Bootstrap bei Logit- und Probitanalyse

P22.3

Kurt Holm

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

2021

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

Weitere Almo-Dokumente

Die folgenden Dokumente können alle von der Handbuchseite in http://www.almo-statistik.de heruntergeladen werden

0. Arbeiten_mit_Almo.PDF (1 MB)
1a. Eindimensionale Tabellierung.PDF (1,8 MB)
1b. Zwei- und drei-dimensionale Tabellierung.PDF (1.1 MB)
2. Beliebig-dimensionale Tabellierung.PDF (1.7 MB)
3. Nicht-parametrische Verfahren.PDF (0.9 MB)
4. Kanonische Analysen.PDF (1.8 MB)
   Diskriminanzanalyse.PDF (1.8 MB)
   enthält: Kanonische Korrelation, Diskriminanzanalyse, bivariate Korrespondenzanalyse, optimale Skalierung
5. Korrelation.PDF (1.4 MB)
6. Allgemeine multiple Korrespondenzanalyse.PDF (1.5 MB)
7. Allgemeines ordinales Rasch-Modell.PDF (0.6 MB)
7a. Wie man mit Almo ein Rasch-Modell rechnet.PDF (0.2 MB)
8. Tests auf Mittelwertsdifferenz, t-Test.PDF (1,6 MB)
9. Logitanalyse.pdf (1,2MB) enthält Logit- und Probitanalyse
9a. Bootstrap bei Logit- und Probitanalyse
10. Koeffizienten der Logitanalyse.PDF (0,06 MB)
11. Daten-Fusion.PDF (1,1 MB)
12. Daten-Imputation.PDF (1,3 MB)
13. ALM Allgemeines Lineares Modell.PDF (2.3 MB)
13a. ALM Allgemeines Lineares Modell II.PDF (2.7 MB)
13b. Bootstrap bei Allgemeinem Linearen Modell III.PDF
14. Ereignisanalyse: Sterbetafel-Methode, Kaplan-Meier-Schätzer, Cox-Regression.PDF (1,5 MB)
15. Faktorenanalyse.PDF (1,6 MB)
15a. Bootstrap bei Faktorenanalyse.PDF (1,7 MB)
16. Konfirmatorische Faktorenanalyse.PDF (0,3 MB)
17. Clusteranalyse.PDF (3 MB)
18. Pisa 2012 Almo-Daten und Analyse-Programme.PDF (17 KB)
19. Guttman- und Mokken-Skalierung.PFD (0.8 MB)
20. Latent Structure Analysis.PDF (1 MB)
21. Statistische Algorithmen in C (80 KB)
22. Conjoint-Analyse (PDF 0,8 MB)
23. Ausreisser entdecken (PDF 170 KB)
24. Statistische Datenanalyse Teil I, Data Mining I
25. Statistische Datenanalyse Teil II, Data Mining II
26. Statistische Datenanalyse Teil III, Arbeiten mit Almo-Datenanalyse-System
27. Mehrfachantworten. Tabellierung von Fragen mit Mehrfachantworten
28. Metrische multidimensionale Skalierung (MDS) (0,4 MB)
29. Metrisches multidimensionales Unfolding (MDU) (0,6 MB)
30. Nicht-metrische multidimensionale Skalierung (MDS) (0,4 MB)
31. Pfadanalyse.PDF (0,7 MB)
32. Datei-Operationen mit Almo (1,1 MB)
33. Wählerstromanalyse und Wahlhochrechnung (1,6 MB)
34. Soziometrie. Auswertung soziometrischer Daten (0,5 MB)

INHALTSVERZEICHNIS

P22.3 Bootstrap bei Logit- und Probitanalyse

Die Logit- und Probitanalyse wird im Almo-Dokument 9 „Logitanalyse“ ausführlich dargestellt. Auf das Bootstrap-Verfahren wird dabei nicht eingegangen. Das soll hier nachgeholt werden.

Das Bootstrap-Verfahrens wurde bereits detailliert im Almo-Dokument 13b „Bootstrap beim *Allgemeinen Linearen Modell*“ erläutert. Wir werden deswegen hier nur sehr kurz die Vorgehensweise und den Kalkül beim Bootstrapping darstellen und dabei immer wieder auf das Almo-Dokument 13b verweisen, jedoch ausführlich die Eingabe-Masken für das Bootstrap-Programm „Prog22m5.Msk“ für das Logit- und Probit-Modell erläutern.

P22.3.1 Vorgehensweise beim Bootstrap

Aus einer vorliegenden Stichprobe (wir nennen sie "originale" Stichprobe) der Größe n werden zufällig n (also gleich viele) Datensätze mit *Zurücklegen* ausgewählt. Dadurch entsteht die Bootstrap-Stichprobe Nr. 1. Das Zurücklegen bewirkt, dass manche Datensätze mehrfach ausgewählt werden und dass manche Datensätze der originalen Stichprobe nicht in die Bootstrap-Stichprobe geraten.

Auf diese Weise werden viele, etwa 1000 Bootstrap-Stichproben erzeugt. Für alle Stichproben werden die Ergebnisse errechnet. In Almo werden zuerst die Ergebnisse für die Original-Stichprobe ausgegeben, danach die aus allen Bootstrapstichproben zusammengefassten Ergebnisse. Das wird noch detailliert gezeigt. Besonders bedeutsam ist, dass aus dem Bootstrapping *empirische* Verteilungen für die verschiedenen Koeffizienten gewonnen werden. Dadurch ist es möglich, *Standardfehler, Signifikanzen* und *Konfidenzintervalle* für diese Koeffizienten zu ermitteln, die keine Verteilungsannahmen erfordern. Das ist der primäre Zweck des Bootstrap-Verfahrens. Es erzeugt "robuste", "verteilungsfreie" Schätzer und löst somit manches statistische Problem.

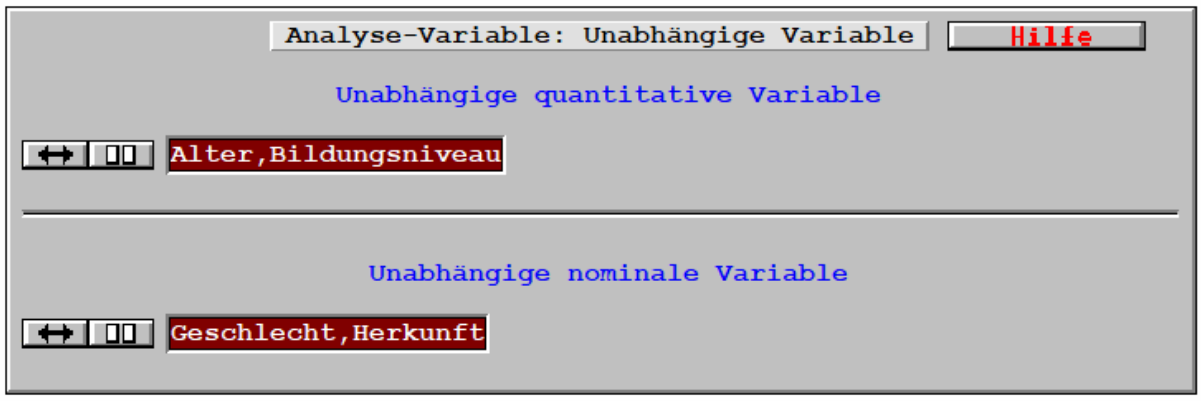
Betrachten wir als Beispiel den Regressionskoeffizienten ß1 für eine unabhängige quantitative Variable x1. Aus den 1000 Bootstrap-Stichproben erhalten wir 1000 Werte für ß1. Wir berechnen deren Mittelwert und ihre Standardabweichung. Die Standardabweichung ist dann der "Standardfehler" von ß1. Die obere und untere Grenze des Konfidenzintervalls für beispielsweise ein Konfidenzniveau von 95% erhalten wir sehr einfach in folgender Weise: Die 1000 ß1-Werte werden der Größe nach (aufsteigend) sortiert. Vom maximalen ß1-Wert werden absteigend 2,5% von 1000 also 25 Werte heruntergezählt. Der dort in Position 975 stehende ß1-Wert ist die obere Intervallgrenze. Entsprechend wird vom minimalen Wert ausgehend 25 Werte hinaufgezählt. So wird der untere Grenzwert gefunden. Zwischen den beiden Grenzwerten befinden sich dann 95% aller Werte und außerhalb der Grenzwerte 5% aller Werte. Fällt die Grenze zwischen zwei ß1-Werte, dann wird interpoliert. Diese sehr einfache Berechnungsweise wird als "Perzentil-Verfahren" bezeichnet. Als alternatives Verfahren wird das PCa-Verfahren empfohlen, dem jedoch vorgeworfen wird, ein zu enges Intervall zu schätzen. Es gibt noch weitere Verfahren. Einen knappen Überblick findet man im englischen Wikipedia. Almo verwendet das beschriebene einfache Perzentil-Verfahren und optional das asymmetrische und symmetrische Perzentil-t - Verfahren.

Wird der Mittelwert aus den 1000 ß1-Werten mit dem ß1-Wert aus der Original-Stichprobe verglichen, so muss man in aller Regel eine kleine "Verzerrung" zur Kenntnis nehmen. Der Mittelwert hat sonst keine Bedeutung. Als bester Schätzer für ß1 wird der ß1-Wert aus der Original-Stichprobe für den Forschungsbericht verwendet. Als seinen Standardfehler wird die aus dem Bootstrap gewonnene Standardabweichung eingesetzt, als seine Signifikanz p und sein Konfidenzintervall wird der aus dem Perzentil-Verfahren errechneten Werte eingesetzt. Alle diese Koeffizienten sind "parameterfrei".

Almo unterzieht nicht nur die Regressionskoeffizienten aus dem Logit- oder Probit-Modell diesem Bootstrap-Kalkül, sondern auch (alternativ) die Risikokoeffizenten exp(ß) und die „paarweisen Kontraste“.

P22.3.2 Unabhängige Variable

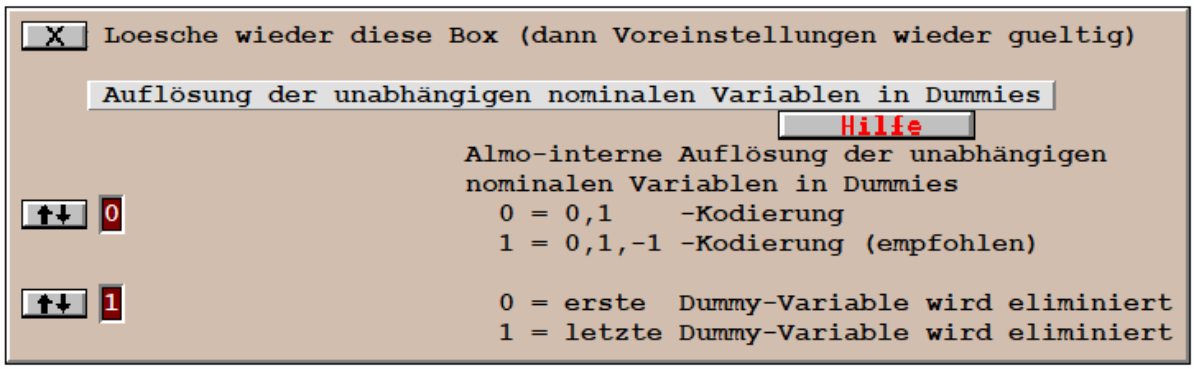
Die Programm-Maske für das Prog22m5.Msk findet man durch Klick auf den Knopf *Verfahren* am Oberrand des Almo-Hauptfensters und dann durch Klick auf die Einträge „Bootstrap“ oder „Logit-,Probit-Analyse“. Prog22m5 ist weitgehend identisch mit dem „reinen“ Logit-, Probitmodell Prog22m (siehe Almo-Dokument 9 „Logitanalyse, Abschnitt P22.2.1) . Wir werden vorrangig jene Eingabeboxen kommentieren, die neu hinzugekommen sind, bzw. einer zusätzlichen Erläuterung für den Bootstrap-Fall bedürfen.



Möglich sind quantitative und nominale Variable. Die nominalen Variablen müssen ganzzahlig mit Schrittweite 1 kodiert sein (z.B. 1,2,3). Sind sie das nicht, dann müssen sie in der Optionsbox „Umkodierungen und Kein-Wert-Angaben“ entsprechend umkodiert werden. Etwa so: `Varname(1.75=1; 4.125=2; 2.03=3)`
Siehe dazu Almo-Dokument „P0 Arbeiten mit Progs“, Abschnitt P0.5.4.2 und P0.5.5

P22.3.2.1 Auflösung der unabhängige nominalen Variablen in Dummies

Die unabhängigen *nominalen* Variablen werden programmintern in Dummies aufgelöst. Dabei kann der Benutzer vorgeben, wie das geschehen soll.



Die Einstellungen, die der Benutzer in dieser Optionsbox vornimmt, gelten sowohl für die originalen Stichproben-Daten als auch für die der Bootstrap-Stichproben.

Der Benutzer kann dabei wählen,

1. ob er die 0,1 -Kodierung oder die 0,1,-1 -Kodierung verwenden möchte (siehe dazu Almo-Dokument 13a, Abschnitt P20.3 und P22)

2. ob er die erste oder letzte Dummy eliminieren möchte. Dies ist nur bei der 0,1 -Kodierung möglich. Bei der 0,1,-1 -Kodierung ist dies gleichgültig.

3. Ist die Optionsbox geschlossen, dann ist die 0,1,-1 -Kodierung voreingestellt.

Bei der 0,1 -Kodierung wird der Regressionskoeffizient der eliminierten Dummy auf .0 gesetzt. Bei der 0,1,-1 -Kodierung erhält jede Dummy einen Regressionskoeffizienten zugeordnet. Die Koeffizienten summieren sich dann zu .0. Wir empfehlen die 0,1,-1 -Kodierung, da hier für alle Dummies Koeffizienten (Beta, Standardfehler etc.) berechnet werden können.
Die 0,1 -Kodierung wird in der Literatur auch "Indikator-Kodierung" (bei SPSS: indicator-contrast) und die 0,1,-1 -Kodierung "Effekt-Kodierung" (bei SPSS: deviation-contrast) genannt.
SPSS verwendet in seiner „multinomialen Logit-Regression“ die 0,1 –Kodierung und eliminiert die letzte Dummy. Sollen die Ergebnisse aus Almo mit denen von SPSS verglichen werden, dann sollte der Benutzer die Optionen in der Box entsprechend einstellen: Im ersten Eingabefeld „0“ und im zweiten „1“.

Dummy-Kodierung	Interpretation der Effekte
0,1-Kodierung Die letzte Ausprägung wird gestrichen. Es werden Effekte für die erste, zweite, usw. bis zur vorletzten Ausprägung berechnet.	Die letzte Ausprägung ist die Referenzgruppe. Die Effekte geben an, ob die Wirkung der anderen Ausprägungen größer/kleiner/ als jene der Referenzgruppe (der letzten Ausprägung) sind.
0/1-Kodierung Die erste Ausprägung wird gestrichen. Es werden Effekte für die zweite, dritte, vierte, usw. Ausprägung berechnet.	Die erste Ausprägung ist die Referenzgruppe. Die Effekte geben an, ob die Wirkung der anderen Ausprägungen größer/kleiner als jene der Referenzgruppe (erste Ausprägung) sind.

0,1.-1 -Kodierung. Es werden die Effekte von allen Ausprägungen berechnet.	Die Effekte der Ausprägungen geben an, ob die Wirkung einer Ausprägung größer/kleiner der durchschnittlichen Wirkung von allen Ausprägungen ist. Die durchschnittliche Wirkung ist .0

Unterschiedliche Ergebnisse.
Die Regressionskoeffizienten der Dummies und nicht nur diese, die aus diesen drei Kodierungsarten hervorgehen, sind verschieden.
Betrachten wir ein Beispiel. In unserem Bootstrap-Programm Prog22m5 ist die abhängige nominale Variable die Wohnlage mit den 3 Ausprägungen (1) Land, (2) Stadtrand, (3) Stadt. Die unabhängigen nominalen Variablen sind das Geschlecht und die soziale Herkunft (mit 2 bzw. 3. Ausprägungen). Wir erhalten folgende drei verschiedene Ergebnisse aus den drei Kodierungsarten. Ausgabe gekürzt.

```
 unabhaengige Variab  Regress.    "Risiko"   Stand.     Wald  Signif. partielle
                      koeffiz.     exp(ß)    Fehler      z*z     p    Korrelat.
                         ß
------------------------------------------------------------------- ---------
                                 0,1 -Kodierung
                                 letzte Dummy-Variable wird eliminiert
                                 ------------------------------------
Konstante              14.23315       -      2.36922   36.090  0.0000      -
A1 Geschlec:   männl   -2.12502    0.11943   0.52408   16.441  0.0001  -0.12258
B1 Herkunft:Untersch    2.04705    7.74503   0.56322   13.210  0.0003   0.10800
B2 Herkunft:Mittelsc    0.78598    2.19455   0.37097    4.489  0.0344   0.05089
  V11 Alter            -0.40542    0.66669   0.06043   45.018  0.0000  -0.21156
   V1 Bildungsniveau    0.33416    1.39676   0.07438   20.184  0.0000   0.13755
=============================================================================
                                 0,1 -Kodierung
                                 erste Dummy-Variable wird eliminiert
                                 ------------------------------------
Konstante              14.15518       -      1.86006   57.913  0.0000      -
A2 Geschlec:   weibl    2.12502    8.37307   0.52408   16.441  0.0001   0.12258
B2 Herkunft:Mittelsc   -1.26108    0.28335   0.49898    6.387  0.0116  -0.06756
B3 Herkunft:Oberschi   -2.04705    0.12912   0.56322   13.210  0.0003  -0.10800
  V11 Alter            -0.40542    0.66669   0.06043   45.018  0.0000  -0.21156
   V1 Bildungsniveau    0.33416    1.39676   0.07438   20.184  0.0000   0.13755
=============================================================================
                                 0,1,-1 -Kodierung
                                 -----------------
Konstante              14.11498       -      2.09682   45.315  0.0000      -
A1 Geschlec:   männl   -1.06251    0.34559   0.26204   16.441  0.0001  -0.12258
A2 Geschlec:   weibl    1.06251    2.89363   0.26204   16.441  0.0001   0.12258
B1 Herkunft:Untersch    1.10271    3.01232   0.33246   11.001  0.0009   0.09677
B2 Herkunft:Mittelsc   -0.15837    0.85354   0.22509    0.495  0.4819  -0.03957
B3 Herkunft:Oberschi   -0.94434    0.38894   0.27095   12.148  0.0005  -0.10275
  V11 Alter            -0.40542    0.66669   0.06043   45.018  0.0000  -0.21156
   V1 Bildungsniveau    0.33416    1.39676   0.07438   20.184  0.0000   0.13755
=============================================================================
```

Man vergleiche die Koeffizienten für die "Mittelschicht" aus diesen drei Ergebnissen. So sind etwa die p-Werte verschieden `0.0344  0.0116  0.4819`

Hingegen sind die paarweisen Vergleich aus den 3 Analysen voll identisch

```
Paarweise Vergleiche (Kontraste) der unabhaengigen nominalen Variablen
fuer 1.Auspraegung "Land" der abhaeng. Var. V4 Wohnlage

 Vergleichs-  Differenz    "Risiko"    Stand.-   z-Wert  Signifikanz
 paar                   exp(Differenz) Fehler            p     (1-p)*100
=========================================================================
                                 0,1 -Kodierung
                                 letzte Dummy-Variable wird eliminiert
```

A1 - A2	-2.1250	0.1194	0.5241	4.055	0.000	100.00
B1 - B2	1.2611	3.5292	0.4990	2.527	0.012	98.84
B1 - B3	2.0471	7.7450	0.5632	3.635	0.000	99.97
B2 - B3	0.7860	2.1945	0.3710	2.119	0.034	96.58

0,1 -Kodierung
erste Dummy-Variable wird eliminiert

A1 - A2	-2.1250	0.1194	0.5241	4.055	0.000	100.00
B1 - B2	1.2611	3.5292	0.4990	2.527	0.012	98.84
B1 - B3	2.0471	7.7450	0.5632	3.635	0.000	99.97
B2 - B3	0.7860	2.1945	0.3710	2.119	0.034	96.58

0,1,-1 -Kodierung

A1 - A2	-2.1250	0.1194	0.5241	4.055	0.000	100.00
B1 - B2	1.2611	3.5292	0.4990	2.527	0.012	98.84
B1 - B3	2.0471	7.7450	0.5632	3.635	0.000	99.97
B2 - B3	0.7860	2.1945	0.3710	2.119	0.034	96.58

Auch und besonders bedeutsam ist, dass alle drei Kodierungsarten exakt dieselben Prognosewerte (vom Modell reproduzierte Wahrscheinlichkeit) je Proband erzeugen. Auch alle Kennwerte für das Gesamtmodell, wie etwa der Log-Maximum-Likelihood-Wert, sind für die drei Analysen identisch.

Umrechnung der Regressionskoeffizienten.
Die aus der 0,1,-1 -Kodierung entstandenen Regressionskoeffizienten können leicht umgerechnet werden auf die beiden durch 0,1-Kodierung erzeugten Regressionskoeffizienten

Ergebnisse für die 1. Ausprägung "Land"
der abhängigen Variablen V4 Wohnlag

0,1 –Kodierung
letzte Dummy eliminiert

unabhaengige Variab	Regress. koeffiz. ß
Konstante	14.23315
A1 Geschlecht: männl	-2.12502
B1 Herkunft: Untersch	2.04705
B2 Herkunft: Mittelsc	0.78598
V11 Alter	-0.40542
V1 Bildungsniveau	0.33416

0,1 –Kodierung
erste Dummy eliminiert

unabhaengige Variab	Regress. koeffiz. ß
Konstante	14.15518
A2 Geschlecht: weibl	2.12502
B2 Herkunft: Mittelsc	-1.26108
B3 Herkunft: Oberschi	-2.04705
V11 Alter	-0.40542
V1 Bildungsniveau	0.33416

0,1,-1 –Kodierung
alle Dummies einbezogen

unabhaengige Variab	Regress. koeffiz. ß
Konstante	14.11498
A1 Geschlecht: männl	-1.06251

```
A2 Geschlecht: weibl   1.06251

B1 Herkunft: Untersch  1.10271
B2 Herkunft: Mittelsc -0.15837
B3 Herkunft: Oberschi -0.94434

V11 Alter             -0.40542
V1  Bildungsniveau     0.33416
-----------------------------
```

Die Regressionskoeffizienten, die aus den drei Versionen hervorgehen, sind verschieden. Die Umrechnung von der 0,1,-1 -Kodierung auf die 0,1 -Kodierung mit eliminierter letzter Ausprägung ist sehr einfach.

Notation:

```
a_letzt = letzter Dummy-Wert aus der 0,1,-1 -Kodierung
a_erst  = erster Dummy-Wert aus der 0,1,-1 -Kodierung
a1,a2,...ai,...a_letzt-1 = die Dummies aus der 0,1,-1 -Kodierung
b1,b2,...bi,...b_letzt-1 = die anderen Dummies aus der 0,1-Kodierung
                           mit letzter eliminierter Dummy
c2,c3,...ci,...c_letzt   = die anderen Dummies aus der 0,1-Kodierung
                           mit erster eliminierter Dummy
```

Die 0,1-kodierten Dummies `bi` bei letzter eliminierter Dummy bzw `ci` bei erster elimierten Variablen entstehen dann aus

0,1-kodiert, *letzte* Dummy eliminiert: `bi = ai - a_letzt`
0,1-kodiert, *erste* Dummy eliminiert: `ci = ai - a_erst`

Wird der *letzte* Dummy-Wert a_letzt subtrahiert von ai, dann entsteht der Dummy-Wert bi
Beipiel: Es soll der Wert der Dummy a1 "Unterschicht" umgerechnet werden von der 0,1,-1-Kodierung auf die entsprechende Dummy b1 aus der 0,1-Kodierung mit letzter eliminierter Dummy. Wie entsteht `b1=2.04705` aus `a1=1.10271`

```
   a1   -  a_letzt  =   b1
1.10271 - -0.94434  = 2.04705
```

Wird der *erste* Dummy-Wert `a_erst` subtrahiert von `ai`, dann entsteht der Dummy-Wert `ci`
Beispiel: Wie entsteht Mittelschicht `c2=-1.26108` in der 0,1-Kodierung mit erster eliminierter Dummy aus `a2=-0.15837` der 0,1,-1-Kodierung

```
   a2    -  a_erst  =    c2
-0.15837 - 1.10271  = -1.26108
```

Interpretation.
Wie müssen die Effekte `ai` bzw. `bi` bzw. `ci` interpretiert werden ?
Soll beispielsweise der Effekt, den die „Mittelschicht“ auf die „Wohnlage: Land“ ausübt, interpretiert werden, dann muss bei der ersten Version formuliert werden: Im Vergleich zur Oberschicht ist der Effekt der Mittelschicht auf die „Wohnlage: Land“ stärker (positives Vorzeichen), aber weniger stark als bei der Unterschicht.

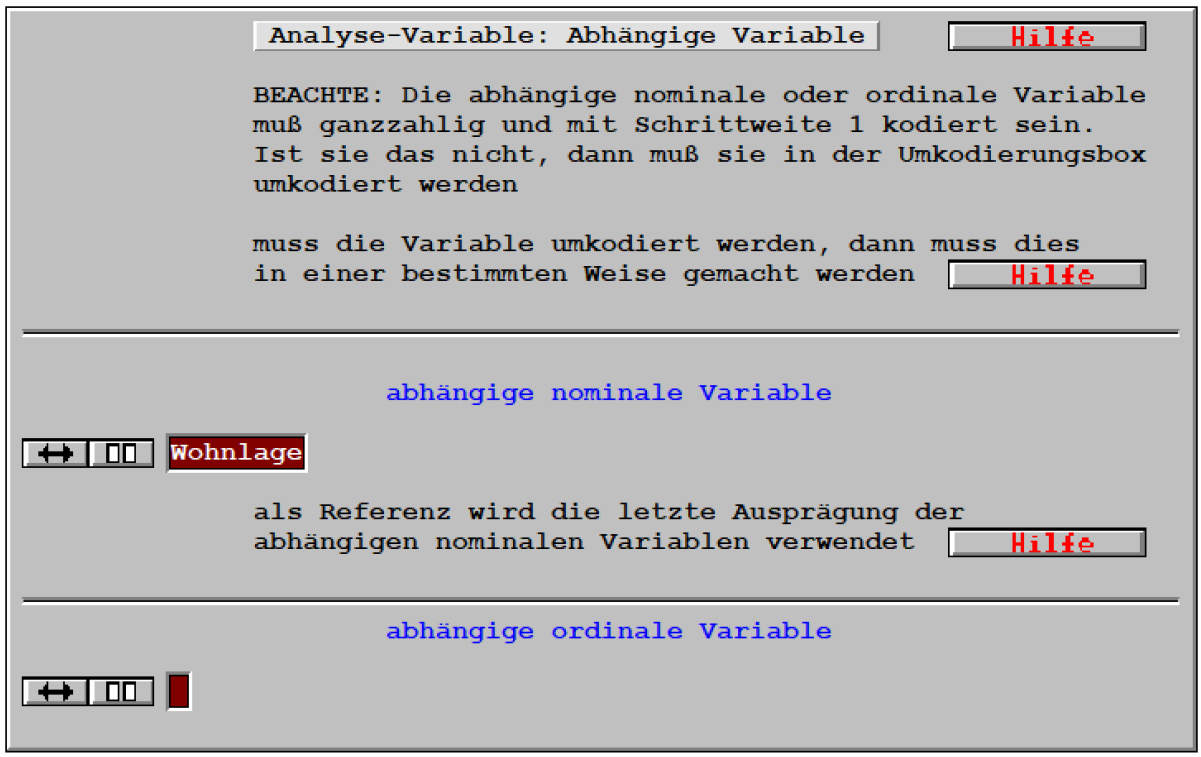
Bei der 2. Version wird die "Unterschicht" zur Referenzgruppe. Im Vergleich zu ihr, ist nun der Effekt der "Mittelschicht" kleiner (negatives Vorzeichen) aber größer als der der Oberschicht.

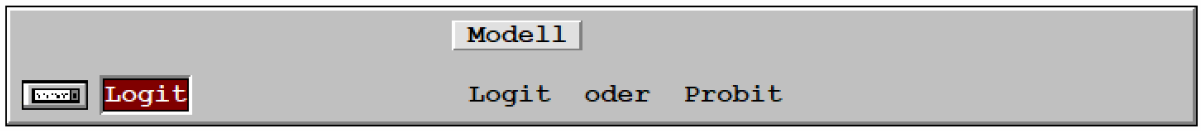
Bei der 3. Version werden die einzelnen sozialen Schichten mit der durchschnittlichen

Wirkung der drei sozialen Schichten auf die „Wohnlage: Land“ verglichen. Der Effekt der Mittelschicht und auch der Oberschicht ist geringer als der durchschnittliche Effekt. Dagegen ist der Effekt der Unterschicht stärker.

Eine sehr ausführliche Darstellung, wie die Logitkoeffizienten zu interpretieren sind, wird im Almo-Dokument 10 „Koeffizienten der Logitanalyse“ vorgetragen. Es erweist sich dabei, dass die "Risikokoeffizienten" exp(ß) möglicherweise inhaltlich besser interpretierbar sind als die Regressionskoeffizienten ß.

P22.3.3 Abhängige Variable





In der zweiten Eingabebox "Modell" kann zwischen dem "Logitmodell" und dem "Probitmodell" gewählt werden. Abhängig vom Messniveau und der Zahl der Ausprägungen der abhängigen Variablen können somit folgende Untermodelle gerechnet werden:

abhängige Variable	Logitanalyse	Probitanalyse
dichotom	binäres Logitmodell	binäres Probitmodell
polytom-nominal	multinomiales Logit-modell	nicht existent
polytom-ordinal	ordinales Logitmodell	ordinales Probitmodell

Eine "dichotome" Variable ist eine Variable mit 2 Ausprägungen, die mit zwei aufeinanderfolgende Ganzzahlen kodiert sind (z.B. 1 und 2, aber auch 33 und 34 wäre

möglich). Wird eine dichotome Variable einmal in das Eingabefeld für *nominale* abhängige Variable eingetragen und in einer zweiten Analyse in das Eingabefeld für *ordinale* Variable, dann entsteht bei der Probitanalyse ein identisches Ergebnis. Bei der Logitanalyse allerdings entstehen geringfügig andere Ergebniswerte (etwa an der 3. Dezimalstelle), wobei auch noch die Vorzeichen umgekehrt sein können. Ob dies bedeutet, dass bei der Logitanalyse der ordinale Charakter der Variablen berücksichtigt wird, ist unklar.

Mit dem Begriff "polytom-nominal" ist eine Variable gemeint, die (1) ganzzahlig kodiert ist, (2) und mit Schrittweite 1 aufsteigend kodiert ist. Die Kodierung muss nicht mit 1 beginnen.

Eine "polytom-ordinale" Variable ist dann eine polytom-nominale Variable, wobei die aufeinander folgenden Zahlen eine *ordinale Folge* ausdrücken.

Entspricht die Kodierung der vorhandenen Daten nicht den angegebenen Vorschriften, dann muss die betreffende Variable in der Umkodierungsbox der Programm-Maske umkodiert werden. Wir werden anschließend zeigen, wie das geschehen muss.

In dem oben abgebildeten Beispiel wird die Variable "Wohnlage" (mit den 3 Ausprägungen: Land, Stadtrand, Stadt) als abhängige nominale Variable eingesetzt. Es soll ein "multinomiales Logitmodell" gerechnet werden.

1. Eingabefeld: Abhängige *nominale* Variable
Hier kann eine Variable mit folgenden Charakteristika eingesetzt werden:

a. Sie muss dichotom oder polytom sein. Ihre Ausprägungen sollen nicht ordinal geordnet sein. Ordinal geordnet wären z.B. die Antworten auf die Frage "Schauen Sie Fernsehen ?" (1) gelegentlich, (2) oft, (3) immer. Nicht-ordinal geordnet wären die Antworten (1) überwiegend Sportsendungen, (2) überwiegend Unterhaltungssendungen, (3) alles.

   Dies gilt für die **Logitanalyse**. Wird für sie eine ordinale Variable in das 1. Eingabefeld eingesetzt, dann wird ihr ordinaler Charakter nicht berücksichtigt. Jede einzelne Ausprägung wird als selbständige abhängige Variable betrachtet, für die eine eigene Analyse gerechnet wird

   Im Fall der **Probitanalyse** ist nur eine dichotome Variable erlaubt.

b. Die Variable muss ganzzahlig und mit Schrittweite 1 kodiert sein. Ist sie das nicht, dann muss sie in der Optionsbox "Umkodierungen und Kein-Wert-Angaben entsprechend umkodiert werden. Wir werden anschließend zeigen, wie das geschehen muss.

2. Eingabefeld: Abhängige *ordinale* Variable
Hier darf nur eine eine ordinale abhängige Variable eingesetzt werden.

a. Sie muss, wie oben unter a) mit Hilfe eines Beispiels beschrieben *ordinal* geordnet sein.
b. und sie muss wie oben unter b) beschrieben ganzzahlig und mit Schrittweite 1 kodiert sein.
   Dies gilt für die Logitanalyse und auch für die Probitanalyse. Für letztere ist also eine dichotome oder eine ordinale Analyse möglich, aber keine multinomiale.

Selbstverständlich gelten diese Eingaben gleichermaßen für die Original-Stichprobe und alle Bootstrap-Stichproben.

P22.3.3.1 Umkodierung der abhängigen Variablen

Betrachten wir ein Beispiel:
Wir wollen im Beispiel, das in der Programm-Maske Prog22m5 gerechnet wird, die "Arbeitsbelastung" als abhängige nominale Variable einsetzen. Die Variable besitzt sehr viele verschiedene Ausprägungen, die zum Teil sogar mit Dezimalzahlen kodiert sind. Wir entschließen uns, die Variable zu dichotomisieren. Die Umkodierungsanweisung dafür lautet

```
Arbeitsbelastung(0:11=1; 11:100=2)
```

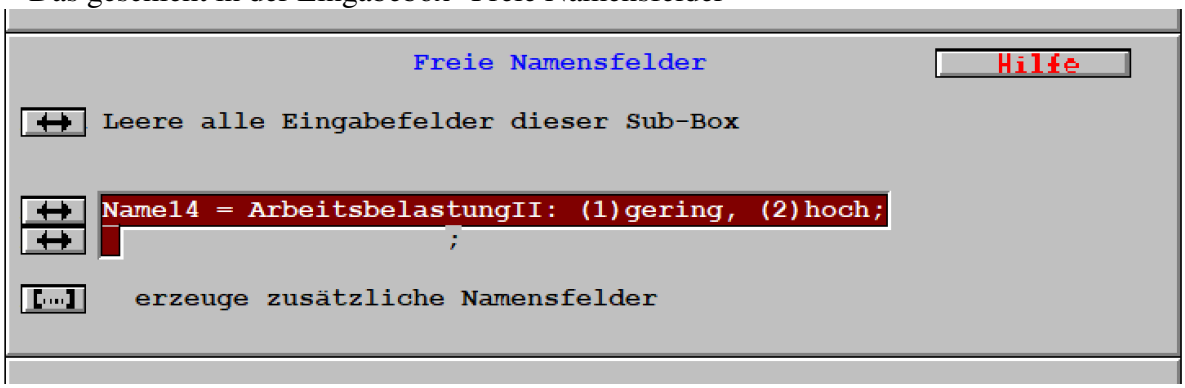
Da die Variable quantitativ ist, wäre es eigentlich günstiger die Variable auf etwa 4-6 Ausprägungen zusammen zu fassen und als ordinale abhängige Variable einzugeben.

Almo verlangt nun, dass die abhängige Variable nach der Umkodierung einer neuen, "freien" Variablen zugewiesen wird. Diese Vorschrift gilt nur für die Logit- und Probitanalyse.

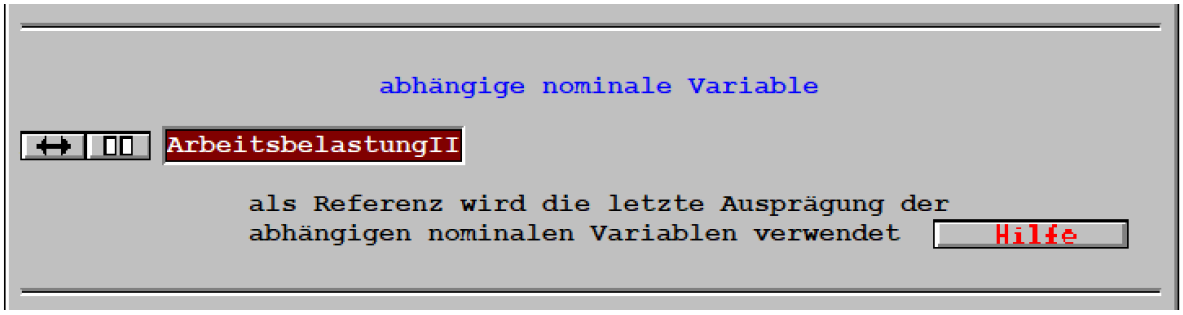
Wenn der eingelesene Datensatz wie in unserem Beispiel die Variablen V1 bis V13 enthält und in der VEREINBARE-Anweisung der Programm-Maske 100 Variable vereinbart wurden, dann sind die Variablen V14 bis V100 noch frei. Man wird also die umkodierte "Arbeitsbelastung" am einfachsten der Variablen V14 zuweisen und dieser auch einen Namen geben, etwa **`ArbeitsbelastungII`**.

Folgende Änderungen müssen in der Programm-Maske vorgenommen werden

1. Die neue, freie Variable 14 erhält den Namen
   `Name 14 = ArbeitsbelastungII: (1)wenig, (2)viel;`
   Das geschieht in der Eingabebox "Freie Namensfelder"



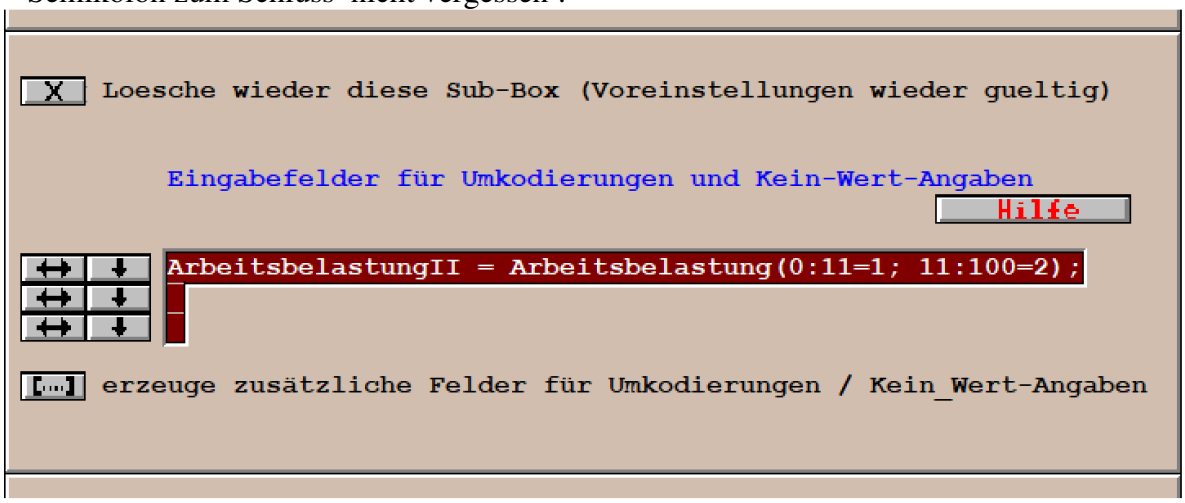
2. Als abhängige nominale Variable wird nun eingesetzt: `ArbeitsbelastungII`
   Das geschieht in der Eingabebox "Abhängige Variable" im Eingabefeld "nominale Variable"



3. Die Optionsbox "Umkodierungen und Kein-Wert-Angaben" wird geöffnet.
   In eines der Eingabefelder wird eingetragen

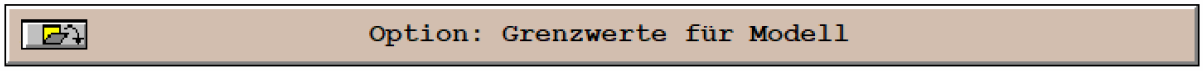
```
ArbeitsbelastungII = Arbeitsbelasung(0:11=1; 11:100=2);
```

Semikolon zum Schluss nicht vergessen !

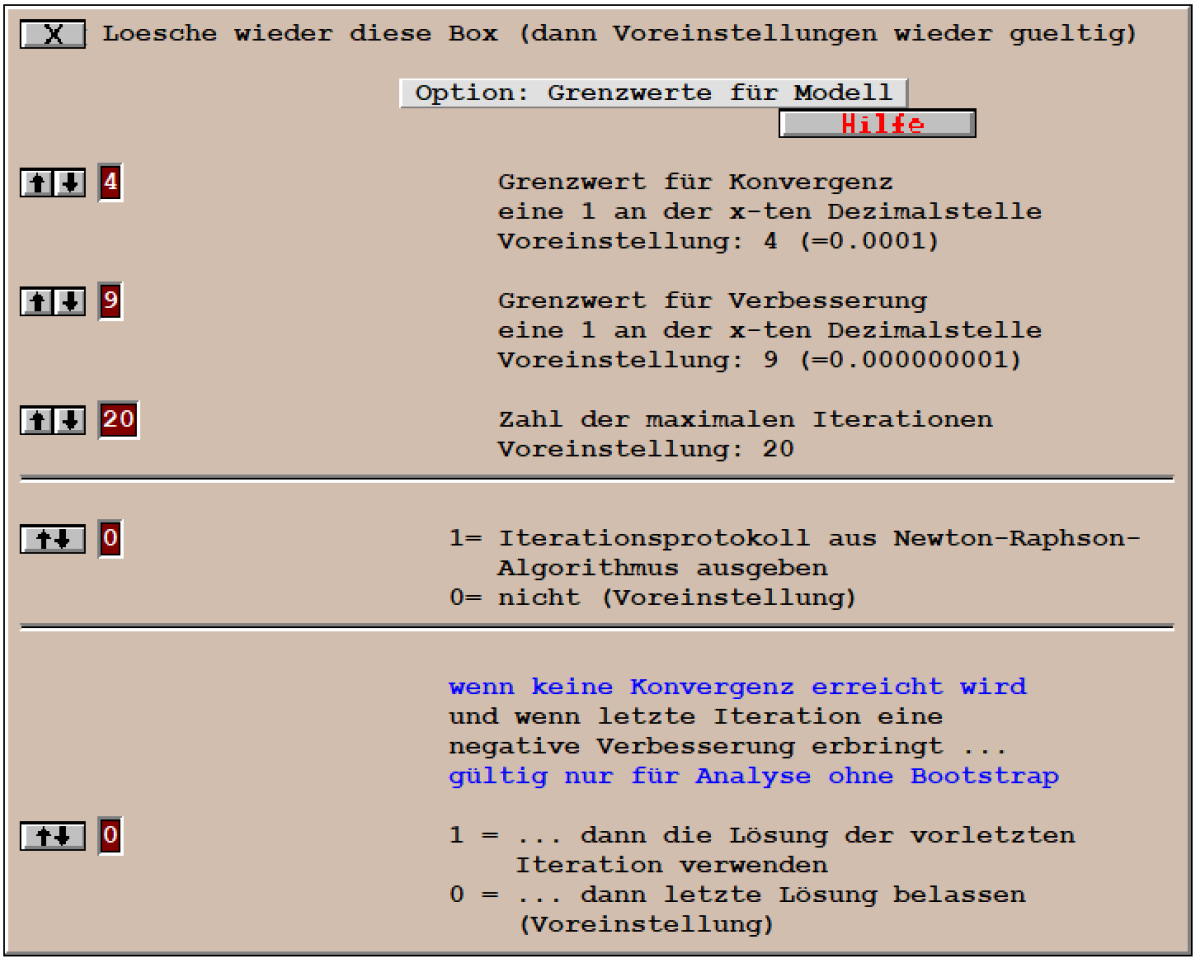


Zu den vielen Möglichkeiten eine Variable umzukodieren. Siehe dazu Handbuch, Teil 2 "Almo-Programmiersprache", Abschnitt 16 oder Almo-Dokument Nr. 0 "Arbeiten mit Progs", Abschnitt P0.5.4.2 und P0.5.5

P22.3.4 Grenzwerte für das Modell



Wird die Optionsbox geöffnet, dann sieht man folgendes



Grundvoraussetzung bei der Logit- und bei der Probitanalyse ist, dass die im Verlauf des Kalküls entstehende "Informationsmatrix" nichtsingulär ist. Sie ist dann nicht invertierbar. Eine Analyse ist nicht mehr durchführbar.

Eine mögliche Abhilfe besteht darin, einige der unabhängigen Variablen herauszunehmen oder durch andere zu ersetzen oder, bei der multinomialen Logitanalyse, die Zahl der Ausprägungen der abhängigen Variablen durch Zusammenfassen zu reduzieren.

Tritt dieser Fall bei einer Bootstrap-Stichprobe auf, dann wird diese Stichprobe ausgeschlossen - d.h. als nicht-existent betrachtet - und das Verfahren mit der nächsten Stichprobe weitergeführt. Die Zahl der Stichproben verringert sich dadurch um 1.

Tritt dieser Fall gleich bei der originalen Stichprobe auf, dann muss das Bootstrap-Verfahren insgesamt sofort abgebrochen werden - nicht jedoch bei einer Nicht-Bootstrap-Analyse

Die Ergebnisse der Logit- und Probitanalyse entstehen in Almo aus dem "Newton-Raphson-Algorithmus". Der Benutzer kann sich im Internet in vielen Einträgen über dieses Verfahren informieren. Es ist eine Iterationsverfahren, das in Almo beendet wird, wenn einer von 3 "Grenzwerten" erreicht bzw. über- oder unterschritten wird. Dies sind

1. der Grenzwert für die Konvergenz,
2. die Verbesserung
3. und die Iterationszahl

In den ersten drei Eingabefeldern der geöffneten Optionsbox kann der Benutzer die in Almo voreingestellten Grenzwerte verändern.

Bei jedem Iterationsschritt wird zuerst überprüft, ob die 1. Ableitung der Likelihoodfunktion den eingestellten Grenzwert von 0.0001 unterschreitet. Dies bezeichnen wir auch etwas vereinfachend als Grenzwert für die "Konvergenz". Ist dies der Fall, dann wird das Iterieren *erfolgreich* beendet - gleichgültig welchen Wert die Verbesserung und die Iterationszahl eingenommen haben. Wird keine Konvergenz erzielt, dann wird überprüft, ob die Verbesserung so minimal geworden ist, dass sich weiteres Iterieren nicht mehr rentiert. und schließlich wird abgebrochen, wenn die Iterationszahl die eingestellte maximale Zahl überschritten hat.

Auch wenn keine Konvergenz erzielt werden konnte, wird weitergerechnet und die endgültigen Ergebnisse der Analyse ermittelt und ausgegeben. Der Benutzer kann in dieser Situation jedoch eingreifen. Das wird noch im Detail weiter unten im Text ausgeführt.

Das Iterieren wird mit *Erfolg* beendet, wenn "Konvergenz" erreicht wird. Das Iterieren wird mit *fehlendem oder mangelhaftem Erfolg* beendet, wenn die Konvergenz (noch) nicht erreicht ist und die "Verbesserung" so minimal ist, oder sogar negativ wird oder wenn die "Iterationszahl" die vorgegebene maximale Zahl überschreitet.

Die drei Kriterien "Konvergenz", "Verbesserung" und "maximale Iterationszahl"
sollen genauer definiert werden.

1. Eingabefeld: Die "Konvergenz"
Bei jedem Iteraionsschritt wird die absolut größte 1. Ableitung der Likelihoodfunktion ermittelt. Siehe dazu im Almo-Dokument Nr. 9 "Logitanalyse", Abschnitt P22.1.5. Diese 1. Ableitung soll mit jedem nachfolgenden Iterationschritt kleiner werden und sich an einen vorgegeben Grenzwert annähern. Wenn dieser erreicht oder sogar unterschritten ist, kann die "Konvergenz" des Newton-Raphson-Kalküls als erfolgreich definiert werden und das Iterieren beendet werden. Auch dann, wenn die im folgenden Punkt beschriebene "Verbesserung" noch unbefriedigend ist oder sogar negativ ist. In Almo ist der Grenzwert mit 0.0001 (1 an der 4. Dezimalstelle) relativ großzügig eingestellt. Der Benutzer kann diesen Wert in der Optionsbox "Grenzwerte für Modell" verkleinern oder vergrößern. Zu diesem Zweck wird angegeben an welcher Dezimalstelle die "1" stehen soll. In der genannten Optionsbox kann auch ein "Iterationsprotokoll" angefordert werden, aus dem ersichtlich wird, wie sich von einem Itertationsschritt zum nächsten die hier beschriebenen Kriterien ihrem Grenzwert nähern. Siehe dazu weiter unten.

2. Eingabefeld: Die "Verbesserung"
Bei jedem nachfolgenden Iteraionsschritt sollte eine immer kleiner werdende "Verbesserung" des "Log-Maximum-Likelikood-Werts" entstehen. Dieser ist ein Indikator für die Güte des Modells. Siehe dazu das Almo-Dokument Nr.9 "Logitanalyse", Abschnitt P22.1.5 und P22.2.3.0. Mit "Verbesserung" ist also die Differenz zwischen dem Log-ML-Wert des Iterationsschritt i gegenüber dem vorhergehenden i-1 gemeint. In Almo ist ein Grenzwert von 1 an der 9. Dezimalstelle (0.000000001) eingestellt, der vom Benutzer in der Optionsbox "Grenzwerte für Modell" verändert werden kann. Wird dieser erreicht oder unterschritten oder sogar negativ, dann wird unterstellt, dass durch weiteres Iterieren keine relevante Verbesserung mehr entsteht. Der Grenzwert definiert also ein gerade noch akzeptables Verbesserungs-Minimum. Wird jedoch "Konvergenz" erzielt bevor dieser

Zustand eingetreten ist, dann wird unabhängig vom Wert der Verbesserung das Iterieren erfolgreich beendet. Die Konvergenz genießt Priorität.

Nicht selten tritt der Fall auf, dass die Verbesserung negativ ist. Ist dann der Konvergenzwert (genauer die 1. Ableitung der Likelihoodfunktion) noch weit vom angestrebten Grenzwert entfernt, dann kann dies ein Hinweis sein, dass die vorhandenen Daten nicht durch das ML-Modell der Logit- bzw. Probitanalyse analysierbar sind.
Möglich wäre es auch den "Likelihood-Ratio-Wert" zu verwenden um die Verbesserung zu bestimmen, wie dies in SPSS in der logistischen Regression getan wird. Dieser Koeffizient wird dort mit "-2 Log-Likelihood" bezeichnet. Er ist -2 mal dem Log-ML-Wert, mit dem Almo die Verbesserung berechnet. Almo gibt diesen Wert im Iterationsprotokoll aus, verwendet ihn jedoch nicht.

3. Eingabefeld: Die maximale Iterationszahl
Wenn nach 20 Iterationen keine "Konvergenz" erreicht und auch die "Verbesserung" noch nicht kleiner gleich dem Grenzwert ist, dann wird das Iterieren abgebrochen. Der Benutzer kann diesen Grenzwert in der Optionsbox "Grenzwerte für Modell ändern" höher oder niedriger einstellen. Ca. "12" sollte nicht unterschritten werden.

Der Idealzustand ist also gegeben, wenn die "Konvergenz" erreicht ist und die "Verbesserung" noch oberhalb des Grenzwerts (0.000000001) liegt. Dieser Fall tritt in der Regel schon nach 4 bis 8 Iterationsschritten ein.

4. Eingabefeld: Das Iterationsprotokoll
Almo gibt für das in Prog22m5 gerechnete Beispiel folgendes Iterationsprotokoll für die **Originalstichprobe** aus

```
Iterationsprotokoll
-------------------
         Konvergenz                                      Likelihood-Ratio
 it     1.Ableitund    Log-ML-Wert   Verbesserung        Testgroesse
 ==========================================================================
  0    1.48700e+003  -4.99869e+002   1.00000e+070    9.99737e+002

       ..........................................................
       Reg.koeff.
           (1)      0.00000    0.00000    0.00000    0.00000
                    0.00000    0.00000
           (2)      0.00000    0.00000    0.00000    0.00000
                    0.00000    0.00000
--------------------------------------------------------------------------
  1    6.18239e+002  -4.22587e+002   7.72818e+001    8.45174e+002
       ..........................................................
       Reg.koeff.
           (1)      9.02359   -1.38361    1.22257    0.59236
                   -0.25303    0.18676
           (2)      4.56202   -0.50697    0.43537    0.48854
                   -0.13226    0.14284
--------------------------------------------------------------------------
  2    9.44194e+001  -4.12828e+002   9.75903e+000    8.25656e+002
       ..........................................................
       Reg.koeff.
           (1)     13.13676   -1.96149    1.69262    0.75903
                   -0.37285    0.29915
           (2)      7.67440   -0.96215    1.00683    0.46069
                   -0.21654    0.22341
--------------------------------------------------------------------------
  3    8.85301e+000  -4.12113e+002   7.14767e-001    8.24226e+002
       ..........................................................
```

```
       Reg.koeff.
           (1)     14.16195   -2.11357    1.99194    0.78491
                   -0.40326    0.33156
           (2)      8.45768   -1.07338    1.26392    0.46884
                   -0.23940    0.24976
-------------------------------------------------------------------
  4    2.15784e-001  -4.12101e+002   1.17675e-002    8.24202e+002
       .....................................................
       Reg.koeff.
           (1)     14.23259   -2.12491    2.04564    0.78597
                   -0.40541    0.33413
           (2)      8.51477   -1.08262    1.31732    0.46945
                   -0.24111    0.25190
-------------------------------------------------------------------
  5    1.28461e-004  -4.12101e+002   6.62117e-006    8.24202e+002
       .....................................................
       Reg.koeff.
           (1)     14.23315   -2.12502    2.04705    0.78598
                   -0.40542    0.33416
           (2)      8.51525   -1.08272    1.31873    0.46945
                   -0.24112    0.25192
-------------------------------------------------------------------
  6    5.38005e-011  -4.12101e+002   3.29692e-012    8.24202e+002
       .....................................................
       Reg.koeff.
           (1)     14.23315   -2.12502    2.04705    0.78598
                   -0.40542    0.33416
           (2)      8.51525   -1.08272    1.31873    0.46945
                   -0.24112    0.25192
-------------------------------------------------------------------
```

Betrachten wir Iterationsschritt 5

```
         Konvergenz                                    Likelihood-Ratio
 it     1.Ableitund    Log-ML-Wert   Verbesserung     Testgroesse
===================================================================
  .
  .
-------------------------------------------------------------------
  5    1.28461e-004  -4.12101e+002   6.62117e-006    8.24202e+002
       .....................................................
       Reg.koeff.
           (1)     14.23315   -2.12502    2.04705    0.78598
                   -0.40542    0.33416
           (2)      8.51525   -1.08272    1.31873    0.46945
                   -0.24112    0.25192
-------------------------------------------------------------------
  .
  .
-------------------------------------------------------------------
```

Zuerst wird die Konvergenz ausgegeben. Mit `0.000128461` ist sie noch minimal größer als der angestrebte Grenzwert von 0.0001. Die Verbesserung hat mit `6.62117e-006` noch nicht den Grenzwert von `1.00000e-009` unterschritten. Es muss also ein weiterer Iterationsschritt gerechnet werden. Beim 6. Schritt wird dann Konvergenz erzielt.
In der ersten Zahlenreihe wird noch der Log-Maximum-Likelihood-Wert und die Likelihood-Ratio-Testgröße ausgegeben. Sie ist gleich dem -2*Log-ML-Wert und könnte somit auch dazu verwendet werden die Verbesserung zu ermitteln. Beides sind Maßzahlen für die Güte des Modells.

In einem 2. Teil werden die Regressionskoeffizienten in fortlaufender Folge für die 1. Ausprägung "Wohnlage: Land" und die 2. Ausprägung "Wohnlage: Stadtrand" ausgegeben - so wie sie sich im 5. Iterationsschritt ergeben haben.

Vergleicht man die aufeinander folgenden 6 Werte für die Konvergenz und für die Verbesserung, dann erkennt man, dass sie sich allmählich den Grenzwerten nähern und im 6. Schritt diese unterschreiten.

```
              Konvergenz
Iteration    1.Ableitund      Verbesserung
---------   ------------      ------------
     1      6.18239e+002      7.72818e+001
     2      9.44194e+001      9.75903e+000
     3      8.85301e+000      7.14767e-001
     4      2.15784e-001      1.17675e-002
     5      1.28461e-004      6.62117e-006
     6      5.38005e-011      3.29692e-012

Grenzwert  1.00000e-004      1.00000e-009
```

Für die **Bootstrap-Stichproben** werden nur Konvergenz, Verbesserung und Iterationszahl für den *letzten* Iterationsschritt ausgegeben, nicht die jeweils errechneten Regressionskoeffizienten. Almo liefert folgende (hier gekürzte) Ausgabe

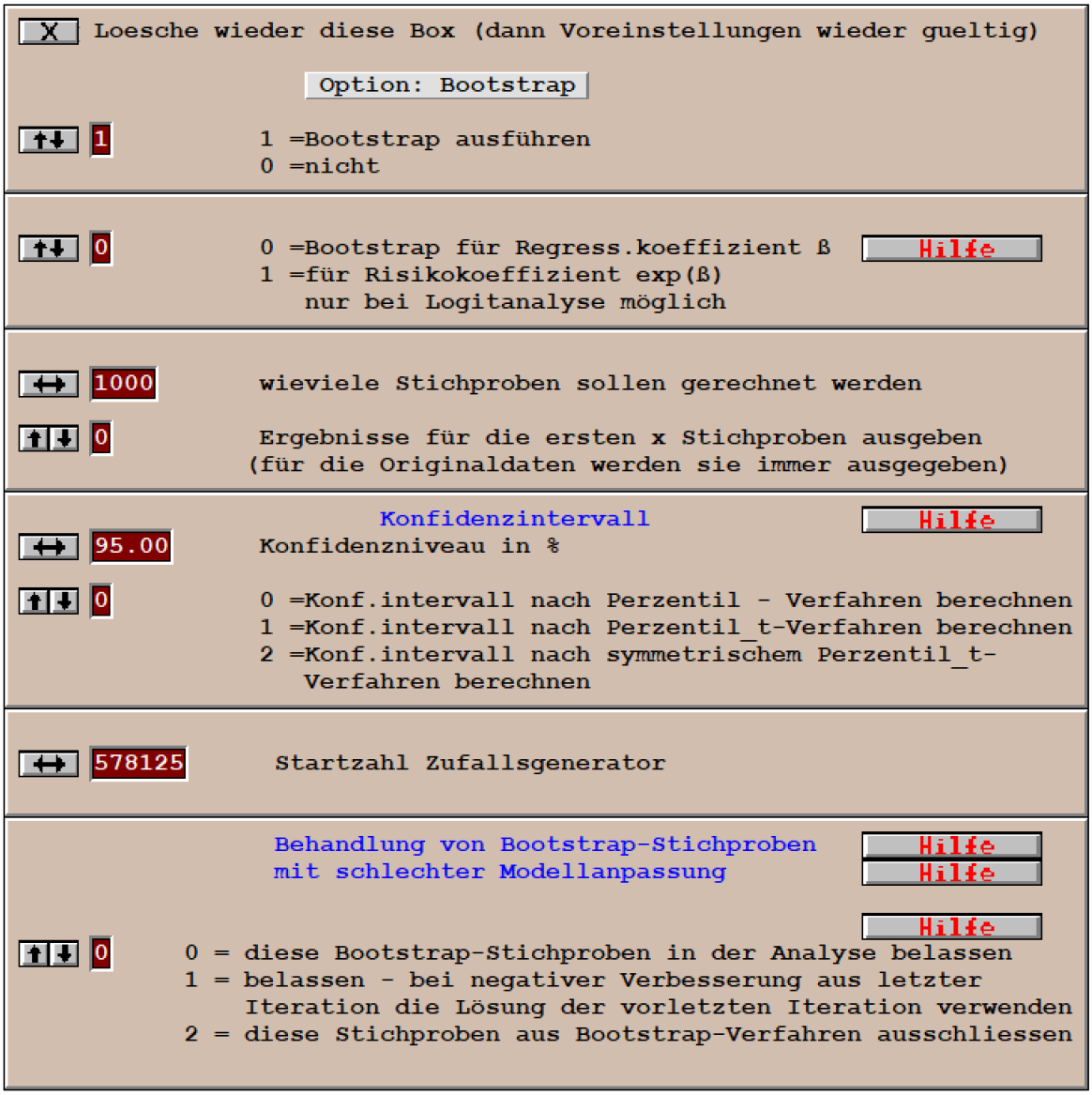
```
---------------------------------------------------------------
Iterationsprotokoll für alle weiteren Stichproben
---------------------------------------------------------------
               Konvergenz
Grenzwerte:   1.Ableitung        Verbesserung       Iterationen
              1.00000e-004      1.00000e-009           20

  *) =bedeutet: keine Konvergenz erzielt
 **) =bedeutet: keine Konvergenz erzielt und letzte Verbesserung ist negativ
 *#) =bedeutet: keine Konvergenz erzielt und letzte Verbesserung ist negativ
      aber vorhergehende Lösung uebernommen

Stichprobe    1.Ableitung       Verbesserung        Iterationen
----------    ------------      ------------        -----------
    2         3.51950e-011      2.55795e-012            6
    3         8.09877e-005      4.64576e-006            5
    4         6.81045e-007      4.16781e-008            5
    .              .                 .                  .
    .              .                 .                  .
  605         2.39623e-008      1.24339e-009            6
  606         1.05244e+001 **) -8.71510e-002            6
  607         1.92338e-008      8.93920e-010            6
    .              .                 .                  .
    .              .                 .                  .
  614         6.74708e+000 **) -1.57516e-001            7
  615         9.96668e-006      4.62474e-007            5
    .              .                 .                  .
    .              .                 .                  .
  998         3.61135e-006      2.01483e-007            5
  999         9.79875e-007      4.99446e-008            6
 1000         3.34763e-011      1.64846e-012            6
```

In Bootstrap-Stichprobe 606 und 614 ist der Konvergenzwert nach 6 bzw. 7 Iterationen noch weit vom Grenzwert entfernt und die Verbesserung ist sogar negativ. In der Bootstrap-Optionsbox, die nachfolgend erläutert wird, wird dem Benutzer eine Möglichkeit angeboten, solche Stichproben aus der Analyse auszuschließen.

P22.3.5 Die Bootstrap-Optionsbox



Folgende Koeffizienten der Logit- bzw. Probitanalyse werdem dem Bootstrap-Verfahren unterworfen:

1. Die Regressionskoeffizienten der unabhängigen quantitativen Variablen
2. Die Regressionskoeffizienten der Dummies der unabhängigen nominalen Variablen
3. Die paarweisen Vergleiche (Kontraste) der Dummies der jeweiligen unabhängigen nominalen Variablen
4. Optional die Risikokoeffizienten exp(ß) aller unabhängigen Variablen. Nicht bei der Probitanalyse möglich.

Für diese Koeffizienten werden der Standardfehler, die Signifikanz p und das Konfidenzintervall durch Bootstrap ermittelt.

Am rechten Rand der oben abgebildeten Optionsbox sind die Nummern der Eingabefelder eingeblendet. Die folgenden Erläuterungen zu den Eingabefeldern sind zum goßen Teil aus dem Almo-Dokument 13b zum Bootstrap beim Allgemeinen Linearen Modell entnommen.

Eingabefeld 1: Bootstrap 1=ausführen, 0=nicht
Wird 0 eingegeben oder die Optionsbox geschlossen, dann wird eine Analyse nur für die originale Stichprobe gerechnet. Gegenüber der Programm-Maske Prog22m ist die Zahl

der Optionen geringer.

Eingabfeld 2: Regressionskoeffizienten ß oder Risikokoeffizienten exp(ß)
Almo berechnet den Standardfehler, die Signifikanz p und das Konfidenzintervall der unabhängigen Dummies und der unabhängigen quantitativen Variablen (Kovariate)

0 = für die Regressionskoeffizienten ß oder alternativ
1 = für die Risikokoeffizienten exp(ß)

Bei der Probitanalyse ist nur die erstgenannte Alternative möglich.

Werden die Risikokoeffizienten exp(ß) anstelle der Regressionskoeffizienten ß dem Bootstrap-Verfahren unterworfen, dann kann nur durch das einfache Perzentil-Verfahren, das Konfidenzintervall und der p-Wert ermittelt werden. Wir werden in Eingabefeld 6 darauf zurückkommen. Almo schaltet automatisch auf das einfache Perzentil-Verfahren um.
Da exp(ß) eine monotone Transformation von ß ist, wird durch das einfache Perzentil-verfahren für beide der gleiche p-Wert errechnet.

Die e-Funktion von ß führt bei hohen ß-Werten zu extrem hohen Werten für den Risikokoeffizienten. Rechnet man das Datenbeispiel in Prog22m5 mit "Urlaub" als abhängiger Variablen, dann erhält man folgendes Bootstrap-Ergebnis für die 1. Ausprägung "Urlaub: zuhause":

	Risikokoeffizient exp(ß)			Standard	Signif.	Konfidenzintervall Konf.niv=0.950	
	original	Bootstrap	Verzerr.	fehler	p	unten	oben
A1 Geschlec: männl	11.331140	74.238416	62.907276	264.44284	0.004000	2.352111	671.5516
B1 Herkunft: Untersch	**0.319631**	**3180.7590**	**3180.4393**	**93758.113**	**0.222000**	**0.029786**	**2.175520**
B2 Herkunft: Mittelsc	0.488378	5694.4862	5693.9978	163253.08	0.410000	0.072199	2.690698
V11 Alter	0.490380	0.476939	-0.013442	0.080641	0.000999	0.332091	0.652928
V1 Bildungsniveau	0.945142	0.939582	-0.005561	0.101574	0.546000	0.765265	1.142980
Konstante	*x)	*x)	*x)	*x)	*x)	*x)	*x)

Betrachten wir die unahängige Dummy "Herkunft: Unterschicht". Der Risikokoeffizient aus der Original-Stichprobe ist `0.319631`. Der Mittelwert aus allen 1000 Bootstrapstichproben ist `3180.7590` und die Verzerrung dadurch `3180.4393`. Wir haben uns alle 1000 ß- und exp(ß)-Werte ausgeben lassen. In der Bootstrapstichprobe 551 finden wir ß=`14.8996` und exp(ß)=`295667.0` und in der Stichprobe 616 für ß=`12.3176` und für exp(ß)=`223604`. Es sind nur diese beiden Bootstrapstichproben, die den extrem hohen Mittelwert zu verantworten haben. Natürlich ist dann auch der Standardfehler durch diese beiden extremen Werte betroffen. Da mit dem "einfachen Perzentil-Verfahren" gerechnet wird, bei dem die hohen Werte nach hinten sortiert werden, nehmen diese beiden Extremwerte die letzten Plätze 999 und 1000 in der Sortierfolge der `exp(ß)`-Werte ein. Die Signifikanz p und die Konfidenzgrenzen werden dadurch nicht beeinflusst. Hier ist ein Ausschnitt der Sortierfolge der ß- und exp(ß)-Werten für die die unahängige Dummy "Herkunft: Unterschicht".

Stichprobe	ß	exp(ß)
1	-4.48033	0.01132
2	-4.34312	0.01299
3	-4.30611	0.01348
4	-4.27619	0.01389
5	-4.22807	0.01458
6	-3.96651	0.01893

```
   .       .            .
   .       .            .
 889    -0.00328       0.99672
 890     0.00077       1.00078
   .       .            .
   .       .            .
 994     1.20894       3.34994
 995     1.44484       4.24116
 996     1.75725       5.79649
 997     1.77333       5.89042
 998     2.25332       9.51928
 999    12.31760   223604.0
1000    14.89960   295667.0
```

Almo-intern wird mit `exp(ß)-1` gerechnet. Zwischen den (sortierten) Stichproben 889 und 890 wird die Stelle `ß=0` bzw. `exp(ß)-1=0` überschritten. Es liegen also 110 Stichproben zwischen 0 und dem oberen Ende der Sortierfolge. Das (optimale) Konfidenzintervall schließt somit `1000-2*110=780` Werte ein. Außerhalb liegen `2*110=220` Werte. Das "optimale Konfidenz*niveau"* op (das 0 gerade noch ausschließt) ist dann `op = 780/1000 = 0.78 (bzw. 78%)`
und die zweiseitige Signifikanz p `p = 1-op = 0.22`

Siehe nachfolgend "Eingabefeld 5" und sehr ausführlich Abschnitt P22.3.6.1

Eingabefeld 3: Zahl der Bootstrap-Stichproben.

Empfohlen sind mindestens 1000. Die Rechenzeit beträgt für unser Beispiel nur einige Sekunden. Sie hängt direkt von der gewählten Stichprobenzahl ab. Erhöht man die Stichprobenzahl von beispielsweise 1000 auf 1500, dann erzeugt man dadurch Bootstrap-Ergebnisse, die vielleicht an der 3. Kommastellen verändert sind. Man darf aber nicht unterstellen, dass sie "besser" sind, d.h. dass sie den "wahren" Werten in der Grundgesamtheit näher kommen. Entscheidend ist die Qualität der Originalstichprobe. Ist sie verzerrt, dann wird sie durch Bootstrapping nicht repräsentativ. Wichtig ist, dass man die Stichprobenzahl an das Konfidenzniveau anpasst. Wir werden bei der Erläuterung zu Eingabefeld 4 (Konfidenzniveau) darauf zurückkommen. Siehe auch Abschnitt P20.25.9. im Almo-Dokument 13b "Bootstrap beim Allgemeinen Linearen Modell". Dort wird gezeigt welche Auswirkungen eintreten, wenn die Stichprobenzahl erhöht wird.

Eingabefeld 4: Ergebnisse ausgeben.

0 = die Ergebnisse für die Originalstichprobe werden ausgegeben. In einem deutlich getrennten 2. Ausgabeteil werden danach die kumulierten Bootstrap-Ergebnisse für die Gesamtzahl der Stichproben ausgegeben.

x = wird beispielsweise 3 eingesetzt, dann werden die Ergebnisse für die Originalstichprobe und zusätzlich für die Bootstrap-Stichproben 1, dann 2, dann 3 und erst danach die zusammengefassten Bootstrap-Ergebnisse für die Gesamtzahl der Stichproben ausgegeben

Eingabefeld 5: Konfidenzniveau für Konfidenzintervall

Der Benutzer bestimmt das Konfidenzniveau. Üblich ist ein Niveau von 95%. Werden die Regressionskoeffizienten aus z.B. 1000 Bootstrap-Stichproben der Größe nach (aufsteigend) sortiert, dann liegen die Grenzwerte des Konfidenzintervalls 2.5%, also 25 Werte unterhalb bzw. oberhalb des maximalen bzw. minimalen Werts. 950 Werte liegen zwischen den Intervallgrenzen. 25 Werte sind wenig, besser wäre es, 2000 Stichproben zu rechnen. Dann liegen 50 Werte ober- und unterhalb der Grenzwerte. Wird ein Konfidenzniveau von 99% gewählt, dann liegen bei 1000 Stichproben nur 5 Werte

ausserhalb der Intervallgrenzen. Erst mit 5000 Stichproben werden 25 Werte und mit 10 000 Stichproben 50 Werte außerhalb der Intervallgrenzen erreicht. Der Benutzer sollte die Stichprobenzahl an das Konfidenzniveau (oder umgekehrt) anpassen. Ziel muss sein, möglichst viele Werte unter- bzw. oberhalb der Grenzwerte des Konfidenzintervalls zu erhalten. Je mehr aufsteigend sortierte Stichprobenwerte vorliegen umso feiner sind die Differenzen von einem Wert zum nächsten, umso genauer können die Grenzwerte bestimmt werden. Je höher auch der Benutzer das Konfidenzniveau ansetzt, umso mehr nähern sich der obere und untere Grenzwert dem maximalen bzw. minimalen Wert an, umso breiter wird das dazwischen liegende Konfidenzintervall.

Eingabefeld 6: Konfidenzintervall

Almo bietet drei Methoden für die Berechnung des Konfidenzintervalls an. Dies sind
0 = das einfache Perzentil-Verfahren
1 = das asymmetrische Perzentil-t -Verfahren
2 = das symmetrische Perzentil-t -Verfahren
Diese drei Verfahren werden in folgenden Abschnitten ausführlich erläutert

Vom gewählten Verfahren hängt auch ab, wie die Signifikanz, genauer der p-Wert des untersuchten Koeffizienten (z.B. des Regressionskoeffizienten) ermittelt wird. Die beiden Perzentil-t -Verfahren gelten hier als dem einfachen Perzentil-Verfahren überlegen. Die drei Verfahren, auch die Berechnung des p-Wertes, werden in den nachfolgenden Abschnitten P22.3.6 und P22.3.7 ausführlich beschrieben.

Eingabefeld 7: Startzahl des Zufallsgenerators.

Der Benutzer kann die Startzahl beliebig verändern. Wird eine zweite Analyse mit der gleichen Startzahl gerechnet, dann entsteht exakt dieselbe Folge von Zufallszahlen, wodurch aus der Originalstichprobe dieselben Probanden für die Bootstrap-Stichprobe ausgewählt werden, wie für die erste Analyse. Damit sind auch die Ergebnisse identisch.

Eingabefeld 8: Behandlung von Bootstrap-Stichproben mit schlechter Modellanpassung

Eine "schlechte Modellanpassungt" liegt vor, wenn der Newton-Raphson-Iterations-Kalkül nicht "konvergiert", d.h. die "absolut größte 1. Ableitung" den Grenzwert 0.0001 (1 an der 4. Dezimalstelle) nicht erreicht oder unterschreitet. Wir haben oben in Abschnitt P22.3.4 diesen Grenzwert ausführlich beschrieben.

Folgende "Behandlungen" sind möglich. Im letzten Eingabefeld der Bootstrap-Optionsbox wird eingesetzt:

0 =Die Ergebnisse der betreffenden Bootstrap-Stichprobe werden für das Bootstrap-verfahren verwendet - auch wenn keine konvergente Lösung entstanden ist - jedoch mit einer Ausnahme. Wenn die aus dem Kalkül hervorgegangene Informationsmatrix singulär ist, dann können Ergebnisse schon gar nicht errechnet werden. Die betreffende Stichprobe wird übersprungen und zur nächsten weiter gegangen.

Bei Eingabe von "0" werden also auch Ergebnisse akzeptiert, die aus einer nicht-konvergenten Lösung stammen. In der Regel musste in diesem Fall das Iterations-verfahren abgebrochen werden, da eine negative "Verbesserung" entstand. Es ist eher selten, dass die vorgegebene maximale Iterationszahl überschritten wurde

1 =wie bei "0". Wurde das Iterieren in der Runde i mit eine negativen Verbesserung beendet, dann werden die Ergebnisse aus der vorherigen Runde i-1 übernommen. Es ist dann möglich, dass die "absolut größte 1. Ableitung" aus dieser vorletztenRunde i-1

vom vorgegebenen Grenzwert (=0.0001) weiter entfernt ist (also schlechter) als die aus der letzten Runde i. Man kann dann allerdings sicher sein, dass bis zu dieser Runde i-1 der Kalkül korrekt verlief. Beachte: Konvergierte der Iterationskalkül mit einer negativen Verbesserung in der letzten Iterationsrunde, dann wird Behandlung 1 nicht durchgeführt.

2 =die Stichprobe wird aus dem Bootstrapverfahren ausgeschlossen, wenn keine Konvergenz erzielt wurde und der Kalkül beendet werden musste, weil eine negativ "Verbesserung" entstand oder die maximale Iterationszahl überschritten wurde.
Kann der Iterations-Algorithmus schon in der originalen Stichprobe nicht konvergieren dann wird bei "2" das Bootstrap-Verfahren sofort abgebrochen.

P22.3.6 Das "einfache Perzentil"-Verfahren

Das Perzentil-Verfahren ist das "Herzstück" des Bootstrapverfahrens. Mit ihm wird das Konfidenzintervall und die Signifikanz p für die Regressionskoeffizienten ß bzw. die Risikokoeffizienten exp(ß) der unabhängigen Variablen berechnet.
Das "einfache Perzentil-Verfahren" verwendet dafür einen sehr einfachen und überschaubaren Kalkül. Das Perzentil-t -Verfahren in seinen zwei Varianten ist komplexer.
Die folgenden Ausführungen sind größtenteils aus dem Almo-Dokument 13b "Bootstrap beim Allgemeinen Linearen Modell", Abschnitt P20.25.5 bis P20.25.6.3 entnommen, können jedoch auf das Logit- und Probitmodell übertragen werden.

P22.3.6.1 Signifikanz p und Konfidenzintervall

Ob ein Koeffizient zweiseitig signifikant ist, wird zunächst daran erkannt, ob das für ihn festgestellte Konfidenzintervall bei dem vom Forscher geforderten Konfidenz*niveau* (von üblicherweise 95%) den Wert 0 einschließt. Ist das nicht der Fall, dann ist der Koeffizient mit `p=1-0.95=0.05` mindestens signifikant. Sein p-Wert kann aber noch kleiner sein. Soll die Signifikanz als genauer p-Wert ermittelt werden, dann geht es darum, dasjenige Konfidenz*niveau* zu finden, das ein Konfidenz*intervall* erzeugt, das gerade noch den Wert 0 unter- oder oberhalb seiner Grenzen positioniert. 1.0 minus diesem Konfidenzniveau/100 ist dann die Signifikanz p. Wir nennen dieses "das optimale Konfidenzniveau".

Wir rechnen eine Analyse mit 10 000 Stichproben, um das Konfidenzintervall und die Signifikanz p mit dem "einfachen Perzentil-Verfahren" möglichst genau bestimmen zu können. Die folgenden Daten entstanden aus dem Allgemeinen Linearen Modell könnten aber so auch aus der Logit- oder Probitanalyse entstanden sein.

Wir betrachten die Hauptdummy `A1 männlich`. Almo liefert diese aufsteigend sortierte Aufeinanderfolge der Effekte dieser Variablen (gemeint sind die Regressionskoeffizienten ß aus den 10000 Stichproben.

```
Bootsrap Stichproben ─────┐        ┌──────── Effekte von Variable A1
aufsteigend sortiert      │        │         aus 10 000 Bootstrap-Stichproben
                        1  -0.367582
                        2  -0.267036
                        3  -0.259893
                        .     .
                        .     .
                      208  -0.000266097
                      209  -0.000042759
                      ---------------- <--- Wert 0.0
                      210   0.000173536 <--- untere "optimale" Konfidenzgrenze
                      211   0.000316841      bei "optimalen" Konf.niveau 0.9582
                        .     .
                        .     .
                      249   0.012127
```

```
                          250   0.0126888
-----------------------------------------
 untere Konf.grenze --->  251   0.013056
 bei Konf.niv. 0.95         .     .
                            .     .
                         4992   0.371055
 Mittelwert von A1  ---> 4993   0.371063
 aus 10 000 Stichpr.     4994   0.371097
                            .     .
                            .     .
 obere  Konf.grenze      9749   0.720293
 bei Konf.niv. 0.95 ---> 9750   0.720791
-----------------------------------------
                         9751   0.720834
                            .     .
                            .     .
                         9999   1.00896
                        10000   1.06902
```

Betrachten wir zunächst die linke Hälfte dieser Tabelle. Sie zeigt wie die untere Grenze des Konfidenzintervalls gefunden wird. Es wurden 10 000 Stichproben gerechnet. Dadurch entstehen für jede Variable 10 000 Effekte, die aufsteigend sortiert wurden. In obiger Tabelle werden diese Werte für die Dummy `A1 Geschlecht männlich` stark gekürzt abgebildet. Das arithmetische Mittel aus dem Bootstrapping für A1 ist `0.371074`. Es liegt zwischen dem `4993.` und dem `4994.` Wert in der Sortierfolge. Im Vergleich dazu ist der Wert aus der Originalstichprobe `0.369082`.
Als Konfidenzniveau wurde `0.95` vorgegeben. Das bedeutet, dass `95%` der Werte zwischen den Intervallgrenzen liegen müssen und je `2.5%` unter- und oberhalb der Intervallgrenzen. Die untere Grenze des Konfidenzintervalls wird - wie ein Blick auf die obige Tabelle zeigt - gefunden, indem die aufsteigend sortierten 10 000 Werte bis zum Wert Nr. 250+1 abgezählt werden. Dort steht der Wert **`0.013056`**. Entsprechend wird vom oberen Ende bis zum Wert Nr. 9750 herunter gezählt. Dort findet man den Wert **`0.720791`**. Das ist der obere Grenzwert. Zwischen den Intervallgrenzen liegen somit 9500 Werte und außerhalb zusammen 500 Werte. Für das Konfidenzintervall wurden so die Grenzen **`0.013056`** bis **`0.720791`** gefunden. Wird das Konfidenzniveau auf z.B. `0.99` gesteigert, dann wird das Intervall breiter. Es würde dann vom **`50.`** Wert bis zum **`9950.`** Wert reichen

Der Wert `0.0` liegt unterhalb des Intervalls. Wir können also konstatieren, dass der Effekt der Variablen A1 mindestens mit `p=1-0.95=0.05` signifikant ist. Die Signifikanz könnte sogar noch besser sein, wenn es gelänge das Konfidenzniveau zu finden, das die untere Intervallgrenze gerade einen Wert über den Wert 0 legt. Das wäre dann das *optimale* Konfidenzniveau und (von 1.0 subtrahiert) die *Signifikanz p* für die Variable `A1`. In obiger Tabelle erkennt man, dass vom `209.` Wert zum `210.` der Null-Wert überschritten wird. Der `210.` Wert mit **`0.000173536`** ist dann der optimale untere Grenzwert. 209 Werte liegen unterhalb des Grenzwertes (und auch entsprechend 209 Werte oberhalb des oberen Grenzwertes).
Die zweiseitige Signifikanz p von `A1` ist dann `2*209/10000 = 0.0418`
und das optimale Konfidenzniveau für `A1` `1-2*209/10000 = 0.9582`

Wir verwenden folgende Notation

ß = Parameterwert aus der Originalstichprobe
ß*= Parameterwert aus einer Bootstrap-Stichprobe

Und können somit den aus dem einfachen Perzentil-Verfahren gewonnenen einseitigen p/2-Wert pragmatisch so definieren:

p/2 ist bei positivem Parameterwert ß die Wahrscheinlichkeit eine Stichprobe zu ziehen mit einem Parameterwert ß* kleinergleich 0, bzw.

bei negativem Parameterwert ß eine Stichprobe zu ziehen mit einem Parameterwert ß* größergleich 0.
Der zweiseitige p-Wert ergibt sich dann einfach aus der Verdopplung.

Wie soll verfahren werden, wenn die aufsteigend sortierten Werte keinen Wert 0 aufweisen bzw. keinen Übergang von negativen Werten zu positiven (oder umgekehrt) enthalten. Das ist dann der Fall, wenn die jeweilige unabhängige Variable die abhängige Variable stark beeinflusst, d.h. wenn der Effekt bzw. Regressionskoeffizient einen großen positiven oder (bei gegenläufigem Einfluß) großen negativen Wert besitzt. Auch wenn sehr viele Bootstrap-stichproben gerechnet werden, tritt keine auf, die für die Variable den Wert 0 oder sogar einen Wert jenseits von 0 aufweist. Das bedeutet, dass die Variable *hoch signifikant* wirkt. In dieser Situation muss der ungünstigste Fall unterstellt werden, dass gerade unterhalb bzw. oberhalb der aufsteigend sortierten Werte der Wert 0 folgen würde - hätte man eine weitere Stichprobe gerechnet. Almo berechnet somit das "optimale" Konfidenzniveau für ein Konfidenzintervall, dessen unterer Grenzwert der erste bzw. niedrigste Wert in der Sortierfolge ist und dessen oberer Grenzwert der letzte bzw. höchste Wert ist. Die zweiseitige Signifikanz ist dann sehr einfach `p=1/(Stichprobenzahl+1)` . Die Signifikanz der Variablen kann dann nur gleich diesem p-Wert oder besser (d.h. kleiner) sein. Sie ist nur durch die Stichprobenzahl bestimmt.

P22.3.7 Das Perzentil-t -Verfahren

Wir verwenden folgende Notation:

```
b  = Regressionskoeffizient einer Kovariaten (oder Effekt einer Dummy) aus der
     originalen Stichprobe
S  = Standardfehler von b aus der Originalstichprobe
b* = Regressionskoeffizient einer Kovariaten (oder Effekt einer Dummy aus den
     Bootstrap-Stichproben
S* = Standardfehler von b* aus den Bootstrap-Stichproben
K  = Konfidenzniveau/100  (wenn z.B. Benutzereingabe = 95, dann ist K=95/100=0.95)
a  = alpha = 1-K
n  = Zahl der Bootstrap-Sichproben
t  = t-Wert der Kovariaten bzw. Dummy aus originaler Stichprobe
t* = Perzentil-t -Wert der Kovariaten bzw. Dummy aus den Bootstrap-Stichproben
```

Für jede der 10 000 Bootstrap-Stichproben muss der `b*`-Wert und sein ihm zugehöriger Standardfehler `S*` erhoben werden. Aus den beiden und dem `b`-Wert aus der originalen Stichprobe wird ein **`t`**-Wert gebildet, den wir mit t* symbolisieren

`t* = (b*-b)/S*`

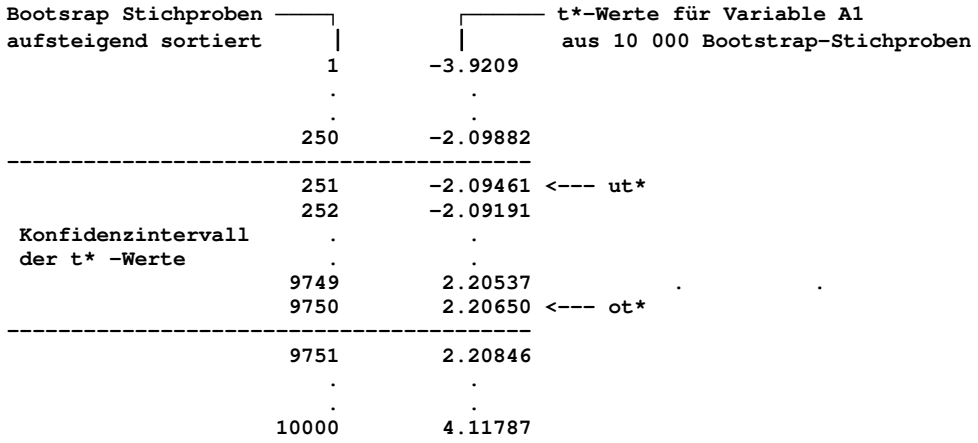
Nach Ablauf des Bootstraps verfügen wir also über 10 000 t*-Werte. Diese Koeffizienten werden, wie in P20.25.6 beim einfachen Perzentil-Verfahren beschrieben, aufsteigend sortiert und auf die Konfidenzgrenzen ausgezählt. Für die 10 000 t*-Werte wird somit, wie für die 10 000 b*-Werte beim einfachen Perzentil-Verfahren, keine t-Verteilung unterstellt.

Beachte: **`(b*-b)`** ist die "Verzerrung". Zu beachten ist auch, dass der **`t*-`**Wert negativ werden kann.

P22.3.7.1 Konfidenzintervall berechnet mit Perzentil-t -Verfahren

In unserem Beispiel soll das Konfidenz*niveau* **`K =95/100 =0.95`** sein. Dann ist **`a=0.05`**
Die 10 000 t*-Werte werden aufsteigend sortiert.

Hier ist die gekürzte Reihenfolge für die t*-Werte der Dummy-Variablen **`A1 (Geschlecht: männlich)`** aus unserem Beispiel



Vom niedrigsten t*-Wert an der Position 1 werden nach oben **n*a/2+1** Werte abgezählt. Das sind **10000*0.05/2+1=250+1** Werte. An der Position 251 steht also der t*-Wert für das untere Konfidenzintervall. Er hat den Wert **-2.09461.** Wir bezeichnen ihn mit **ut*.**
Vom höchsten t*-Wert an der Position 10000 werden nach unten **n*a/2 = 250** Werte abgezählt. Es wird also der **9750.** t* -Wert herausgegriffen. Wir bezeichnen ihn mit **ot***
ut* ist kleiner als ot*. Außerhalb des Intervalls befinden sich dann **5% =500** Werte und innerhalb **95% =9500** Werte. Aus **ut*** und **ot*** werden dann die Konfidenzgrenzen für A1 nach folgenden sehr einfachen Formeln berechnet.

Die untere Konfidenzgrenze ist **b – S * ot***
=0.369082–0.155913*2.20650 =0.025060

und die obere **b – S * ut***
=0.369082–0.155913*–2.09461=0.695659

b = das ist der Effekt von A1 aus der originalen Stichprobe
S = das ist dessen Standardfehler
Das Intervall ist nicht symmetrisch um **b** herum**.**

Im Vergleich dazu wurde mit dem einfachen Perzentil-Verfahren das Intervall **0.013056 – 0.720791** gefunden.

P22.3.7.2 Signifikanz p berechnet mit Perzentil-t -Verfahren

Die Perzentil-t -Werte **t*** aus den Bootstrapstichproben für die unabhängige Variablen (in unserem Beispiel: die Dummy A1) werden quadriert. Wir bezeichnen sie mit **(t*)²**. Ebenso wird der eine t-Wert für die Dummy A1 aus der originalen Stichprobe quadriert. Wir bezeichnen ihn mit **t²**. Dann wird gezählt: Wie oft ist **(t*)²** größer/gleich **t²**. Wir bezeichnen das Zählergebnis mit **z**.

Die Signifikanz p ist dann **p = z/n**

Für die Dummy A1 wird ein Wert von **p = 0.031** berechnet. mit dem einfachen Perzentil-Verfahren wurde **p=0.0418** ermittelt.

Ist **z=0** dann wird gerechnet (wie beim einfachen Perzentil-Verfahren) **p = 1/(n+1)**
In diesem Fall muss interpretiert werden, dass der tatsächliche p-Wert mindestens **1/(n+1)** ist oder kleiner, d.h. signifikanter. **p** ist dann nur durch die Stichprobenzahl bestimmt.

P22.3.7.3 Das symmetrische Perzentil-t -Verfahren

Der oben definierte t* -Wert wird absolut gesetzt

t' = abs(t*)

Die t' -Werte werden aufsteigend sortiert. Der **n*K = n*(1-a) = 1000*0.95 = 950.** Wert wird herausgegriffen. Wir bezeichen ihn mit **ot'.** Für die Dummy A1 beträgt er **2.15277**

Die untere Konfidenzgrenze ist dann **b - S * ot'**
=0.369082 - 0.155913*2.15277=0.033439

und die obere **b + S * ot'**
=0.369082 + 0.155913*2.15277=0.704725

Das Intervall liegt symmetrisch um **b**

Literatur zu den hier beschriebenen Perzentil-Verfaren

C.J. Elias beschreibt in seiner Arbeit zu "Percentile and Percentile-t Bootstrap Confidence Intervals" (2013) die drei hier dargestellten Verfahren. Er führt ein Simulationsexperiment durch, bei dem er feststellt, dass die beiden Perzentil-t -Verfahren, das symmetrische und das asymmetrische, dem einfachen Perzentil-Verfahren überlegen sind. Dieses Ergebnis darf aber keinesfalls verallgemeinert werden. In der Literatur sind viele derartige Simulationen zu finden, die andere Ergebnisse erbracht haben, insbesondere auch Ergebnisse, die das einfache Perzentil-Verfahren favorisierten. Der Leser suche im Internet unter dem Suchwort "Percentil Bootstrap".

P22.3.8 Bootstrap-Ergebnisse

Almo gibt zuerst die Ergebnisse für die Originalstichprobe aus. Sie werden ausführlich im Almo-Dokument 9 "Logitanalyse" Abschnitt P22.2.3 erläutert. Sofern in der Bootstrap-Optionsbox angefordert wurde, auch für die ersten x Bootstrapstichproben die vollen Ergebnisse anzuzeigen, dann werden auch diese ausgegeben. Danach folgen die Bootstrap-Ergebnisse. Im Daten-Beispiel in der Programm-Maske Prog22m5 wird eine multinomiale Logitnanalyse gerechnet. Die abhängige Variable besitzt 3 Ausprägungen. Die 3. und letzte Ausprägung wird als Referenz verwendet mit der die ersten beiden Ausprägungen verglichen werden. Für die 3. Ausprägung entstehen keine Ergebnisse. Hier sind die Ergebnisse aus Prog22m5.

```
=========================================================================
             Ergebnisse aus Bootstrap mit 1000 Stichproben
=========================================================================

                                        ******************** WARNUNG
                                     Zahl der Bootstrap-Stichproben
             bei denen der Newton-Raphson-Algorithmus nicht konvergierte
                          und mit einer negativen Verbesserung bendetwurde
             die aber trotzdem in die Auswertung miteinbezogen wurden: 3

Zahl der angeforderten Bootstrap-Stichproben-Ergebnisse: 1000
Zahl der ausgewerteten Bootstrap-Stichproben-Ergebnisse: 1000


Einstellungen
-------------
Nominales Logitmodell für abhängige nominale (dichotome oder polytome) Variable
Bootstrap der Regressionskoeffizienten ß und der paarweisen Vergleiche

Startzahl für Zufallsgenerator: 578125
```

Konfidenzintervall u. Signifikanz p werden
bei Bootstrap der Regr.koeffiz. und
paarweisen Vergleiche nach dem einfaches Perzentil-Verfahren berechnet

Konfidenzniveau: 95%

kleinst möglicher berechenbarer p-Wert=0.000999
gerundet=0.0010

Bootstrap-Ergebnisse fuer Regressionskoeffizienten ß
fuer 1. Ausprägung "Land "
der abhaengigen Variablen V4 Wohnlage
(als Referenz wird die letzte Ausprägung "Stadt" verwendet)

	*a			*b	*c	*d Konfidenzintervall
	Regressionskoeffizient ß			Standard	Signif.	Konfniv=0.950
	original	Bootstrap	Verzerr.	fehler	p	unten oben
A1 Geschlec: männlich	-2.125021	-2.172206	-0.047185	0.551098	0.000999	-3.243195 -1.110504
B1 Herkunft: Untersch	2.047052	2.126773	0.079721	0.683010	0.000999	0.976283 3.444383
B2 Herkunft: Mittelsc	0.785977	0.781353	-0.004624	0.400132	0.070000	-0.048515 1.533557
V11 Alter	-0.405425	-0.414768	-0.009343	0.063056	0.000999	-0.551138 -0.305340
V1 Bildungsniveau	0.334158	0.341163	0.007005	0.073561	0.000999	0.200114 0.488134
Konstante	14.233147	14.581941	0.348795	2.430605	0.000999	10.091353 19.660570

Paarweise Vergleiche (Kontraste) der Regressionskoeffizienten

	*a			*b	*c	*d
	Regressionsdifferenz			Standard	Signif.	Konfidenzinzervall
	original	Bootstrap	Verzerr.	fehler	p	unten oben
A1 - A2	-2.125021	-2.172206	-0.047185	0.551098	0.000999	-3.243195 -1.110504
B1 - B2	1.261075	1.345420	0.084345	0.648200	0.006000	0.299599 2.614418
B1 - B3	2.047052	2.126773	0.079721	0.683010	0.000999	0.976283 3.444383
B2 - B3	0.785977	0.781353	-0.004624	0.400132	0.070000	-0.048515 1.533557

Bootstrap-Ergebnisse fuer Regressionskoeffizienten ß
fuer 2. Ausprägung "Stadtrand "
der abhaengigen Variablen V4 Wohnlage
(als Referenz wird die letzte Ausprägung "Stadt" verwendet)

	*a			*b	*c	*d Konfidenzintervall
	Regressionskoeffizient ß			Standard	Signif.	Konfniv=0.950
	original	Bootstrap	Verzerr.	fehler	p	unten oben
A1 Geschlec: männlich	-1.082716	-1.101908	-0.019192	0.545182	0.064000	-2.111354 0.027748
B1 Herkunft: Untersch	1.318728	1.401908	0.083180	0.653521	0.016000	0.302466 2.792565
B2 Herkunft: Mittelsc	0.469454	0.473803	0.004349	0.318560	0.150000	-0.178096 1.058110
V11 Alter	-0.241124	-0.246344	-0.005220	0.053792	0.000999	-0.353384 -0.143249
V1 Bildungsniveau	0.251916	0.257552	0.005636	0.062289	0.000999	0.131894 0.382448
Konstante	8.515251	8.702150	0.186900	2.205475	0.000999	4.461253 13.095238

Paarweise Vergleiche (Kontraste) der Regressionskoeffizienten

	*a			*b	*c	*d
	Regressionsdifferenz			Standard	Signif.	Konfidenzinzervall
	original	Bootstrap	Verzerr.	fehler	p	unten oben
A1 - A2	-1.082716	-1.101908	-0.019192	0.545182	0.064000	-2.111354 0.027748
B1 - B2	0.849273	0.928104	0.078831	0.626030	0.074000	-0.107246 2.145541
B1 - B3	1.318728	1.401908	0.083180	0.653521	0.016000	0.302466 2.792565
B2 - B3	0.469454	0.473803	0.004349	0.318560	0.150000	-0.178096 1.058110

***a "original" bezeichnet den Wert aus der Originalstichprobe**

```
   "Bootstrap" bedeutet Mittelwert aus 1000 Bootstrap-Stichproben
   "original" bezeichnet den Wert aus Originalstichprobe
   mit "Verzerr." (Verzerrung) wird die Differenz zwischen dem Mittelwert aus den
   Bootstrap-Stichproben minus dem Wert aus der Originalstichprobe bezeichnet
 *b Der Standardfehler ist gleich der Standardabweichung der Werte aus den
   Bootstrap-Stichproben
 *c Berechnet wird die zweiseitige Signifikanz p
   Beim symmetrischen und asymmetrischen Perzentil-t -Verfahren entsteht der
   gleiche p-Wert. Beim einfachen Perzentil-Verfahren entsteht ein geringfügig
   anderer p-Wert
 *d Konfidenzintervall (nach vom Benutzer vorgegebenen Konfidenz-Niveau)
   Das Konfidenzniveau ist 95.00%. Beim "einfachen" Perzentil-Verfahren bedeutet
   das: Von den aufsteigend sortierten 1000 Werten aus den Bootstrap-Stichproben
   befinden sich
   95.00% der Werte zwischen den Konfidenzgrenzen und je 2.50% oberhalb und
   unterhalb der Konfidenzgrenzen
*e) Die alfa-Schwellenwerte werden im Almo-Dokument Nr.9 "Logitanalyse", Abschnitt
   P22.2.6 ausführlich erläutert
*x) Wird das Bootstrap-Verfahren auf den Risikokoeffizienten exp(ß) angewendet
   dann koennen die so markierten Koeffizienten nicht berechnet werden
```

P22.3.8.1 Inhaltliche Interpretation der "paarweisen Vergleiche"

Betrachten wir die obige 1. Zeile

	Regressionsdifferenz			Standard	Signif.	Konfidenzinzervall	
	original	Bootstrap	Verzerr.	fehler	p	unten	oben
A1 - A2	-2.125021	-2.172206	-0.047185	0.551098	0.000999	-3.243195	-1.110504

Der Vergleich zwischen Männer (A1) und Frauen(A2) bezieht sich auf die abhängige Variable "Wohnlage: Land" im Vergleich zu "Wohnlage: Stadt". Wir müssen also einen doppelten Vergleich anstellen. Die Regressionsdifferenz ist negativ. Das bedeutet, dass Männer einen niedrigeren Wert haben als Frauen. Man wäre nun geneigt zu interpretieren:

> **Frauen haben eine stärkere Tendenz als Männer ihren Wohnsitz auf dem Land zu haben**

Diese Formulierung ist jedoch falsch. Korrekt ist es so zu formulieren:

> **Frauen haben eine stärkere Tendenz als Männer ihren Wohnsitz auf dem Land zu haben *als in der Stadt***

Die Referenzkategorie *muss* angegeben werden.
Diese Tendenz ist mit **-2.125021** recht stark und mit **p=0.000999** hoch signifikant.
Wird anstelle der "Stadt" der "Stadtrand" als Referenzkategorie für die unabhängige Variable "Wohnlage: Land" eingesetzt, dann entsteht folgender paarweiser Vergleich

A1 - A2	-1.042304	-1.070298	-0.027993	0.343463	0.002000	-1.747470	-0.417145

Die Tendenz geht mit **-1.042304** und **p=0.002000** zwar noch in dieselbe Richtung, ist jetzt nur noch halb so stark. Es ist durchaus möglich, dass beim Wechsel der Referenz die Tendenz (das Vorzeichen) umgedreht wird.

Anmerkung: Almo verwendet in der Bootstrap-Programm-Maske Prog22m5 und der Standard-Programm-Maske Prog22m die letzte Ausprägung der abhängigen Variablen als Referenz. Der Benutzer kann die Referenz dadurch wechseln, dass er die abhängige Variable umkodiert. Die Referenzkategorie wird an die letzte Stelle gesetzt. In unserem Fall müssten die 2. und 3. Ausprägung vertauscht werden. Die Umkodierungs-anweisung würde so lauten:

```
WohnlageII = Wohnlage(2=3; 3=2);
```

Auch die Ausprägungsnamen müssen vertauscht werden:

```
Name 14=WohnlageII:(1)Land, (2)Stadt, (3)Stadtrand;
```

"Stadtrand" ist jetzt die letzte Ausprägung. Siehe die besondere Art und Weise, wie die abhängige Variable bei Logit- und Probitanalyse umkodiert werden muss, in Abschnitt P22.3.3.1.

P22.3.8.2 Vergleich mit SPSS

Zum Vergleich das Bootstrap-Ergebnis aus SPSS

Bootstrap für Parameterschätzer

wohnlage		B	Bootstrap[a] Verzerrung	Standardfehler	Sig. (2-seitig)	95% Konfidenzintervall Unterer	Oberer
1,00	Konstanter Term	14,233	,381	2,472	,001	9,847	19,841
	alter	-,405	-,010	,064	,001	-,545	-,293
	bildungsniveau	,334	,008	,077	,001	,196	,505
	[geschlecht=1,0	-2,125	-,058	,542	,001	-3,258	-1,175
	[geschlecht=2,0	0	0	0		0	0
	[herkunft=1,00]	2,047	,098	,793	,002	1,107	3,470
	[herkunft=2,00]	,786	-,001	,377	,030	,057	1,579
	[herkunft=3,00]	0	0	0		0	0
2,00	Konstanter Term	8,515	,257	2,188	,001	4,740	13,367
	alter	-,241	-,008	,053	,001	-,361	-,153
	bildungsniveau	,252	,010	,064	,001	,140	,395
	[geschlecht=1,0	-1,083	-,025	,530	,034	-2,122	-,111
	[geschlecht=2,0	0	0	0		0	0
	[herkunft=1,00]	1,319	,090	,804	,009	,346	2,675
	[herkunft=2,00]	,469	,002	,312	,120	-,104	1,100
	[herkunft=3,00]	0	0	0		0	0

a. Sofern nicht anders angegeben, beruhen die Bootstrap-Ergebnisse auf 1000 Bootstrap-Stichproben

Die Regressionskoeffizienten aus der Originalstichprobe von Almo und SPSS (in Spalte B) stimmen exakt überein. Die verschiedenen Bootstrap-Ergebnisse sind minimal verschieden. Die muss so sein, da die beiden Statistikprogramme verschiedene Zufallsgeneratoren verwenden. Die "paarweisen Vergleiche" werden in SPSS nicht ermittelt. Das SPSS-Syntaxprogramm kann aus Almo unter dem Namen "SPSS_Logit.sps" aus dem Ordnet "Testdat" geladen werden. Dort befinden sich auch unter dem Namen "Adat.sav" die Daten zu dieser Analyse.

Literatur zu Bootstrap

Davison, A.C. & Hinkley, D.V.: Bootstrap methods and their application, 2006, 8th Edn. Cambridge University Press

Efron, Bradley, and Robert Tibshirani: An Introduction to the Bootstrap. Chapman and Hall/CRC, 1994.

Elias, Christopher J.: Percentile and Percentile-t Bootstrap Confidence Intervals: A Practical Comparison, Journal of Econometric Methods, 2013, Band 4, Heft 1, S 153-161

Wilcox, Rand R. Introduction to Robust Estimation and Hypothesis Testing. 4th edition. Academic Press, 2017.