diese Residuen können dann auf Ausreisser vom Typ 2 untersucht werden. Betrachten Sie das Beispielprogramm "Ausreis2.Alm". Sie finden es im Menü "Almo / Liste aller Almo-Programme".

Wann kann Prog20bm zur Ausreisser-Identifizierung angewendet werden ?

Prog20bm liefert neue "Ausreisser-bereinigte" Datensätze. Wann sollen diese zur Analyse verwendet werden ? Sie können verwendet werden für 1. bivariate Korrelationen, z.B. mit den Programmen zur Korrelation 2. multiple Korrelationen, z.B. mit den Programmen zum ALM 3. einfache und multiple ALM, mit den Programmen zum ALM

Probleme mit Prog20bm

Werden mit Prog20bm 2 Analysen gerechnet, wobei für die Zielvariable y einmal x1 und einmal x2 als unabhängige Variable eingesetzt wird, dann kann es geschehen, dass y nicht in

denselben Datensätzen als Ausreisser identifiziert wird. In der Analyse mit x1 kann y beispielsweise im 10. Datensatz als Ausreisser erkannt werden, nicht jedoch in der Analyse mit x2. Das ist normal und kein Problem. Dies ist jedoch ein Hinweis darauf, dass die von Prog20bm gelieferte "Ausreisser-bereingte" Datei nicht generell, sondern nur für Analysen benutzt werden darf, in denen sich x1 und y gegenüberstehen.

Ein spezielles Problem tritt auf, wenn eine m*m-Korrelationsmatrix mit m groesser 2 mit "Ausreisser-bereinigten" Daten gerechnet werden soll. Es müsste für jede bivariate Korrelation mit Prog20bm eine eigene Datei gebildet werden, aus der die "bereinigte" bivariate Korrelation errechnet wird. Das ist im Prinzip machbar, aber sehr mühsam. Es bietet sich an, Prog20bm für diesen Fall umzuprogrammieren. Gegenwärtig liegt aber ein solches Programm noch nicht vor. Das beschriebene Problem gilt natürlich auch für Verfahren, die Korrelationsmatrizen als input benötigen - wie etwa die Faktorenanalyse

Prog20bm iterieren

Der Gedanke liegt nahe, die von Prog20bm gelieferte "Ausreisser-bereinigte" Datei nochmals mit Prog20bm auf Ausreisser durchzusuchen. Das kann mehrfach gemacht werden, so oft bis keine Ausreisser mehr auftreten oder so oft bis 2 aufeinderfolgende Korrelationen von x und y sich nur noch um einen geringen Schwellenwert unterscheiden. Diese Vorgehensweise kann natürlich zu einer massiven Manipulation empirischer Daten entarten.

Literatur

Grubbs, Frank E.: Sample criteria for testing outlying observations,
The Annals of Mathematical Statistics 21(1), 1950, S.27-58

NIST-Agency: Engineering statistics (e-Handbook of Statistical Methods)
Das Handbuch im htm-Format ist im Internet zu finden unter
http://www.itl.nist.gov/div898/handbook/index.htm
Das Kapitel über Ausreisser unter
http://www.itl.nist.gov/div898/handbook/eda/section3/eda35h.htm