Bootstrap bei Allgemeinem Linearen Modell

Allgemeines Lineares Modell III

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
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)
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

P20.25 Bootstrap bei Allgemeinem Linearen Modell

P20.25.1 Vorgehensweise

Aus einer vorliegenden Stichprobe (wir nennen sie "originale" Stichprobe) der Größe n werden zufällig n 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. So stellt sich etwa das Problem der Heteroskedastizität beim ALM nicht mehr. Siehe dazu Almo-Dokument 13 "Allgemeines lineares Modell", Abschnitt P20.6.8.1

Betrachten wir als Beispiel den Regressionskoeffizienten b1 für eine unabhängige quantitative Variable x1. Aus den 1000 Bootstrap-Stichproben erhalten wir 1000 Werte für b1. Wir berechnen deren Mittelwert und ihre Standardabweichung. Die Standardabweichung ist dann der "Standardfehler" von b1. Die obere und untere Grenze des Konfidenzintervalls für beispielsweise ein Konfidenzniveau von 95% erhalten wir sehr einfach in folgender Weise: Die 1000 b1-Werte werden der Größe nach (aufsteigend) sortiert. Vom maximalen b1-Wert werden absteigend 2,5% von 1000 also 25 Werte heruntergezählt. Der dort in Position 975 stehende b1-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 b1-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 Perzentil-Verfahren und optional das asymmetrische und symmetrische Perzentil-t -Verfahren.
Wird der Mittelwert aus den 1000 b1-Werten mit dem b1-Wert aus der Original-Stichprobe verglichen, so muss man in aller Regel eine kleine "Verzerrung" zur Kenntnis nehmen. Der

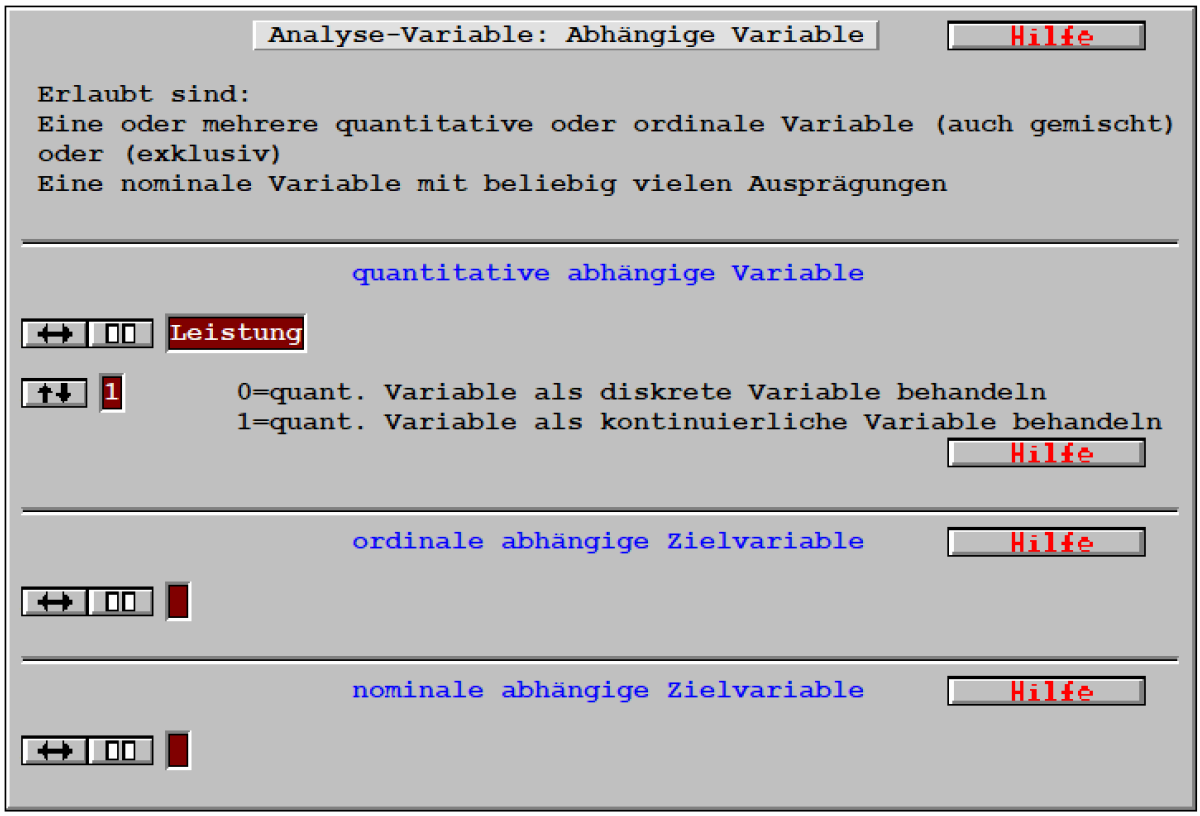
Mittelwert hat sonst keine Bedeutung. Als bester Schätzer für b1 wird der b1-Wert aus der Original-Stichprobe für den Forschungsbericht verwendet. Als seinen Standardfehler wird die aus dem Bootstrap gewonnene Standardabweichung eingesetzt, als seine Signifikanz p und sein Konfidenzintervall wird der aus dem Perzentil-Verfahren errechneten Werte eingesetzt. Alle diese Koeffizienten sind "parameterfrei".

Almo unterzieht nicht nur die Regressionskoeffizienten diesem Bootstrap-Kalkül sondern auch weitere Koeffizienten, die im Rahmen des ALM berechnet werden. Das wird in den nachfolgenden Abschnitten noch gezeigt.

P20.25.2 Die Almo-Eingabemaske

Das Bootstrap-Verfahren beim ALM ist in der Maske "Prog20my.Msk" realisiert. Man findet die Maske durch Klick auf den Knopf "Verfahren/Allgemeines lineares Modell" am Oberrand des Almo-Fensters. Die Maske stimmt weitgehend mit der ALM-Standard-Maske Prog20mo.Msk überein - so dass wir hier nur die Eingabemaske für die Analysevariablen und die Optionsbox "Bootstrap" erläutern werden, sowie die Eingabeboxen einiger Optionen, die in einer speziellen Weise auf das Bootstrapping einwirken.
Die Eingabefelder unserer Maske "Prog20my.Msk" sind für ein Beispiel ausgefüllt, das zuvor kurz vorgestellt werden soll. Die Daten für dieses Beispiel sind unter dem Namen "Adat.fre" im Ordner "Testdat" enthalten, ebenso die Datei der Variablen-Namen "Adat.nam". Die Daten sind konstruiert, also nicht empirisch gewonnen. So können die verschiedenen Möglichkeiten und Probleme gut an ihnen demonstriert werden Es wird folgende Kovarianzanalyse gerechnet:

Die abhängige Variable in unserem Beispiel ist die Leistung (in irgend einem Test)



Die unabhängigen Variablen sind

1. die Hauptdummies der 3 nominalen Variablen

   A Geschlecht: A1 männl, A2 weibl
   B Herkunft: B1 Unterschicht, B2 Mittelschicht, B3 Oberschicht
   C Wohnlage: C1 Land, C2 Stadtrand, C3 Stadt

2. die Interaktionsdummies der 2-er Interaktionen: AB, AC, BC
3. die Interaktionsdummies der 3-er Interaktion: ABC
4. die 3 Kovariaten: Alter, Bildungsniveau, Berufsqualifikation

Die jeweils letzte, redundante Dummy jeder nominalen Variablen wird gestrichen. Die wirksamen unabhängigen Variablen sind dann folgende

lfde. Nummer	Bezeichnung	Name
1	A1	männl
2	B1	Unterschicht
3	B2	Mittelschicht
4	C1	Land
5	C2	Stadtrand
6	A1 B1	
7	A1 B2	
8	A1 C1	
9	A1 C2	
10	B1 C1	
11	B1 C2	
12	B2 C1	
13	B2 C2	
14	A1 B1 C1	
15	A1 B1 C2	
16	A1 B2 C1	
17	A1 B2 C2	

```
18    V11        Alter
19    V1         Bildungsniv.
20    V2         Berufsqualif.
```

Die laufende Nummer "**lfde.Nummer**" der wirksamen, nicht-redundanten Variablen wird später im Zusammenhang mit dem Problem der Multkollinearität noch bedeutsam sein.

Interaktionen
Interaktionen höher als 2. Ordnung sind fast immer inhaltlich nicht interpretierbar. Sie erhöhen beträchtlich den Speicherbedarf und die Rechenzeit und verursachen oft Multikollinearitäten. Wir werden auf dieses Thema zurückkommen. Ihr einziger Vorteil ist es, dass sie die multiple Korrelation des Gesamtmodells vergrößern - dies jedoch durch die Einführung nicht interpretierbarer Variablen.

Problem: Ordinale Variable
Die Einbeziehung ordinaler Variable ist problematisch. Siehe dazu Abschnitt P20.6.9 im Almo-Dokument Nr. 13 "Allgemeines lineares Modell". Je nach Variablenkonstellation können der Speicherbedarf und die Rechenzeit extrem anwachsen. In der Literatur sind keine Analysen zu finden, bei denen ordinale Variable in das Bootstrap-Verfahren mit einbezogen wurden.

P20.25.2.1 Die Optionsboxen

Die in der Programm-Maske angebotenen Optionen leisten folgendes: (a) Sie manipulieren die Daten und (b) sie legen fest, welche Kalkül-Varianten anstelle des voreingestellten Kalküls gewählt werden sollen. Betrachten wir zunächst die Gruppe a der Optionen.

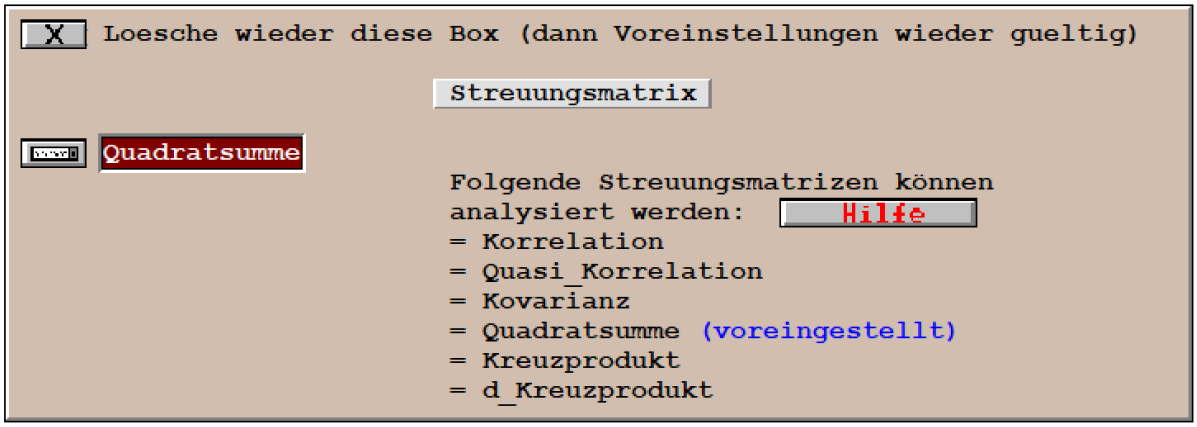
Optionen, die die Daten manipulieren
(1) Optionsbox *Ein- und Ausschluss von Untersuchungseinheiten.*
Entsprechend der Benutzer-Anweisungen werden Datensätze eliminiert bzw. nur unter bestimmten Bedingungen für die weitere Analyse beibehalten.
(2) Optionsbox *Umkodierungen und Kein-Wert-Angaben*
Umkodierungen von Variablen, Gleichungen, Anweisungen vom Typ "Wenn...Dann" usw. werden gemäß der Anweisung des Benutzers durchgeführt.
(3) Optionsbox *Ausreisser vom Typ 1 identifizieren.* Untersuchungsobjekte, die in Analysevariablen Werte einnehmen, die nicht mehr plausibel sind, werden identifiziert und ausgeschlossen oder in einer speziellen Weise behandelt.
(4) Optionsbox *Untersuchungseinheiten gewichten;* beispielsweise weil sie in manchen Variablen in den Stichprobendaten gegenüber der Grundgesamtheit unter- oder über-repräsentiert sind.

Diese Datenmanipulationen werden in der dargestellten Reihenfolge, eine nach der anderen, ausgeführt. Dies geschieht in einem vorausgehenden Schritt. Almo schreibt die veränderten Daten - aber nur die der Analysevariable - in eine spezielle Datei auf einem externen Speichermedium und in eine interne Datenmatrix. Aus ersterer werden dann die Daten für die originale Stichprobe entnommen. Die Daten für die x Bootstrap-Stichproben werden aus der internen Datenmatrix entnommen (wodurch die Rechengeschwindigkeit erheblich beschleunigt wird).

Diese vier Optionen werden nach Klick auf die Hilfeknöpfe in den jeweiligen Optionsboxen ausführlich erläutert. Auch im Dokument Nr. 0 "Arbeiten mit Almo" und im Dokument Nr.13 "Allgemeines Lineares Modell" im Abschnitt P20.8.1 werden sie ausführlich beschrieben.

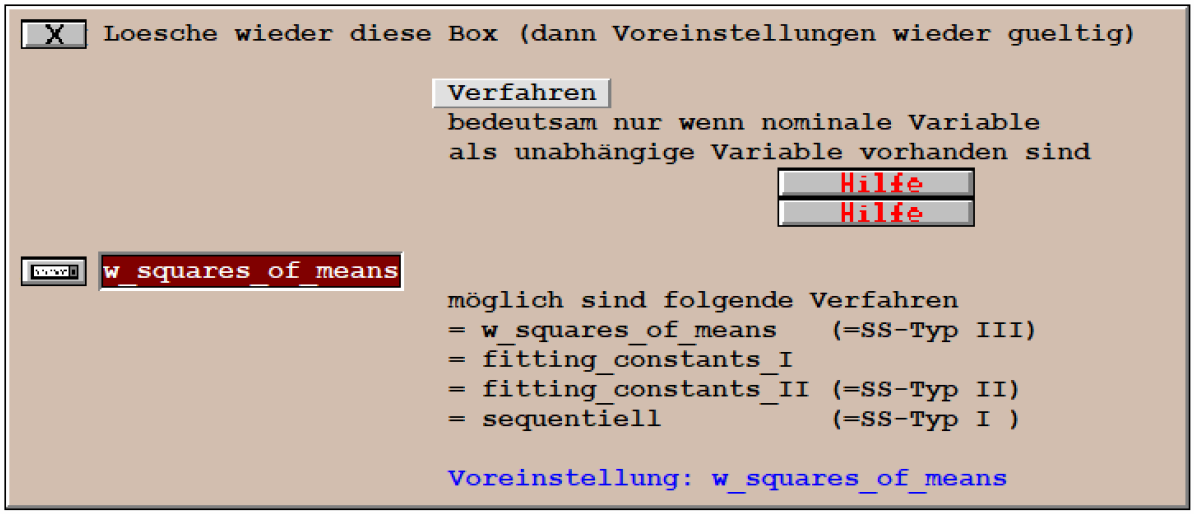
Optionen, die den Kalkül steuern
Die Optionsboxen werden in Almo-Dokument 13 "Allgemeines Lineares Modell", Abschnitt P20.8.1, ca. S. 98. ausführlich beschrieben

(1) Optionsbox *Streuungsmatrix*.



Standardmäßig verwendet Almo die Quadratsummenmatrix. Dabei tritt beim Boostrapping ein kleines Problem auf: Das Konfidenzintervall für die Konstante kann nur nach dem einfachen Perzentil-Verfahen aber nicht nach dem Perzentil-t -Verfahren berechnet werden. Wird die Quadratsummenmatrix verwendet, dann ermittelt Almo für die Konstante zwar deren Wert, aber nicht deren Standardfehler, der für das Perzentil-t -Verfahren gebraucht wird. Soll dieses aber doch eingesetzt werden, dann muss die "Kreuzprodukte"- oder besser die "d_Kreuzprodukte"-Matrix in der Optionsbox gewählt werden. Dann ermittelt Almo für die Konstante auch deren Standardfehler. "Kreuzprodukt" oder "d_Kreuzprodukt" sollten dann aber nur als *zusätzliche 2. Analyse* eingesetzt werden - eben nur um den Konstanten-Standardfehler zu gewinnen. Siehe Almo-Dokument 13, Abschnitt P20.8.1.1, ca. S. 98.

(2) Optionsbox Verfahren.



Standardmäßig verwendet Almo "w_squares_of_means". Es ist identisch mit SS-Typ III bei SPSS bzw. SAS. Ausnahmsweise eingesetzt werden können auch fitting constants I oder II. Dann dürfen jedoch keine Interaktionen angefordert werden. Die beiden Verfahren sind in diesem Fall identisch. Ist nur eine unabhängige nominale Variable vorhanden, dann sollte fitting constants sogar "w_squares_of_means" vorgezogen werden. Das sequentielle

Verfahren kann für Bootstrap nicht verwendet werden. Siehe Almo-Dokument 13, Abschnitt P20.7.1

(3) Optionsbox Nenner für Varianz, Kovarianz
(4) Optionsbox Behandlung eventueller Multikollinearität.
(5) Optionsbox Spezielle Programm-Optionen.

Die Optionsboxen 3, 4 und 5 werden in Almo-Dokument 13, Abschnitt P20.7.1 erläutert.

P20.25.2.2 Die Bootstrap-Eingabebox

Wird die Optionsbox "Bootstrap" geöffnet, dann sieht man folgendes

Eingabefeld 1: Bootstrap 1=ausführen, 0=nicht

Eingabefeld 2: Zahl der Bootstrap-Stichproben.
Empfohlen 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 nachfolgenden Abschnitt P20.25.9. Dort wird gezeigt welche Auswirkungen eintreten, wenn die Stichprobenzahl erhöht wird.

Eingabefeld 3: 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 danach die zusammengefassten Bootstrap-Ergebnisse für die Gesamtzahl der Stichproben ausgegeben.

Eingabefeld 4: Konfidenzniveau für Konfidenzintervall
Der Benutzer bestimmt das Konfidenzniveau. Üblich ist ein Niveau von 95%. Wir haben in Abschnitt P20.7.9.1 am Beispiel eines Regressionskoeffizienten bereits gezeigt, was dies bedeutet. 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 es 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 5: 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 P20.25.4 bis P20.25.6 ausführlich beschrieben.

Eingabefeld 6: 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 mit der Folge, dass aus der Originalstichprobe dieselben Probanden für die Bootstrap-Stichprobe ausgewält werden wie für die erste Analyse. Damit sind auch die Ergebnisse identisch.
Siehe dazu Abschnitt P20.25.9.

In Eingabefeld 7 und 8 wird das Problem der Multikollinearität beim Bootstrapping behandelt. Wir werden hier die beiden Eingabefelder nur kurz und oberflächlich definieren. In Abschnitt P20.7.9.3 werden wir ausführlich darstellen, wie Almo dieses Problem löst.

Eingabefeld 7: Behandlung von Multikollinearität in den Bootstrap-Stichproben
Der Forscher kann in das Eingabefeld diejenigen Variablen eintragen, die lineare Abhängigkeiten verursachen und von Almo eliminiert werden müssen. Almo schlägt vor, welche das sein sollten. Dann wird eine 2. Analyse gerechnet.

Eingabefeld 8: Wird 1 eingesetzt, dann werden die Nummern der Stichproben, in denen sich Multikollinearitäten ereignet haben, mitgeteilt.

Eingabefeld 9: Bootstrap für Wilks Lambda, Pillais Spur und Korrelation
1 = ja, berechnen und Ergebnis ausgeben
0 = nein, nicht berechnen
Siehe dazu Abschnitt P20.25.11

Eingabefeld 10: Ergebnisse aus Bootstrap für univariate Analysen für alle abhäng. Variablen
1 = ja, berechnen und Ergebnis ausgeben
0 = nein, nicht berechnen

P20.25.3 Die Bootstrap-Ergebnisse

Rechenzeit
Vergleich mit SPSS
Almo gibt zuerst die Ergebnisse für die Originalstichprobe aus. Hat der Benutzer im Eingabefeld 3 angefordert auch die Ergebnisse der z.B. ersten 3 Bootstrap-Stichproben auszugeben, so werden deren Ergebnisse eine nach der anderen präsentiert. Wir werden diese Ergebnisse hier nicht erläutern. Sie werden in Abschnitt P20.9 des Almo-Dokuments Nr. 13 detailliert interpretiert.

Im 2. Teil der Ergebnisliste werden die kumulierten Bootstrap-Ergebnisse vorgetragen, die in unserem Beispiel mit einer Fehlermeldung beginnen - die den Benutzer aber nicht erschrecken sollte.

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

                                                ********** FEHLER
```

```
In manchen Bootstrap-Stichproben entstand eine Datenkonstellation, bei
der unabhaengige Variable wegen linearer Abhaengigkeit eliminiert werden
    mussten. Die Folge davon ist, dass die Stichproben unterschiedliche
Zahlen und Konstellationen von unabhaengigen Variablen besitzen können
 Die Stichproben des Bootstrap-Verfahrens sind nicht mehr vergleichbar
```

Die Ausgabe beginnt mit einer Fehlermeldung. Es wird mitgeteilt, dass in einigen Bootstrap-Stichproben lineare Abhängigkeiten (=Multikollinearitäten) aufgetreten sind, die Almo beseitigt hat, indem es Variable, die dafür verantwortlich sind, aus der Analyse ausgeschlossen hat. Die *einzelnen* Stichproben werden also korrekt ausgewertet. Es gibt aber nun ein Problem. Es ist problematisch die 1000 Stichproben-Ergebnisse zu einem gemeinsamen Bootstrap-Ergebnis zu kumulieren, da sie teilweise auf verschiedenen Variablenkonstellationen beruhen, also nicht *vergleichbar* sind. Almo schlägt eine zweite veränderte Analyse vor, bei der die verantwortlichen Variablen aus allen 1000 Analysen ausgeschlossen sind.. Wir werden im folgenden Abschnitt P20.7.9.3 ausführlich darauf eingehen. Almo rechnet nicht automatisch diese zweite Analyse. Es überlässt dem Benutzer zu entscheiden, was geschehen soll. Es rechnet weiter und kumuliert ein abschließendes Bootstrap-Ergebnis. Das kann akzeptiert werden, wenn nur einige wenige unabhängige Variablen in nur einigen wenigen Bootstrap-Stichproben lineare Abhängigkeiten erzeugten und eliminiert wurden. In der letzten Spalte der Ergebnistabelle für das Bootstrapping listet Almo auf, wie viele Fälle von Multikollinearität je Variable aufgetreten sind. Wir zeigen einen Ausschnitt aus der Tabelle

	Regress.koeff/Effekte original	Bootstrap	Variable eliminiert in x Stichproben
Haupteffekte			
Interaktionseffekte			
A1 männl	0.373667	0.374458	–
A2 weibl	−0.373667	−0.374458	–
B1 Unterschic	−0.158575	−0.131731	–
B2 Mittelschi	0.293393	0.294919	–
B3 Oberschich	−0.134817	−0.163188	–
.	.	.	.
.	.	.	.
.	.	.	.
B2 C1	0.150083	0.154270	–
B2 C2	0.023437	0.020398	3
B2 C3	−0.173521	−0.174606	–
B3 C1	−0.092119	−0.092519	–
B3 C2	0.407962	0.431752	–
B3 C3	−0.315842	−0.339233	–
A1 B1 C1	0.458397	0.457822	–
A1 B1 C2	0.024182	0.027550	3
A1 B1 C3	−0.482579	−0.485290	–
A1 B2 C1	−0.281430	−0.280906	22
A1 B2 C2	0.044724	0.040872	244
A1 B2 C3	0.236706	0.249311	22
A1 B3 C1	−0.176968	−0.183097	–

Man erkennt, dass z.B. die 3-er Interaktion `A1B1C2` in nur `3` Stichproben und `A1B2C2` in `244` Stichproben eine lineare Abhängigkeit erzeugt hat und dadurch von Almo eliminiert werden musste. Für `A1B1C2` stehen also nicht `1000` Stichproben mit ihren Ergebnissen zur Verfügung sondern nur `997`, für `A1B2C2` kann Almo die Bootstrap-Ergebnisse nur aus `756` Stichproben kumulieren. 3 lineare Abhängigkeiten wäre zu akzeptieren, aber nicht 224. Es *muss* eine 2. Analyse mit Bootstrap gerechnet werden. Almo teilt mit, welche Variable der Benutzer dabei ausschließen muss.

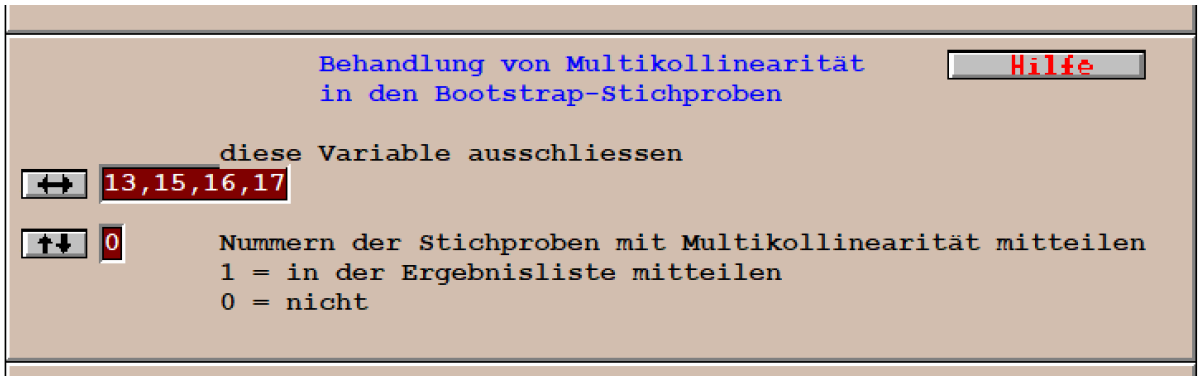
```
********** WARNUNG
In den Bootstrap-Stichproben wurden fuer
```

```
                                        folgende nicht-redundante Dummies und Kovariate
                                                        lineare Abhaengigkeiten entdeckt
                                                  13,15,16,17

        Dies sind die laufenden Nummern der nicht-redundanten Dummies und Kovariaten
           in der Reihenfolge der unabhaengigen Variablen, nicht die Variablennummer
-------------------------------------------------------------------------------------------
Rechnen Sie eine 2. Analyse, bei der in der Optionsbox für Bootstrap diese Nummern
eingetragen sind
-------------------------------------------------------------------------------------------
```

Die 2. Analyse

Wir folgen der Aufforderung von Almo und rechnen eine 2. Analyse, wobei wir jetzt nicht 1000 sondern 10 000 Stichproben rechnen. Die Rechenzeit wird dadurch um den Faktor 10 verlängert (auf unserem alten Computer auf 200 sec). Dadurch wird es möglich, das Konfidenzintervall präziser zu bestimmen. In der Optionsbox für das Bootstrap werden die oben angegebenen laufenden Nummern eingetragen



Als Ergebnis aus den 10 000 Bootstrap-Stichproben gibt Almo eine sehr breite Tabelle aus, die hier - um sie abbilden zu können - in zwei Teile zerschnitten werden muss. Im 1. Teil werden die aus dem Bootstrappping hervorgegangenen Effekte und Regressionskoeffizienten ausgegeben und im 2. Teil die Korrelationen.

P20.25.4 Die Ergebnistabelle Teil 1: Effekte und Regressionskoeffizienten

```
Einstellungen
-------------
Startzahl fuer Zufallsgenerator: 578125
Berechnung von Konfidenzintervall u. Signifikanz p bei univariater Analyse:
           (1) bei Effekte, Regr.koeffiz. und Konstante: einfaches Perzentil-Verfahren
           (2) bei paarweisen Mittelwertsvergleichen:    einfaches Perzentil-Verfahren
           (3) bei Eta-Korrelation und multiplem R:      einfaches Perzentil-Verfahren
Konfidenzniveau: 95%


Bootstrap-Ergebnisse fuer univariate Analyse aus 10 000 Stichproben fuer
abhaengige Variable:                                        V1 Leistung
================================================================================

                        Effekte / Regressionskoefizienten, Standardfehler
                     Signifikanz, optimales Konfidenzniveau, Konfidenzintervall

                                                                              *e
                                  *a        *b      *c       *d  Konfidenzintervall
                 Regress.koeff/Effekte  Standard  Signif  optimal       Konfniv=0.950
                   original  Bootstrap    fehler     p     Konfniv   unten       oben
-------------------------------------------------------------------------------------
Haupteffekte
Interaktionseffekte
-------------------
A1  männl          0.369082   0.371074   0.17999  0.0418  0.9582   0.013056   0.720791
```

A2	weibl	-0.369082	-0.371074	0.17999	0.0418	0.9582	-0.720791	-0.013056
B1	Unterschic	-0.141332	-0.123639	0.19977	0.5208	0.4792	-0.505368	0.276996
B2	Mittelschi	0.309515	0.316566	0.13498	0.0198	0.9802	0.051663	0.578614
B3	Oberschich	-0.168183	-0.192927	0.24954	0.4362	0.5638	-0.690430	0.287545
C1	Land	-0.059068	-0.067426	0.20757	0.7582	0.2418	-0.477625	0.327728
C2	Stadtrand	0.236925	0.239525	0.11660	0.0412	0.9588	0.011257	0.467974
C3	Stadt	-0.177857	-0.172099	0.22125	0.4332	0.5668	-0.612169	0.262323
A1 B1		0.022939	0.027294	0.14935	0.8542	0.1458	-0.262406	0.325269
A1 B2		-0.079384	-0.071080	0.14345	0.6047	0.3953	-0.347667	0.220457
A1 B3		0.056445	0.043779	0.21739	0.8330	0.1670	-0.386726	0.471195
A2 B1		-0.022939	-0.027294	0.14935	0.8542	0.1458	-0.325269	0.262406
A2 B2		0.079384	0.071080	0.14345	0.6047	0.3953	-0.220457	0.347667
.		.	.	.	.	.	.	.
.		.	.	.	.	.	.	.
.		.	.	.	.	.	.	.
A1 B1 C1		0.421090	0.432833	0.27043	0.1014	0.8986	-0.081953	0.963235
A1 B1 C2		-	-	-	-	-	-	-
A1 B1 C3		-0.421090	-0.432833	0.27043	0.1014	0.8986	-0.963235	0.081953
A1 B2 C1		-	-	-	-	-	-	-
A1 B2 C2		-	-	-	-	-	-	-
A1 B2 C3		-	-	-	-	-	-	-
A1 B3 C1		-0.421090	-0.432833	0.27043	0.1014	0.8986	-0.963235	0.081953
.		.	.	.	.	.	.	.
.		.	.	.	.	.	.	.
.		.	.	.	.	.	.	.
A2 B3 C3		-0.421090	-0.432833	0.27043	0.1014	0.8986	-0.963235	0.081953
nominale Variable und Interaktionen (Dummies zusammengefasst)								
V9 Geschlech		-	-	-	-	-	-	-
V5 Herkunft		-	-	-	-	-	-	-
V4 Wohnlage		-	-	-	-	-	-	-
V9*V5		-	-	-	-	-	-	-
V9*V4		-	-	-	-	-	-	-
V5*V4		-	-	-	-	-	-	-
V9*V5*V4		-	-	-	-	-	-	-
Kovariante und Konstante								
Alter		0.523529	0.524910	0.03837	0.0001	0.9999	0.450941	0.601617
Bildungsniveau		0.279852	0.280774	0.03893	0.0001	0.9999	0.205095	0.358734
Berufsqualifik		0.315820	0.317833	0.04135	0.0001	0.9999	0.237140	0.398653
Konstante		-19.85130	-19.93475	1.23127	0.0001	0.9999	-22.41889	-17.55386
multiple Korr. R		-	-	-	-	-	-	-

*x) Wird der p-Wert und das Konfidenzintervall der Konstanten gebraucht, dann muss in der Optionsbox "Streuungsmatrix" umgeschaltet werden auf "Kreuzprodukt" bzw. "d_Kreuzprodukt" oder es muss in der Bootstrap-Optionsbox auf das einfache Perzentil-Verfahren umgeschaltet werden

*a "original" bezeichnet den Wert aus der Originalstichprobe
mit "Bootstrap" wird der Mittelwert aus den Bootstrap-Stichproben bezeichnet

*b Der Standardfehler ist gleich der Standardabweichung der Werte aus den Bootstrap-Stichproben

*c Berechnet wird die zweiseitige Signifikanz p
Beim einfachen Perzentil-Verfahren ist sie gleich 1.0-"optimales Konfidenzniveau"

*d Optimales Konfidenzniveau. Entsteht nur bei einfachem Perzentil-Verfahren
Es erzeugt ein Konfidenzintervall, in dem der Wert 0 gerade nicht mehr enthalten ist. Er befindet sich in den aufsteigend sortierten Bootstrap-Werten gerade unterhalb der Intervall-Untergrenze bzw. gerade oberhalb der Intervall-Obergrenze

*e Vom Benutzer vorgegebenes Konfidenzintervall
Das Konfidenzniveau ist 95.00%. Beim "einfachen" Perzentil-Verfahren bedeutet das: Von den aufsteigend sortierten 10000 Werten aus den Bootstrap-Stichproben befinden sich 95.00% der Werte zwischen den Konfidenzgrenzen und je 2.50% oberhalb und unterhalb der Konfidenzgrenzen

*f Dummies und Kovariate koennen wegen linearer Abhaengigkeit in x Stichproben eliminiert worden sein. Die Bootsrap-Ergebnisse dieser Variablen beruhen so auf 10000 Stichproben minus der in der letzten Spalte der Tabelle angegebenen Zahl x

***g Dies ist die multiple partielle Korrelation der nicht-redundanten Dummies der nominalen Variablen bzw. der Interaktionsvariablen hinsichtlich der abhängigen Variablen**

Die Spalten der Tabelle 1.

1. die ***Effekte*** der Dummies bzw. die ***Regressionskoeffizienten*** der Kovariaten. Sie entstanden als Mittelwerte aus den entsprechenden Koeffizienten aus den 10 000 einzelnen Bootstrap-Stichproben. Sie stehen in der Spalte **`Bootstrap`**. In der 1.Spalte wird zum Vergleich der Koeffizient aus der Originalstichprobe angegeben. Die Differenz "**`Bootstrap minus Original`**" zwischen den beiden beträgt z.B. für die Hauptdummy `A1 männlich = 0.000791`. Dies ist die *Verzerrung*. Als bester Schätzer, somit als Endergebnis für den Forschungsbericht wird der Wert aus der Originalstichprobe verwendet.

Zum Begriff der ***Effekte*** und zum Vergleich mit den entsprechenden Parametern bei **SPSS** siehe Abschnitt P20.7.5 im Almo-Dokument Nr.13 zum ALM. Im Anhang "Kovarianzanalyse mit SPSS und Almo" zu diesem Dokument wird gezeigt, dass die Effekte in Almo identisch sind mit den Abweichungskontrasten in SPSS. In Almo wird auch für die *letzte*, eigentlich redundante Dummy ein Effekt berechnet. Die Effekte in Almo werden auf Summe 0 normiert. So ergibt sich der Effekt der letzten Dummy durch Subtraktion der Summe der *vorderen* Effekte von 0. In Abschnitt P20.7.5 des Dokuments Nr. 13 haben wir gezeigt, dass sich im einfachen Fall der Varianzanalyse mit gleichen Zellenhäufigkeiten der Effekt ergibt als *Abweichungskontrast* des Ausprägungsmittelwerts der Dummy-Variablen vom Gesamtmittelwert. Bei der Kovarianzanalyse ist dieser Sachverhalt nicht ganz so übersichtlich. Die Effekte in Almo bzw. die Abweichungskontraste bei SPSS werden noch an die Kovariaten angepasst.

2. der ***Standardfehler*** des Effekts bzw. des Regressionskoeffizienten ist gleich der Standardabweichung seiner Werte aus den 10 000 Bootstrapstichproben. Für die dummy-Variable `A1` (Geschlecht: männlich) ist der Standardfehler mit `0.187518` halb so groß wie der Effekt mit `0.373667`, so dass für diesen Effekt angenommen werden kann, dass er die abhängige Variable "Leistung" signifikant beeinflusst. Der aus den Bootstrapstichproben gewonnene „verteilungsfreie“ Standardfehler ist häufig größer als der „parametrische“ Standardfehler, der aus den Daten der Original-Stichprobe berechnet wurde.

3. ***Konfidenzintervall, "optimales Konfidenzniveau" , Signifikanz p***
Um die Signifikanz p eines Koeffizienten bestimmen zu können, müssen wir eine bestimmte Verteilungsfunktion unterstellen. In der Regel ist dies die Normalverteilung oder die t-Verteilung. Eine solche Unterstellung ist jedoch "wider den Geist des Bootstrappings", dessen Vorteil gerade die "Verteilungsfreiheit" ist. In Almo wird deswegen die Signifikanz nach dem "Perzentil"-Verfahren ermittelt, mit dessen Hilfe auch das Konfidenzintervall berechnet wird. Almo bietet als Alternative noch das "Perzentil-t"-Verfahren in einer symmetrischen und asymmetrischen Variante an. Die Bezeichnung „Perzentil_**t“** legt den Verdacht nahe, dass hier die t-Verteilung unterstellt werden muss. Das ist nicht der Fall, wie aus der Konstruktion dieses Verfahrens noch sichtbar werden wird. Wir werden die beiden Verfahren anschließend darstellen.

Die Zeilen der Tabelle 1.

Zuerst werden *Haupteffekte* ausgegeben, dann die *Interaktionseffekte*. Betrachten wir die Dummies der nominalen Variablen "Herkunft". Die dritte Dummy `B3` ist redundant. Sie wird im Regressionskalkül nicht verwendet. Ihren Effekt von **`-0.168183`** erhält sie residual aus

```
0-(B1+B2) = 0-(-0.141332+0.309515) = -0.168183
```

Anders formuliert: Die Effekte einer nominalen Variablen sind auf Summe .0 normiert. Das gilt allerdings nur, wenn das ALM mit dem Verfahren der "weigthed squares of means" (SS-Typ III) gerechnet wurde. Dies ist das Standardverfahren. Siehe Abschnitt P20.7.3.1.

Betrachtet man die Interaktionseffekte, so wird ersichtlich, dass besonders diejenigen hoher Ordnung, häufig - vielleicht sogar in aller Regel - Artefakte sind, d.h. inhaltlich nicht interpretierbar sind. So hat etwa der Interaktionseffekt 3. Ordnung **A1B1C1** einen Wert von 0.421090, also einen höheren Wert als irgend einer der Haupteffekte. Es wäe deswegen ohnehin sinnvoll gewesen, die Interaktionen 3. Ordnung nicht in das Modell aufzunehmen.

Danach folgen die *nominalen Variablen und ihre Interaktionen*. Sie bestehen aus den zusammengefassten Dummies. Für sie werden erst in Tabelle 2 Ergebnisse angezeigt.

Schließlich werden die *Kovariaten* und die *Konstante* ausgegeben. Für beispielsweise die Variable des Alters wird ein Mittelwert aus den 10 000 Stichproben von 0.524910 ausgegeben, der gegenüber dem Wert aus der originalen Stichprobe nur um 0.001381 verzerrt ist. Der Standardfehler ist mit 0.03837 sehr gering. Das 95%-ige Konfidenzintervall reicht von 0.450941 bis 0.601617.

P20.25.5 Das einfache Perzentil-Verfahren

Das einfache Perzentil-Verfahren liefert durch einen sehr einfachen und überschaubaren Kalkül die Signifikanz p (den p-Wert) und das Konfidenzintervall für die Effekte bzw. Regressionskoeffizienten und die Eta-Korrelationen der unabhängigen Variablen.

P20.25.5.1 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 Signifikanzniveau (von üblicherweise 95%) den Wert 0 einschließt. Ist das nicht der Fall, dann ist der Koeffizient signifikant. 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 rechnen die von Almo geforderte 2. Analyse. Dabei rechnen wir mit 10 000 Stichproben, um das Konfidenzintervall und die Signifikanz p mit dem "einfachen Perzentil-Verfahren" möglichst genau bestimmen zu können. Wir betrachten die Hauptdummy A1 männlich. Almo liefert diese aufsteigend sortierte Aufeinanderfolge der Effekte dieser Variablen aus den 10 000 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 originale Wert `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`

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.

P20.25.6 Das Perzentil-t -Verfahren

Wir verwenden folgende Notation:

```
b  = Regressionskoeffizient einer Kovariaten (oder Effekt einer Dummy) aus der
     originalen Stichprobe
S  = Standardfehler von b
b* = Regressionskoeffizient einer Kovariaten (oder Effekt einer Dummy aus den
     Bootstrap-Stichproben
S* = Standardfehler von b*
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.

P20.25.6.1 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

```
Bootsrap Stichproben ─────┐          ┌─────── t*-Werte für Variable A1
aufsteigend sortiert      │          │        aus 10 000 Bootstrap-Stichproben
                        1        -3.9209
                        .           .
                        .           .
                      250        -2.09882
-----------------------------------------
                      251        -2.09461 <--- ut*
                      252        -2.09191
 Konfidenzintervall     .           .
 der t* -Werte          .           .
                     9749         2.20537           .          .
                     9750         2.20650 <--- ot*
-----------------------------------------
```

```
 9751        2.20846
    .           .
    .           .
10000        4.11787
```

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.

P20.25.6.2 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.

P20.25.6.3 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".

P20.25.7 Die Ergebnistabelle Teil 2: Korrelationskoeffizienten

In Almo wird die Bootstrap-Ergebnistabelle in einem Stück ausgegeben. Wir werden hier nun den hinteren Teil dieser Tabelle abbilden und kommentieren

```
Bootstrap-Ergebnisse fuer univariate Analyse aus 10 000 Stichproben fuer
abhaengige Variable:                                   V1 Leistung
================================================================================

                |             partielle Korrelation Eta                |
                |         Standardfehler und Konfidenzintervall        |
                |                                                      |
                |                                            *e        |       *f
                |                 *a          *b   Konfidenzintervall  | Variable
                |    partielles Eta     Standard        Konfniv=0.950  | eliminiert in
                |  original  Bootstrap    fehler    unten       oben   | x Stichproben
----------------|------------------------------------------------------|-------------
Haupteffekte    |
Interaktionseffekte
----------------|
A1  männl       |  0.112904   0.110235   0.053843   0.003481   0.213784        -
A2  weibl       | -0.112904  -0.110235   0.053843  -0.213871  -0.003487        -
B1  Unterschic  | -0.035791  -0.030566   0.047909  -0.123873   0.062866        -
B2  Mittelschi  |  0.106499   0.105598   0.044670   0.016796   0.190963        -
B3  Oberschich  | -0.040751  -0.042827   0.055131  -0.147743   0.065973        -
C1  Land        | -0.016209  -0.016318   0.053184  -0.118089   0.086996        -
C2  Stadtrand   |  0.087170   0.086768   0.042380   0.004019   0.169087        -
C3  Stadt       | -0.044234  -0.041302   0.052204  -0.143945   0.058679        -
A1 B1           |  0.007333   0.008655   0.046019  -0.079955   0.099853        -
A1 B2           | -0.027786  -0.025153   0.048320  -0.120732   0.068820        1
A1 B3           |  0.015320   0.011963   0.055186  -0.096046   0.120568        -
A2 B1           | -0.007333  -0.008655   0.046019  -0.100107   0.079946        -
A2 B2           |  0.027786   0.025153   0.048320  -0.068829   0.120489        1
    .           |    .          .          .          .          .             .
    .           |    .          .          .          .          .             .
    .           |    .          .          .          .          .             .
A1 B1 C1        |  0.093473   0.089614   0.054460  -0.017483   0.193995        -
A1 B1 C2        |    -          -          -          -          -         10000 *h
A1 B1 C3        | -0.093473  -0.089614   0.054460  -0.194056   0.017424        -
A1 B2 C1        |    -          -          -          -          -         10000 *h
A1 B2 C2        |    -          -          -          -          -         10000 *h
A1 B2 C3        |    -          -          -          -          -         10000 *h
A1 B3 C1        | -0.093473  -0.089614   0.054460  -0.194056   0.017424        -
    .           |    .          .          .          .          .             .
    .           |    .          .          .          .          .             .
    .           |    .          .          .          .          .             .
A2 B3 C3        | -0.093473  -0.089614   0.054460  -0.194056   0.017424        -
................................................................................
                |
nominale Variab|  partielle multiple *g
```

und Interaktion Dummies zusamme	Eta-Korrelation original	Bootstrap			
V9 Geschlech	0.112904	0.110977	0.052295	0.012727	0.213597
V5 Herkunft	0.106555	0.118731	0.041410	0.039449	0.201442
V4 Wohnlage	0.087248	0.102794	0.039683	0.028331	0.182444
V9*V5	0.027791	0.064702	0.034496	0.011214	0.142045
V9*V4	0.017522	0.062956	0.032772	0.011426	0.135314
V5*V4	0.119105	0.139056	0.044940	0.053673	0.227650
V9*V5*V4	0.093473	0.091945	0.050424	0.006032	0.193671
Kovariante und Konstante	partielles Eta original	Bootstrap			
Alter	0.627383	0.628612	0.033246	0.561159	0.691198 -
Bildungsniveau	0.345871	0.346294	0.042909	0.260774	0.427224 -
Berufsqualifik	0.370950	0.372433	0.040548	0.288882	0.449759 -
Konstante	-	-	-	-	-
multiple Korr.	0.942498	0.944608	0.005134	0.934043	0.954085

*h Vom Benutzer eliminierte Variable

Im 2. Teil der Tabelle werden die Korrelationen der unabhängigen Variablen ausgegeben. Alle Korrelationen in dieser Tabelle sind nach dem PRE-Prinzip entstanden. Siehe Almo-Dokument 13, Abschnitt P20.6.3. Sie sind also auch partielle Koeffizienten. In der Varianz-Kovarianzanalyse werden diese Koeffizienten üblicherweise "Eta"-Korrelationen genannt.

In den Spalten der Tabelle werden ausgegeben: Die *partielle Eta-Korrelation* der Variablen aus der originalen Stichprobe und der *Mittelwert der Eta-Korrelationen* aus den 10 000 Bootstrap-Stichproben. Die Differenz zwischen den beiden (die *Verzerrung)* z.B. der Dummy A1 ist mit `0.00267` gering. Der *Standardfehler* der Eta-Korrelation ist gleich der Standardabweichung der Eta-Korrelationen aus den 10 000 Bootstrap-Stichproben. Die beiden Grenzwerte für das *Konfidenzintervall* werden immer durch das einfache Perzentil-Verfahren bestimmt. Der Benutzer hat keine Wahlmöglichkeiten. Die beiden Perzentil-t - Verfahren sind nicht verfügbar. In der letzten Spalte wird angezeigt, bei wie vielen Bootstrap-Stichproben die betreffende Variable eine lineare Abhängigkeit erzeugt hat und deswegen aus der Analyse eliminiert werden musste. So musste beispielsweise die Interaktionsdummy `A1B2` und `A2B2` bei nur einer Bootstrap-Stichproben ausgeschlossen werden. Das kann beruhigt negiert werden. Almo ermittelt für jede Variable die Zahl der Bootstrapstichproben, in denen sie einen Wert abgeliefert hat bzw. nicht abliefern konnte, da sie eliminiert war.

Die Zeilen der Tabelle wurden schon oben beschrieben. Nach den Korrelationen der Haupt- und Interaktionsdummies werden die Korrelationen der nominalen Variablen und Interaktionen angezeigt. Diese ergeben sich aus der Zusammenfassung ihrer jeweiligen Dummies und sind somit *partielle multiple Korrelationen.* Betrachten wir z.B. die nominale Variable der "`Herkunft`". Hier ein Ausschnitt aus der Tabelle

	*a partielles Eta original	Bootstrap	*b Standard fehler	*e Konfidenzintervall Konfniv=0.950 unten	oben	*f Variable eliminiert in x Stichproben
V9 Geschlech	0.112904	0.110977	0.052295	0.012727	0.213597	
V5 Herkunft	0.106555	0.118731	0.041410	0.039449	0.201442	
V4 Wohnlage	0.087248	0.102794	0.039683	0.028331	0.182444	
V9*V5	0.027791	0.064702	0.034496	0.011214	0.142045	
V9*V4	0.017522	0.062956	0.032772	0.011426	0.135314	
V5*V4	0.119105	0.139056	0.044940	0.053673	0.227650	
V9*V5*V4	0.093473	0.091945	0.050424	0.006032	0.193671	

Die nominale Variable `Herkunft` besteht aus den zwei nicht-redundanten Dummies `B1` und `B2`. Die dritte Dummy `B3` ist redundant. Sie wird im Kalkül nicht verwendet. Die durch `B1` und `B2` gemeinsam hinsichtlich der abhängigen Variablen erklärte Streuungen werden ermittelt. Sie werden in die PRE-Formel eingesetzt (siehe Almo-Dokument 13, P20.6.3). Es entsteht die *partielle multiple Korrelation* für die Variable `Herkunft`. Das geschieht in allen 10 000 Bootstrap-Stichproben und wird dann abschließend zum Wert `0.118731` gemittelt.

P20.25.8 Mittelwertsvergleiche

Nach den großen Tabellen der Effekte/Regressionskoeffizienten und der Korrelationen gibt Almo noch eine Tabelle der Mittelwertsdistanzen zwischen den Ausprägungen der nominalen Variablen und ihrer Signifikanzen aus - sofern der Benutzer diese in der Eingabemaske angefordert hat. Es geht beispielsweise um die wichtige Frage "unterscheiden sich `A1` Männer und `A2` Frauen signifikant in ihrer Leistung (in einem Test) ?" Oder: "Wie unterscheiden sich Menschen verschiedener sozialer Herkunft in ihrer Test-Leistung - und sind diese Unterschiede signifikant ?". Die Almo-Ausgabe ist folgende

```
================================================================================
Ergebnisse aus Bootstrap mit 10 000 Stichproben fuer paarweise Mittelwertsvergleiche
hinsichtlich der abhaengigen Variablen:           V6 Leistung
================================================================================

Fuer die Vergleichspaare werden folgende Bezeichnungen und Namen verwendet

  A  V9    Geschlecht
           A1  männl
           A2  weibl

  B  V4    Wohnlage
           B1  Land
           B2  Stadtrand
           B3  Stadt

  C  V5    Herkunft
           C1  Unterschicht
           C2  Mittelschicht
           C3  Oberschicht

Bootstrap: Paarweise Mittelwertsvergleiche fuer Variable A   V9 Geschlecht

                          *a         *b           *c                  *e   Variable   *f
            Mittellwertsdistanz   Standard-  Signifikanz  Konfidenzinzervall   eliminiert in
            original  Bootstrap   fehler         p        unten         oben   x Stichproben
-------------------------------------------------------------------------------------------
A1 - A2     0.738164   0.742149   0.359979    0.041800    0.026112   1.441581       -
-------------------------------------------------------------------------------------------

Bootstrap: Paarweise Mittelwertsvergleiche fuer Variable B   V4 Wohnlage

                          *a         *b           *c                  *e   Variable   *f
            Mittellwertsdistanz   Standard-  Signifikanz  Konfidenzinzervall   eliminiert in
            original  Bootstrap   fehler         p        unten         oben   x Stichproben
-------------------------------------------------------------------------------------------
B1 - B2    -0.295993  -0.306951   0.253791    0.222800   -0.817651   0.177970       -
B1 - B3     0.118789   0.104673   0.412897    0.791200   -0.710507   0.897521       -
B2 - B3     0.414782   0.411624   0.286376    0.147600   -0.143109   0.976911       -
-------------------------------------------------------------------------------------------

Bootstrap: Paarweise Mittelwertsvergleiche fuer Variable C   V5 Herkunft

                          *a         *b           *c                  *e   Variable   *f
            Mittellwertsdistanz   Standard-  Signifikanz  Konfidenzinzervall   eliminiert in
            original  Bootstrap   fehler         p        unten         oben   x Stichproben
-------------------------------------------------------------------------------------------
C1 - C2    -0.450848  -0.440205   0.232345    0.057200   -0.894306   0.015418       -
C1 - C3     0.026851   0.069288   0.431429    0.881400   -0.759196   0.935502       -
C2 - C3     0.477698   0.509494   0.347955    0.141400   -0.154591   1.189964       -
-------------------------------------------------------------------------------------------
```

Nur Männer und Frauen unterscheiden sich mit p=**0.0418** signifikant von einander in ihrer Leistung. Im Konfidenzintervall von **0.026112** bis **1.441581** ist die 0 nicht eingeschlossen. Bei allen anderen Vergleichen ist dies jedoch der Fall und der p-Wert ist immer größer als 0.05.

P20.25.9 Die Zahl der Bootstrap-Stichproben

Wie ändern sich die Ergebnisse des Bootstrappings wenn die Stichprobenzahl verändert wird. Wir haben dies für die Dummy A1 männlich getan und dabei die zweiseitige Signifikanz p und das Konfidenzintervall des Effekt dieser Variablen untersucht. Gerechnet wurde das einfache Perzentil-Verfahren.

Stich-proben-zahl	Signif p	optimal Konfniv	Konfidenzintervall Konfniv=0.950 unten	oben
100	0.020	0.980	0.096652	0.613246
500	0.048	0.952	0.013475	0.667787
1000	0.046	0.954	0.013475	0.711313
2000	0.048	0.952	0.005291	0.731829
5000	0.046	0.954	0.007103	0.720098
10000	0.042	0.958	0.013056	0.720791
15000	0.041	0.959	0.015180	0.724802
20000	0.042	0.958	0.014530	0.722788

Die Startzahl für den Zufallsgenerator war immer dieselbe. Das wirkt sich so aus dass z.B. bei Stichprobenzahl 2000 die ersten 1000 Stichproben identisch mit denen bei Stichprobenzahl 1000 sind. Die Werte-Änderung bei Stichprobenzahl 2000 im Vergleich zu 1000 Stichproben ist also dadurch entstanden, dass um 1000 Stichproben erweitert wurde.
Ab 1000 Stichproben sind die Unterschiede gering.

In nachfolgender Tabelle wird gezeigt, dass bei Änderung der Startzahl des Zufallsgenerators die Ergebnisse nur zufällig verschieden sind - was nicht anders zu erwarten war. Die größte Differenz beträgt rund 0.009.

10 000 Stichproben mit verschiedener Startzahl für Zufallsgenerator

Signif p	optimal Konfniv	Konfidenzintervall Konfniv=0.950 unten	oben
0.0418	0.9582	0.013056	0.720791
0.0444	0.9556	0.008719	0.714514
0.0456	0.9544	0.009512	0.716102
0.0420	0.9580	0.012772	0.715375
0.0426	0.9574	0.004243	0.718730

Mess-"Feinheit"
Wie wirkt sich eine Erhöhung der Stichprobenzahl auf den Kalkül für die Signifikanz und das Konfidenzintervall aus. Wir vergleichen die aufsteigend sortierten Effekte der Variablen A1 aus 1000 und 10 000 Stichproben

1000 Stichproben		10 000 Stichproben		
1	−0.2598930	1	−0.3675820	
2	−0.1551420	2	−0.2670360	
.	.	.	.	
.	.	.	.	
23	−0.0010547	209	−0.0000428	
				<--- 0.0
24	0.0033431	210	0.0001735	
		.	.	

```
                                .       .
      25    0.0052912           250    0.0126888
      --------------            ----------------
      26    0.0134753           251    0.0130560  <--- untere Konfidenzgrenze
      .      .                    .      .
      .      .                    .      .
    1000    0.7319800         10000    0.7359790
```

Die Werte aus den beiden Analysen überstreichen einen Bereich von grob -0.26 bis 0.73. Bei 10 000 Stichproben kann der 1. Wert mit **-0.3675820** als Ausreisser betrachtwt werden. Bei der Analyse mit 1000 Stichproben ist dieser Bereich durch 1000 Werte unterteilt, bei der Analyse mit 10 000 Stichproben durch 10 000 Werte also *"feiner"*. Der Wert 0 wird bei 1000 Stichproben zwischen Position 23 und 24 überschritten. Dem würde bei 10 000 Stichproben die Positionen von 230 bis 240 entsprechen. Tatsächlich wird er jedoch zwischen Position 209 und 210 überschritten. Durch die Position von 0 wird das optimale Konfidenzniveau und die Signifikanz p festgelegt. Die Signifikanz beträgt für 1000 Stichproben 0.046 und für 10 000 Stichproben 0.042. Der Unterschied von 0.004 ist allerdings vernachlässigbar. Aber wir können behaupten, er ist messgenauer.

P20.25.10 Behandlung von Multikollinearität in den Bootstrap-Stichproben

In manchen Bootstrap-Stichproben kann, bedingt durch die Zufallsauswahl der Probanden, eine Datenkonstellation entstehen, bei der unabhängige Variable wegen linearer Abhängigkeit eliminiert werden müssen, damit Almo weiter rechnen kann. Dieser Fall tritt hauptsächlich auf, wenn bei Varianz-, Kovarianz-Analysen Interaktionen höherer Ordnung mit eingeschlossen werden. Er tritt fast immer auf, wenn Interaktionen 3. oder noch höherer Ordnung angefordert werden. Werden bestimmte Variable in *allen* Stichproben eliminiert, dann ist das kein Problem. Werden sie nur in *einigen* eliminiert, dann hat das zur Folge, dass die Stichproben unterschiedliche Variablen-Konstellationen besitzen. Die Bootstrapstichproben sind dann nicht mehr korrekt vergleichbar. Almo meldet daraufhin einen Fehler - rechnet aber weiter und weist den Benutzer in mehrfacher Weise auf diese Situation hin. Tritt der Fall der Multikollinearität nur bei sehr wenigen Stichproben auf, dann wirkt sich dies beim schlussendlichen Kumulieren der Ergebnisse nur an hinteren Dezimalstellen aus und kann negiert werden.

Wird im Eingabefeld 7 der Bootstrap-Optionsbox eine 1 eingetragen, dann wird in der Ergebnisliste von Prog20my dem Benutzer mitgeteilt, in welcher Bootstrap-Stichprobe welche Variablen lineare Abhängigkeiten verursachten und eliminiert werden mussten, damit weiter gerechnet werden konnte

```
Variable wegen linearer Abhaengigkeiten eliminiert in Bootstrap-Stichprobe
-------------------------------------------------------------------------
Bootstrap-   lfde.Nummer       Bezeichnung
Stichprobe   der unabhaeng.    der unabhaeng.
   Nr.       Variablen         Variablen
----------   ----------        ----------
      5           17            A1 B2 C2
      6           17            A1 B2 C2
      8           17            A1 B2 C2
     23           17            A1 B2 C2
     25           17            A1 B2 C2
     28           17            A1 B2 C2
     35           13            B2 C2
     35           17            A1 B2 C2
     46           17            A1 B2 C2
      .            .               .
      .            .               .
      .            .               .
```

Daran anschließend wird noch die Warnung gebracht:

```
                                  ********** WARNUNG
          Zahl der Bootstrap-Stichproben, in denen für eine oder
        mehrere Variable lineare Abhaengigkeiten gefunden wurden
      und diese Variable demzufolge eliminiert werden mussten:  244
```

In Bootstrap-Stichprobe 5 musste die Interaktionsdummy 3. Ordnung `A1B2C2` eliminiert werden. In der Hintereinander-Reihung der unabhängigen Variablen ist sie die 17. In der Stichprobe 35 wurden 2 dummy Variable, die Interaktionsdummy 2. Ordnung `B2C2` und 3. Ordnung `A1B2C2` eliminiert. Insgesamt gab es 272 Eliminierungen. In der Warnung wird mitgeteilt diese 272 in 244 Stichproben sich ereignen. Ganz überwiegend waren das Interaktionsdummies 3. Ordnung. Die Interaktionsdummy 2. Ordnung `B2C2` musste nur in 3 Stichproben ausgeschlossen werden.

Hier wird ersichtlich, dass das Problem der Multikollinearität in den meisten Fällen durch den Forscher selbst leicht zu beheben ist. In unserem Beispiel hätte es genügt in der Eingabebox **Analyse-Variable: Unabhängige Variable** nur Interaktionen 2. Ordnung anzufordern - und das Problem wäre gelöst gewesen. Drei Eliminierungen von `B2C2` hätte man verschmerzen können. Es ist ohnehin empfehlenswert nur Interaktionen zu akzeptieren, die inhaltlich interpretierbar sind. Für Interaktionen 3. Ordnung und höher ist das kaum mehr möglich. Die einzige (eigentlich illegitime) "Lebensberechtigung" für Interaktionen solch hoher Ordnung ist, dass sie die Fehlerstreuung des Modells reduzieren

Almo liefert dem Forscher auch noch weitere Informationen zum Thema der Multikollinearität. In der Tabelle zu den Effekten bzw. Regressionskoeffizienten der Variablen wird eine letzte Spalte angehängt, in der zu jeder Variablen angegeben wird, ob und wie häufig sie eliminiert werden musste. Wir zeigen hier einen stark reduzierten Ausschnitt

```
                                Korrelation            Variable****
                                Konfidenzintervall**   eliminiert in
                                unten        oben      x Stichproben
 ____________________   .....  ____________________________________
 Haupteffekte
 Interaktionseffekte
 --------------------
 A1                     ..... -0.003183   0.202510        -
 A2                     ..... -0.204525   0.001851        -
 B1                     ..... -0.116775   0.053876        -
  .                     .....     .          .
  .                     .....     .          .
 B2 C2                  ..... -0.085432   0.093854        3
  .                     .....     .          .
  .                     .....     .          .
 A1 B2 C1               ..... -0.154182   0.042020       22
 A1 B2 C2               ..... -0.077285   0.093348      244
 A1 B2 C3                     -0.046739   0.136007       22
 A1 B3 C1                     -0.142801   0.080889        -
```

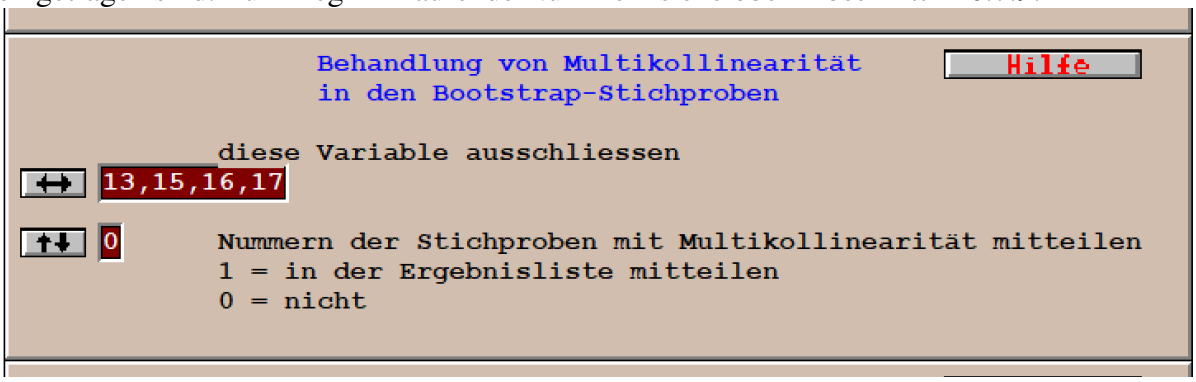
Man erkennt, dass die Dummy `B2C2` in 3 Stichproben und die Dummy `A1B2C2` in 244 Stichproben eliminiert wurde. Will der Benutzer auch noch wissen, in welchen Stichproben die jeweilige Variable ausgeschlossen wurde, dann muss er im nachfolgend abgebildeten Ausschnitt im 2. Eingabefeld eine 1 einsetzen.

Almo meldet dann noch folgende Warnung und empfiehlt folgende Vorgehensweise:

```
                                  ********** WARNUNG
                In den Bootstrap-Stichproben wurden fuer
          folgende nicht-redundante Dummies und Kovariate
                        lineare Abhaengigkeiten entdeckt
              13,15,16,17
```

```
        Dies sind die laufenden Nummern der nicht-redundanten Dummies und Kovariaten
           in der Reihenfolge der unabhaengigen Variablen, nicht die Variablennummer
-------------------------------------------------------------------------------
Rechnen Sie eine 2. Analyse, bei der in der Optionsbox für Bootstrap diese Nummern
eingetragen sind
-------------------------------------------------------------------------------
```

Der Forscher soll also eine 2. Analyse rechnen, bei der keine Eingabe geändert wird, jedoch die vier Nummern **13,15,16,17** in das Eingabefeld 6 eingetragen werden sollten. Hier ist ein Ausschnitt aus der Optionsbox für das Bootstrapping, in dem diese 4 Nummern eingetragen sind. Zum Begriff "laufende Nummer" siehe oben Abschnitt P20.7.9.2



Almo rechnet dann eine Analyse, in der diese Variable nicht enthalten sind und somit keine Variable eliminiert werden muss und dadurch die Bootstrap-Stichproben vergleichbar sind und ihre Ergebnisse korrekt kumuliert werden können.

P20.25.11 Bootstrap bei multivariater Analyse

Zuerst sollen die in Almo zentralen Koeffizienten der multivariaten Analyse kurz erläutert werden. Dies sind *Wilks Lambda* und *Korrelation*, *Pillais Spur* und *Pillais Korrelation*. Siehe dazu auch die ausführliche Darstellung in Almo-Dokument Nr. 13a "Allgemeines lineares Modell II", Abschnitt P20.9.4

Notation

Wir verwenden folgende Notation, die sich von der im Almo-Dokument geringfügig unterscheidet:

```
m = Gesamtzahl der unabhängigen Variablen, Kovariate und/oder Dummies
z = Zahl der Untermenge x von unabhängigen Variablen, deren Erklärungs-
    fähigkeit ermittelt werden soll. Ist x eine einzelne Variable dann ist
    z=1
w = Zahl der abhängigen Variablen
Q = das ist die w*w Matrix der Streuungen der w abhängigen Variablen
    In der Diagonale stehen die Streuungen, außerhalb der Diagonalen die
    Ko-Streuungen zwischen den w abhängigen Variablen
V = das ist die w*w Matrix der Streuungen und Ko-Streuungen zwischen den w
    abhängigen Variablen, die durch die m unabhängigen Variablen erklärt
    wird
W = das ist die w*w Matrix der Fehlerstreuungen, die in den w abhängigen
    Variablen verbleibt, nachdem die m unabhängigen Variablen eingeführt
    wurden und Erklärung leisten konnten
L = Wilks Lambda
P = Pillais Spur
```

Als "Streuung" wird im ALM die Abweichung-Quadratsumme, gelegentlich auch die Varianz/Kovarianz verwendet.

P20.25.11.1 Wilks Lambda und Korrelation

Es gilt:

(1) **Q** = **V** + **W**

Es werden die zwei Determinanten det(**Q**) und det(**W**) gebildet. Dies sind nun Skalare und nicht mehr Matrizen. Sie werden als "generalisierte Streuungen" bezeichnet.

```
det(Q) = das ist die generalisierte Gesamtstreuung der w abhängigen
         Variablen
det(W) = das ist die generalisierte Fehlerstreuung der w abhängigen
         Variablen nachdem m erklärende Variable eingeführt wurden
```

Das Wilk'sche Lambda L für das Gesamtmodell ist dann

(2) L = det(**W**)/det(**Q**)

Wir betrachten nun eine einzelne unabhängige Variable x. Es wird eine zweite Analyse gerechnet. Dabei wird aus den m unabhängigen Variablen die einzelne Variable x herausgenommen. x könnte auch eine Untermenge aus der Gesamtmenge der unabhängigen Variablen sein, z.B. Kovariate, die in einer allgemeinen Gesundheitsstudie den besonderen psychischen Gesundheitszustand eines Probanden beschreiben. Es könnten auch die Dummies einer unabhängigen nominal-polytomen Variablen sein.

Es entstehen die Matrizen **W~x** und **V~x**. Mit dem angehängten Symbol **~x** soll ausgedrückt werden, dass x aus der Menge der unabhängigen Variablen herausgenommen wurde.

```
W~x = das ist die Matrix der Fehlerstreuungen, die in den w abhängigen
      Variablen verbleibt nachdem aus den m unabhängigen Variablen die
      einzelne Variable x (oder die Untermenge x) herausgenommen wurde.
      Die Fehlerstreuungswerte in W~x sind in der Regel größer als die in
W,
      da weniger unabhängige Variable Erklärung leisten durften.
```

Es gilt:

(3) **V~x** = **W~x – W**

```
V~x = ist die w*w Matrix der durch die einzelne Variable x in den w
      abhängigen Variablen erklärte Streuung. Sie entsteht aus der
      Differenz zweier Fehlerstreuungsmatrizen aus zwei aufeinander
      folgenden Analysen
```

Wilks Lambda L(x) für eine einzelne unabhängige Variable x entsteht dann aus den Determinanten der beiden Fehlerstreuungsmatrizen. Die Determinanten werden auch als "generalisierte Streuung" bezeichnet.

(4) L(x) = det(**W**) / det(**W~x**)

Wilks Lambda drückt die Relation zweier generalisierter Streuungen aus. Es bewegt sich zwischen 0 und 1. Es ist 1, wenn x in den w abhängigen Variablen nichts erklärt. **W~x** ist dann gleich **W**. Lambda verhält sich also wie ein "umgedrehter" Korrelationskoeffizient.

Wilks Lambda L bzw. L(x) erfüllt einen doppelten Zweck:

(1) Aus Wilks Lambda kann ein Korrelationskoeffizient abgeleitet werden, wodurch die Wirkungsstärke der unabhängigen Variablen eingeschätzt werden kann und (2) es kann ein F-Wert entwickelt werden, der es ermöglicht zu prüfen, ob die unabhängigen Variablen *signifikant* auf die abhängigen Variablen einwirken.

1. Es kann ein F-Wert entwickelt werden. Dadurch wird es möglich die Signifikanz von `L` bzw `L(x)` zu prüfen.

```
L, L(x)  Wilks Lambda
F        F-Wert
m,z,w    siehe oben
df1,df2  Freiheitsgrade für F
N        Stichprobengröße

(4a) F = (1.0-L**(1.0/s))*df2 / (L**(1.0/s)*df1)
Wird der F-Wert für eine einzelne unabhängige Variable x bzw.
eine Untermenge x gebraucht, dann ist in Formel 4a L durch L(x)
zu ersetzten

**  = bedeutet "potenzieren"
df1 = w*z
df2 = f*s-k

s   = sqrt( (w²*z²-4) / (w²+z²-5) )     ist w*z=2 dann ist s=1
k   = (w*z-2)/2
f   = N-1-m-(w-z+1)/2

Soll der F-Wert für das Gesamtmodell ermittelt werden, dann ist
z=m zu setzen.
```

   Aus Wilks Lambda kann, wie wir anschließend zeigen werden, ein Korrelationskoeffizient abgeleitet werden, der wie gewohnt vom Forscher interpretiert werden kann . Wilks Lambda als selbständiger Koeffizient liefert dem Forscher nur dann wertvolle Information, wenn es in einen F-Wert überführt wird und darüber informiert, wie signifikant die unabhängigen Variablen auf die abhängigen einwirken.

2. Es kann, wie bereits angedeutet, ein Korrelationskoeffizient entwickelt werden, der die Wirkungsstärke der unabhängigen Variablen x ausdrückt. Die Formel für die (quadrierte) Korrelation nach Wilks ist

```
(5)  rw²    = 1.0 - L**(1/t)
(6)  rw²(x) = 1.0 - L(x)**(1/t)

rw²    = multiple quadrierte Korrelation nach Wilks für die Gesamtmenge
         der unabhängigen Variablen
rw²(x) = quadrierte Korrelation von x mit den w abhängigen Variablen -
t      = ist die kleinere Zahl von w oder m, bzw. w oder z
**     = dieses Symbol bedeutet "potenzieren".
```

   `L**(1/t)` bedeutet also: Das Wilks'sche Lambda wird mit 1/t potenziert. Entsprechend `L(x)**(1/t)`

   Ist x eine einzelne Variable, dann ist `1/t = 1` und Wilks quadrierte Korrelation `rw²(x) = 1.0-L(x)`

P20.25.11.2 Pillais Spur und Korrelation

Zu Wilks Lambda gibt es mehrere Alternativen. Bedeutsam ist "Pillais Spur" P und abgeleitet aus ihr "Pillais Korrelation". In Almo kann dann noch über eine Option die Hotelling-Lawley-Spur errechnet werden. Für Pillais Spur gilt:

```
(7)  P    =  Spur( V * inv(W) )          für das Gesamtmodell
(8)  P(x) =  Spur( V~x * inv(W~x) )      für eine einzelne unabhängige
                                         Variable x oder eine Untermenge

P      = Pillais Spur für die Gesamtmenge der unabhängigen Variablen
P(x)   = Pillais Spur für eine einzelne Variable x oder eine Untermenge x
         von unabhängigen Variablen
inv(W) = das ist die Inverse der Matrix W (bzw W~x
Spur   = das ist die Summe der Eigenwerte einer Matrix, hier der Matrix
         V*inv(W) bzw. V~x*inv(W~x)
```

Auch Pillais Spur kann zu einem F-Wert transformiert werden und dadurch auf seine Signifikanz geprüft werden. Siehe dazu Almo-Dokument 13a, Abschnitt P20.9.4.
Pillais quadrierte Korrelation ist

```
(9)   rp²    =  P/t           für das Gesamtmodell
(10)  rp²(x) =  P(x)/t        für eine einzelne unabhängige Variable x oder
                              eine Untermenge

rp²    = multiple quadrierte Korrelation nach Pillai für die Gesamtmenge
         der unabhängigen Variablen
rp²(x) = quadrierte Korrelation von x mit den w abhängigen Variablen –
t      = ist die kleinere Zahl von w oder m.
```

Ist x eine einzelne Variable aus einer Menge von mehreren unabhängigen Variablen, dann ist Pillais Spur gleich dem "umgedrehten" Wilks'schem Lambda und Pillais Korrelation gleich der Wilks'schen Korrelation. Sie sind dann "partielle" quadrierte Korrelationskoeffizienten.

```
(11)  P(x)   = 1.0 - L(x)
(12)  rp²(x) = rw²(x)
```

Ist x eine Untermenge von unabhängigen Variablen, dann sind sie verschieden, wenn auch nur geringfügig.

```
(13)  P(x)   != 1.0 - L(x)
(14)  rp²(x) != rw²(x)

Das Symbol != bedeutet "ungleich".
```

Diese Ungleichheit erklärt sich, wenn man eine "kanonische Korrelationsanalyse" rechnet. Pillais Spur ist dann die Summe der Eigenwerte der kanonischen Faktoren. Wilks Lambda wird im Rahmen der kanonischen Korrelationsanalyse für jeden einzelnen kanonischen Faktor berechnet. Wilks Lambda, wie wir es aus der multivariaten Analyse im Rahmen des Allgemeinen Linearen Modells erhalten, ist identisch mit Wilks Lambda für den 1. kanonischen Faktor, wie wir es im Rahmen der kanonischen Korrelationsanalyse (Prog29m1) erhalten. Die anderen kanonischen Faktoren werden nicht berücksichtigt. Das ist der Grund, warum in Almo der auf Pillais Spur beruhende Korrelationskoeffizient vorgezogen wird. Pillais Korrelation berücksichtigt alle kanonischen Faktoren. Hieraus folgt auch, dass die Korrelationskoeffizienten berechnet aus Pillais Spur und Wilks Lambda gleich sind, wenn nur ein kanonischer Faktor existiert. Und das ist der Fall wenn nur eine unabhängige quantitative oder auch eine in mehrere Dummies aufgelöste nominale Variable vorhanden ist. D.h. bei der einen nominalen Variablen spielt es dabei keine Rolle wie viele Ausprägungen sie besitzt.

P20.25.12 Ergebnisse aus Bootstrap bei multivariater Analyse

Wir rechnen die multivariate Analyse mit 2 abhängigen Variablen, `V6 Leistung` und `V12 Stressbelastung`. Almo ermittelt zuerst die univariaten Ergebnisse für `Leistung` und `Stressbelastung` und erst danach die multivariaten Ergebnisse für folgende Koeffizienten

Wilks Lambda,
Pillais Spur
Pillais Korrelation.

Als Ergebnis des Bootstrapping entstehen für dieser drei Koeffizienten
die Mittelwerte aus den Bootstrap-Stichproben
deren Standardfehler und
die Konfidenzintervalle.

Diese Koeffizienten werden für alle einzelnen unabhängigen Variablen, d.h. für die Kovariaten und die Dummies der nominalen Variablen und auch für das Gesamtmodell ermittelt.

Die Frage, ob durch Bootstrap zusätzlich zu den Ergebnissen aus den Originaldaten wertvolle Informationen gewonnen werden, kann nicht generell beantwortet werden.

Der Standardfehler aus den Bootstrap-Stichproben für die verschiedenen Koeffizienten wird im Rahmen des ALM nicht weiter gebraucht. Er informiert den Forscher darüber, wie stark die Mittelwerte der Bootstrap-Stichproben streuen.

Aus Wilks Lambda in den Originaldaten kann, wie wir oben gezeigt haben, ein Korrelationskoeffizient abgeleitet werden, der wie gewohnt vom Forscher interpretiert werden kann. Almo zieht es allerdings vor, den Korrelationskoeffizienten nach Pillai zu errechnen. Siehe die nachfolgende Tabelle 2. Pillais Korrelation und Wilks Korrelation sind für einzelne unabhängige Variable (Kovariate oder Dummies) identisch. Sie divergieren nur bei der multiplen Korrelation für das Gesamtmodell. In diesem Falle wird Pillais Korrelation vorgezogen, da sie (quadriert) sich, als „durchschnittliche (quadrierte) kanonische Korrelation“ interpretieren lässt. Pillais Spur und Korrelation kann nicht negativ sein. So kann auch das im Bootstrapverfahren gewonnene Konfidenzintervall mit seinem unteren Grenzwert nicht den Wert .0 unterschreiten. Nur wenn der untere Grenzwert gleich .0 ist, kann konstatiert werden, dass der jeweilige Korrelationskoeffizient bezüglich des vorgegebenen Vertrauensniveaus (von üblicherweise 95%) nicht signifikant ist.

Wilks Lambda als selbständiger Koeffizient liefert dem Forscher nur dann wertvolle Information, wenn es in einen Korrelationskoeffizienten und einen F-Wert überführt wird und darüber informiert, wie stark und wie signifikant die unabhängigen Variablen auf die abhängigen einwirken. Das durch Bootstrap gewonnene Konfidenzintervall von Wilks Lambda kann zwar wie gewohnt interpretiert werden, liefert aber keine besonders interessierende Information.

Die Ergebnisse aus dem Bootstrapping des multivariaten ALM werden in einer Tabelle ausgegeben, die hier in zwei Teile getrennt abgebildet wird.

Tabelle 1 aus multivariater Abalyse

	Wilks Lambda	Standardfehler und Konfidenzintervall

	Wilks Lambda *c		Standard	Konfidenzintervall *b	
	original	Bootstrap*a	fehler	unten	oben
Haupteffekte					
Interaktionseffekte					
A1 männl	0.724098	0.733795	0.049429	0.638163	0.835634
A2 weibl	0.724098	0.733795	0.049429	0.638163	0.835634
B1 Land	0.978048	0.973196	0.016310	0.932552	0.995429
B2 Stadtrand	0.992398	0.988223	0.008415	0.967703	0.999000
B3 Stadt	0.979878	0.975476	0.016923	0.933157	0.998118
C1 Unterschic	0.998066	0.993292	0.006414	0.977009	0.999822
C2 Mittelschi	0.984738	0.980786	0.013177	0.946952	0.998660
C3 Oberschich	0.993621	0.987413	0.011977	0.952769	0.999757
A1 B1	0.997650	0.991652	0.008129	0.971095	0.999742
A1 B2	0.991235	0.987030	0.009876	0.962903	0.999376
A1 B3	0.998491	0.991930	0.008100	0.970465	0.999836
A2 B1	0.997650	0.991652	0.008129	0.971095	0.999742
A2 B2	0.991235	0.987030	0.009876	0.962903	0.999376
A2 B3	0.998491	0.991930	0.008100	0.970465	0.999836
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
A1 B1 C1	0.986405	0.981224	0.011139	0.954578	0.997202
A1 B1 C2	-	-	-	-	-
A1 B1 C3	0.986405	0.981224	0.011139	0.954578	0.997202
A1 B2 C1	-	-	-	-	-
A1 B2 C2	-	-	-	-	-
A1 B2 C3	-	-	-	-	-
A1 B3 C1	0.986405	0.981224	0.011139	0.954578	0.997202
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
A2 B3 C3	0.986405	0.981224	0.011139	0.954578	0.997202
nominale Variable					
und Interaktionen					
(Dummies zusammengefasst)					
V9 Geschlech	0.724098	0.733795	0.049429	0.638163	0.835634
V4 Wohnlage	0.969591	0.960867	0.019635	0.914241	0.989495
V5 Herkunft	0.982556	0.973061	0.015898	0.932724	0.994554
V9*V4	0.990555	0.979764	0.011783	0.949304	0.996399
V9*V5	0.993685	0.983575	0.011812	0.954760	0.998094
V4*V5	0.972413	0.958043	0.019207	0.913664	0.988062
V9*V4*V5	0.986405	0.981224	0.011139	0.954578	0.997202
Kovariante					
Alter	0.181884	0.178963	0.018807	0.145544	0.219224
Bildungsniveau	0.877402	0.874599	0.029329	0.810929	0.925738
Berufsqualifik	0.862363	0.857589	0.029222	0.798148	0.912388
	Pillais Spur				
	original	Bootstrap *a			
alle unabhaengig Variablen zusamm	1.451122	1.468743	0.030277	1.412622	1.529937

***a** mit "Bootstrap" wird der Mittelwert aus den Bootstrap-Stichproben bezeichnet "original" bezeichnet den Wert aus Originalstichprobe

***b** Das Konfidenzniveau ist 95.00%. Das bedeutet: Von den aufsteigend sortierten 1000 Mittelwerten aus den Bootstrap-Stichproben befinden sich 95.00% der Werte zwischen den Konfidenzgrenzen und je 2.50% oberhalb bzw. unterhalb der Konfidenzgrenzen

***c** für alle Variable dieser Tabelle - ausser fuer "alle unabhaengigen Variablen zusammen": gilt: Pillais Spur = 1.0 - Wilks Lambda

***d** Dummies und Kovariate koennen wegen linearer Abhaengigkeit in x Stichproben ausgeschlossen werden. Die Bootstrap-Ergebnisse dieser Variablen beruhen so auf 1000 Stichproben minus der in der letzten Spalte der Tabelle angegebenen Zahl x

Tabelle 2 aus multivariater Abalyse

partielle Korrelationen (nach Pillai)
Standardfehler und Konfidenzintervall

	partielle Pillai-Korr. original	Bootstrap*a	Standard fehler	Konfidenzintervall *b unten	oben	Variable *d eliminiert in x Stichproben
Haupteffekte Interaktionseffekte						
A1 männl	0.525264	0.513608	0.049137	0.405420	0.601529	–
A2 weibl	0.525264	0.513608	0.049137	0.405420	0.601529	–
B1 Land	0.148163	0.156105	0.049376	0.067612	0.259708	–
B2 Stadtrand	0.087187	0.101435	0.038596	0.031621	0.179715	–
B3 Stadt	0.141851	0.146880	0.054341	0.043378	0.258540	–
C1 Unterschic	0.043980	0.073179	0.036801	0.013338	0.151627	–
C2 Mittelschi	0.123540	0.129890	0.048426	0.036607	0.230321	–
C3 Oberschich	0.079867	0.100250	0.050394	0.015576	0.217327	–
A1 B1	0.048477	0.081603	0.041123	0.016064	0.170014	–
A1 B2	0.093624	0.105161	0.043738	0.024984	0.192607	–
A1 B3	0.038847	0.079234	0.042347	0.012788	0.171858	–
A2 B1	0.048477	0.081603	0.041123	0.016064	0.170014	–
A2 B2	0.093624	0.105161	0.043738	0.024984	0.192607	–
A2 B3	0.038847	0.079234	0.042347	0.012788	0.171858	–
.	.	.	.	.	.	
.	.	.	.	.	.	
.	.	.	.	.	.	
A1 B1 C1	0.116597	0.130733	0.041073	0.052899	0.213125	–
A1 B1 C2	–	–	–	–	–	1000
A1 B1 C3	0.116597	0.130733	0.041073	0.052899	0.213125	–
A1 B2 C1	–	–	–	–	–	1000
A1 B2 C2	–	–	–	–	–	1000
.	.	.	.	.	.	
.	.	.	.	.	.	
.	.	.	.	.	.	
A2 B3 C3	0.116597	0.130733	0.041073	0.052899	0.213125	–
nominale Variable und Interaktionen (Dummies zusammenge						
V9 Geschlech	0.525264	0.513608	0.049137	0.405420	0.601529	
V4 Wohnlage	0.123660	0.135871	0.035118	0.072506	0.207706	
V5 Herkunft	0.093465	0.111259	0.033641	0.052211	0.184002	
V9*V4	0.068727	0.096359	0.029244	0.042448	0.159343	
V9*V5	0.056205	0.085046	0.031515	0.030869	0.150406	
V4*V5	0.117667	0.141434	0.033622	0.077331	0.209106	
V9*V4*V5	0.116597	0.130733	0.041073	0.052899	0.213125	
Kovariante						
Alter	0.904498	0.906051	0.010408	0.883615	0.924368	–
Bildungsniveau	0.350141	0.351690	0.041430	0.272511	0.434823	–
Berufsqualifik	0.370994	0.375334	0.039201	0.295994	0.449280	–
alle unabhaengig Variablen zusamm	0.851799	0.856909	0.008828	0.840423	0.874625	

P20.25.13 "Plausible Values", Rubin-Kalkül und Bootstrap

Im Almo-Dokument Nr. 12 "Daten-Imputation", Abschnitt P45.7.4 wird das Konzept der "plausible values" und der Kalkül nach Rubin sehr ausführlich dargestellt. Als Beispiel wurde die Pisa-Studie zur Leistung von Schülern in Mathematik, Lesen und Wissenschaft verwendet. Wir werden dieses Beispiel wieder aufgreifen und im Folgenden mit der Programm-Maske "Bootstrap_Pisa.Alm" ein ALM mit folgender Variablen- und Analyse-Konstellation rechnen:

1. Die abhängige Variable ist die Schüler-Kompetenz in Mathematik, vertreten durch 10 "plausible values".
2. Als ursächliche Variable für diese werden eingesetzt
   a. die unabhängigen nominalen Variablen Geschlecht und Immigrationsstatus

b. die unabhängige quantitative Variable soziale Herkunft, gemessen durch die Schulbildung des Vaters

3. Es wird ein multivariates ALM gerechnet. Es liefert die Ergebnisse aus 10 univariaten Analysen für jeden einzelnen plausible value und auch eine multivariate Analyse, die wir in diesem Zusammenhang nicht benötigen (und durch eine Option ausblenden können)
4. Gemäß dem Kalkül nach Rubin werden aus den 10 univariaten Analysen für die unabhängigen Variablen die durchschnittlichen Effekte und Regressionskoeffizienten sowie deren Standardfehler berechnet.
5. Diese Koeffizienten werden dann dem Bootstrap-Verfahren unterzogen, das uns je Effekt bzw. Regressionskoeffizient den Standardfehler, das Konfidenzintervall und die Signifikanz p liefert.

P20.25.13.1 Die Daten für Programm-Maske "Bootstrap_Pisa.Alm"

In "www.almo-statistik.de" kann der Ordner "Pisa2015a" herunter geladen werden. Er enthält die Almo-Datei

"Schueler_A_B_CH_D_FIN_FRA_ITA_NL.dir"

für eine Reihe europäischer Staaten. Die Datei umfasst 57980 Datensätze (Schüler) und 932 Variable. Die dazu gehörende Datei der Variablennamen ist

"Schueler_A_B_CH_D_FIN_FRA_ITA_NL.nam"

Sie umfasst die Variablennamen der 932 Variablen

Die abhängige Variable in "Bootstrap_Pisa.Alm" ist die Kompetenz der Schüler in Mathematik. Sie wird durch folgende 10 "plausible values" repräsentiert. In der großen Namensdatei "Schueler_A_B_CH_D_FIN_FRA_ITA_NL.nam" erhalten sie folgende Variablen-Nummern und -Namen:

```
Name810 = PV1MATH.Plausible.Value.1.in.Mathematics;
Name811 = PV2MATH.Plausible.Value.2.in.Mathematics;
    .       .                       .
    .       .                       .
Name818 = PV9MATH.Plausible.Value.9.in.Mathematics;
Name819 = PV10MATH.Plausible.Value.10.in.Mathematics;
```

Die unabhängige nominale Variable haben in oben angegebenen umfangreichen Dateien folgende Variablen-Nummern und -Namen: Nominale unabhängige Variable:

```
Name29  = ST004D01T.Geschlecht:(1)Female,(2)Male;
Name660 = IMMIG.Index.Immigration.status:
              (1)Native,
              (2)Second-Generation,
              (3)First-Generation;
```

Mit "Second-Generation" ist gemeint: "bereits im Inland geboren", mit "First-Generation" "im Herkunftsland geboren". Als quantitative unabhängige Variable wird die Schulbildung des Vaters eingesetzt

```
Name35 =
ST007Q01TA.What.is.the.highest.level.of.schooling.completed.by.your.father;
```

Aus der großen Schülerdatei "Schueler_A_B_CH_D_FIN_FRA_ITA_NL.dir" haben wir zufällig 1000 Datensätzen herausgeschnitten und so eine verkürzte Datei für Deutschland und Österreich gebildet, dabei aber nur die angegebenen 13 Analysevariablen V29,35,660,810:819 übernommen.

In der Programm-Maske "Bootstrap_Pisa.Alm" werden somit folgende Daten verwendet:

1. die Schülerdaten aus Pisa 2015 für Deutschland bzw. Austria - reduziert auf 1000 Datensätze und 13 Variable mit dem Namen **"Schueler_D1000.fre" bzw. "Schueler_A1000.fre".** Die Daten wurden zufällig aus "Schueler_A_B_CH_D_FIN_FRA_ITA_NL.dir" entnommen sofern sie aus Deutschland bzw. Österreich waren, dabei wurden jedoch Datensätze mit fehlenden Werten übersprungen
2. die Datei der Variablennamen ("**Schueler_AD.nam**") für die 13 Variablen

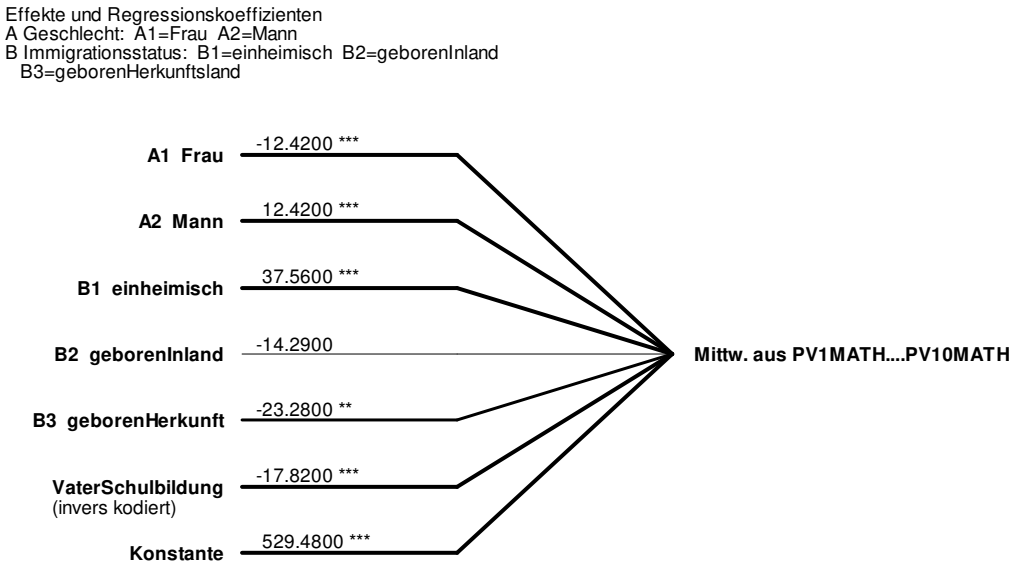
Wir vereinfachen die Namen aus den originalen Pisa2015-Daten und beginnen die Variablennummerierung bei V1.

```
Name 1 = Geschlecht:(1)weiblich, (2)männlich;
Name 2 = SchulbildungVater;
Name 3 = Immigrationsstaus:(1)einheimisch,
                           (2)im Inland geboren,
                           (3)im Herkunftsland geboren,
Name 4 = PVMathe1;
Name 5 = PVMathe2;
   .          .
   .          .
Name13 = PVMathe10;
```

Wir rechnen also folgendes Modell:

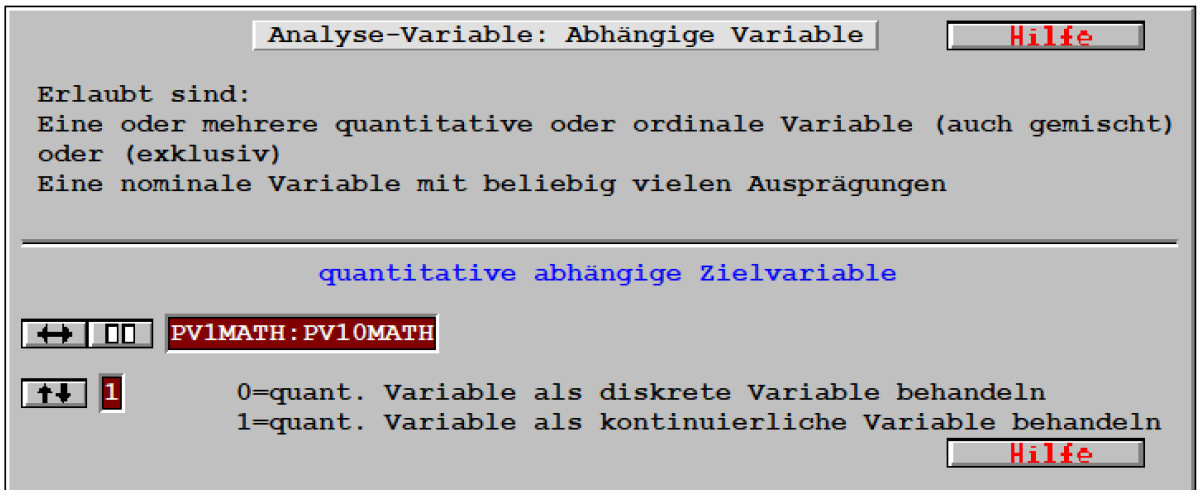
```
Leistung = b1*weiblich    + b2*männlich +
           b3*einheim.    + b4*Inlandgeboren + b5*Herkunftgeboren +
           ß1*SchuleVater + Konstante
```

Als Flussdiagramm der Effekte/Regressionskoeffizienten dargestellt:

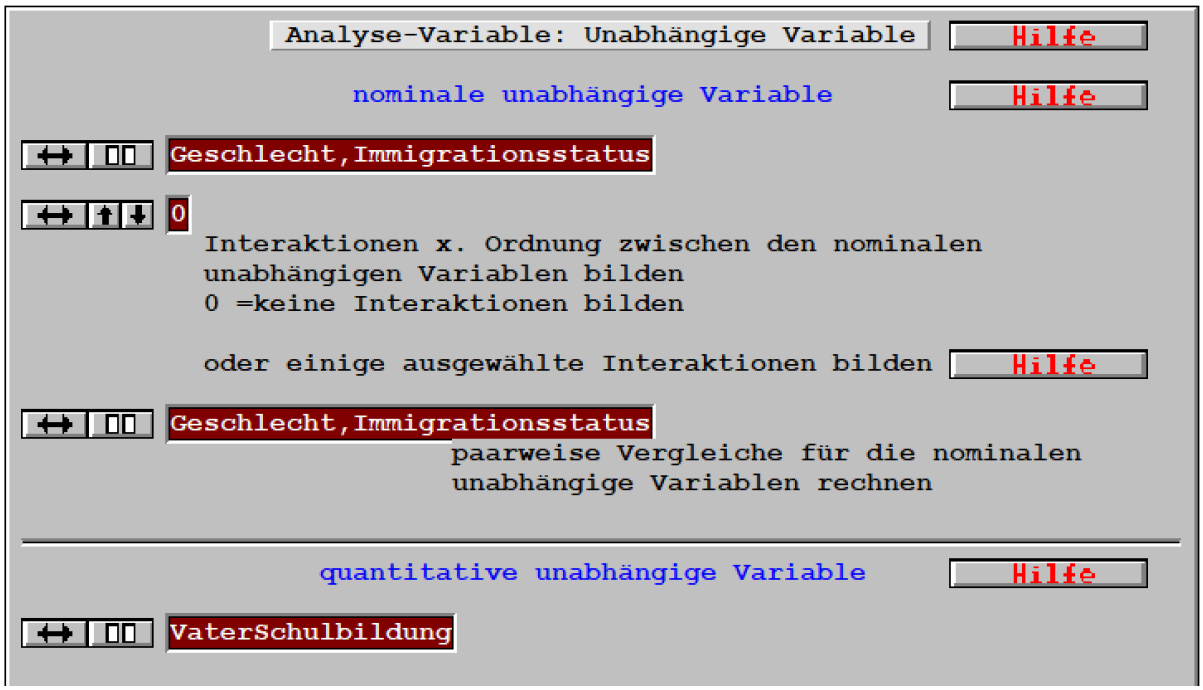


P20.25.13.2 Eingabe- und Optionsboxen von "Bootstrap_Pisa.Alm"

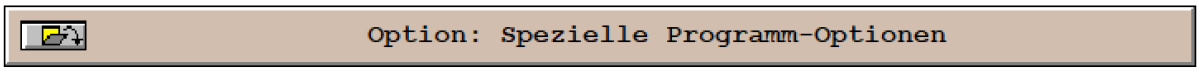
Das Programm ist identisch mit dem bereits ausführlich erläutertem Standardprogramm Prog20my. Es werden deswegen nur die Eingabeboxen interpretiert, die spezifisch für das Beispiel sind.
Die abhängigen Variablen sind die 10 "plausible values" der Kompetenz in Mathematik.



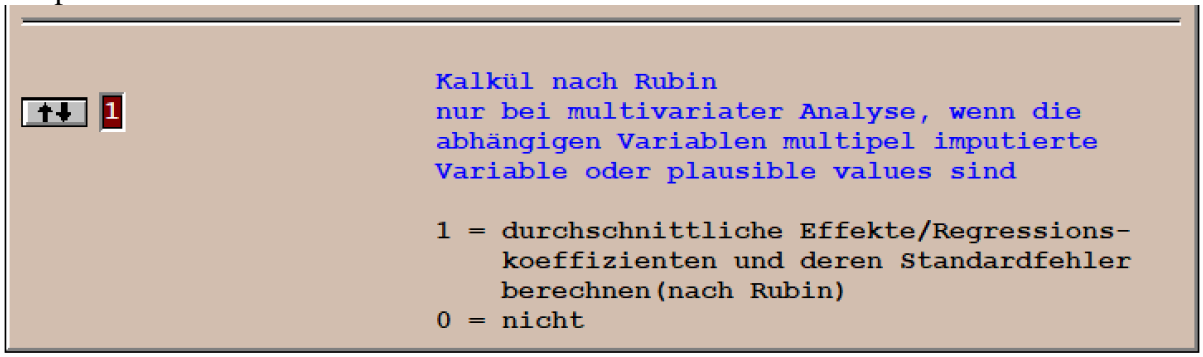
Die unabhängigen Variablen sind:



Die nachfolgende Optionsbox muss geöffnet werden



In der geöffneten Box werden drei Optionen angeboten, von denen nur die 3. für unser Beispiel bedeutsam ist



Wird "1" eingegeben, dann werden, wie im nachfolgenden Abschnitt gezeigt, die Ergebnisse nach dem Rubin-Kalkül für die Original-Stichprobe und die Ergebnisse nach dem Bootstrap ausgegeben.

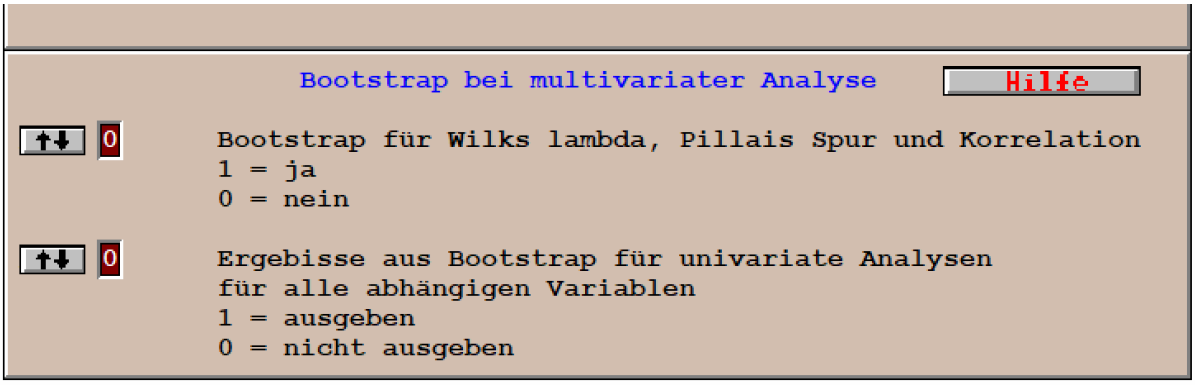
P20.25.13.2 Ergebnisse aus Programm-Maske "Bootstrap_Pisa.Alm"

Wir zeigen hier nur die aus dem Rubin-Kalkül hervor gegangenen Ergebnisse. Für die Originalstichprobe werden zuerst die Ergebnisse aus den univariaten Analysen hinsichtlich aller abhängigen Variablen ausgegeben. Dann wird folgende Tabelle ausgegeben:

```
Rubin-Verfahren:
Durchschnittliche Effekte und Regressionskoeffizienten von unabhaengigen Variablen
hinsichtlich multipel imputierter Variabler bzw. hinsichtlich von "plausible values"

                    durchschnittliche
                              Effekte                          Signifikanz
                        Regress.koeff     Standardfehler      p     (1-p)*100
                        ---------------------------------------------------
A1  Frau                   -12.418438        3.121861      0.000070     99.99
A2  Mann                    12.418438        3.121861      0.000070     99.99
B1  einheimisch             37.562675        6.912138      0.000000    100.00
B2  geborenInland          -14.286552        8.521269      0.093678     90.63
B3  geborenHerkunftland    -23.276123       12.911849      0.071486     92.85
V2  VaterSchulbildg.       -17.819710        2.922916      0.000000    100.00
Konstante                  529.480367        9.842523      0.000000    100.00
```

Nach dem Bootstrap mit 1000 Stichproben entsteht eine sehr umfangreiche Ergebnisliste. So werden die 10 univariaten Ergebnisse für jede der 10 plausible values ausgegeben. Das kann verhindert werden, wenn in der Bootstrap-Optionsbox im letzten Abschnitt eine "0" eingesetzt wird.



Das Bootstrap-Ergebnis für das Rubin-Verfahren ist folgendes

```
Bootstrap-Ergebnisse aus "Plausible-Values"-Analyse nach Rubin aus 1000 Stichproben
================================================================================

                    Effekte / Regressionskoefizienten, Standardfehler
                  Signifikanz, optimales Konfidenzniveau, Konfidenzintervall

                                                                          *e
                                 *a        *b       *c       *d  Konfidenzintervall
                Regress.koeff/Effekte  Standard  Signif  optimal       Konfniv=0.950
                  original  Bootstrap    fehler     p     Konfniv   unten       oben
--------------------------------------------------------------------------------

Haupteffekte
------------
A1  Frau         -12.41844  -12.46170   2.26328  0.0010  0.9990  -17.01203  -8.119289
A2  Mann          12.41844   12.46170   2.26328  0.0010  0.9990    8.11929  17.012025
B1  einheimisch   37.56268   37.61556   6.13854  0.0010  0.9990   25.14615  48.946853
B2  geborenInl.  -14.28655  -14.24335   7.37097  0.0420  0.9580  -28.82707  -0.988905
B3  geborenHerk. -23.27612  -23.37221  11.39838  0.0520  0.9480  -44.30976   0.280203
................................................................................
```

Kovariante und Konstante							
VaterSchulbildg.	-17.81971	-17.75576	2.34720	0.0010	0.9990	-13.19736	-22.09134
Konstante	529.48037	529.21005	8.35631	0.0010	0.9990	513.52944	545.50942

*a ...*f siehe bei Ausgabe der univariaten Analyse

Mit letzterem Ergebnis verfügen wir für die originalen Effekte bzw. Regressionskoeffizienten je über ein Konfidenzintervall und einen p-Wert der "verteilungsfrei" ist, der also nicht auf der üblichen Normalverteilungsannahme basiert.

Werden die Konfidenzintervalle und p-Werte mit dem Perzentil-t -Verfahren gerechnet, dann entstehen nur minimal andere Werte. Der p-Wert für "geboren im Herkunftsland" ist dann allerdings mit p=0.031 signifikant.

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.