



Logitanalyse Probitanalyse

P22

Johann Bacher, Kurt Holm, Heinrich Potuschak

Almo Statistik-System

<http://www.almo-statistik.de>

holm@almo-statistik.de

kurt.holm@jku.at

Autoren: Prof. Dr. Johann Bacher, Universität Linz, Österreich
em. Prof. Dr. Kurt Holm, Universität Linz, Österreich
Dr. Heinrich Potuschak, Universität Linz, Österreich

Der vorliegende Text zur Logit- und Probit-Analyse ist eine überarbeitete Version des Kapitels P22 aus dem Almo-Handbuch "Teil 4: Fortgeschrittene Verfahren". Siehe auch das Almo-Dokument Nr.9b "Bootstrap bei der Logit- und Probitanalyse" sowie Nr.10 "Koeffizienten der Logitanalyse" und die kompakte Darstellung der Logitanalyse im Almo-Dokument Nr.25 "Statistische Datenanalyse Teil II". Im Text wird häufig auf das Dokument **P0** Bezug genommen. Dabei handelt es sich um das Almo-Dokument "Arbeiten mit Almo" (Dokument 0).

Weitere Almo-Dokumente

Die folgenden Dokumente können alle kostenlos von der Handbuchseite in <http://www.almo-statistik.de> heruntergeladen werden. Siehe insbesondere das Dokument Nr.10 Koeffizienten der Logitanalyse.PDF, in dem die Ergebnisse der Logitanalyse ausführlich beschrieben 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) und 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
- 9b. Bootstrap bei der 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 Linearem Modell III.PDF
14. Ereignisanalyse: Sterbetafel-Methode, Kaplan-Meier-Schätzer,
Cox-Regression.PDF (1,5 MB)
15. Faktorenanalyse.PDF (1,6 MB)
- 15a. Bootstrap bei Faktorenanalyse.PDF (1,7 MB)
16. Konfirmatorische Faktorenanalyse.PDF (0,3 MB)
17. Clusteranalyse.PDF (3 MB)
18. Pisa 2012 Almo-Daten und Analyse-Programme.PDF (17 KB)
19. Guttman- und Mokken-Skalierung.PFD (0.8 MB)
20. Latent Structure Analysis.PDF (1 MB)
21. Statistische Algorithmen in C (80 KB)
22. Conjoint-Analyse (PDF 0,8 MB)
23. Ausreisser entdecken (PDF 170 KB)
24. Statistische Datenanalyse Teil I, Data Mining I
25. Statistische Datenanalyse Teil II, Data Mining II
26. Statistische Datenanalyse Teil III, Arbeiten mit Almo-
Datenanalyse-System
27. Mehrfachantworten, Tabellierung von Fragen mit Mehrfach-
antworten (0.8 MB)
28. Metrische multidimensionale Skalierung (MDS) (0,4 MB)
29. Metrisches multidimensionales Unfolding (MDU) (0,6 MB)
30. Nicht-metrische multidimensionale Skalierung (MDS) (0,5 MB)
31. Pfadanalyse.PDF (0,7 MB)
32. Datei-Operationen mit Almo (1,1 MB)
33. Wählerstromanalyse und Wahlhochrechnung.PDF (1,6 MB)

34. Soziometrie. Auswertung soziometrischer Daten (0,5 MB)
 35. Konfidenzintervall und p-Wert beim Bootstrap-Verfahren (200 KB)

Inhaltsverzeichnis

P22 Logit- und Probitanalyse.....	1
<i>P22.0 Einleitung</i>	<i>4</i>
P22.0.1 Lineare Wahrscheinlichkeitsanalyse.....	4
P22.0.2 Logit-Analyse	6
P22.0.3 Probit-Analyse	8
P22.0.4 Logt-, Probit-Analyse als Kleinste-Quadrate-Schätzung	10
<i>P22.1 Das Modell der Logit- und Probitanalyse</i>	<i>13</i>
P22.1.1 Das Modell der binären und multinomialen Logitanalyse und der binären Probitanalyse.....	13
P22.1.2 Binäre Logit- und Probitanalyse mit unabhängigen nominalen und quantitativen Variablen.....	16
P22.1.3 Die Logitanalyse mit nominal-polytomer abhängiger Variablen	22
P22.1.4 Die Logit- und Probitanalyse mit ordinaler abhängiger Variablen.....	25
P22.1.5 Die Maximum-Likelihood-Schätzung	25
<i>P22.2 Eingabe und Ausgabe.....</i>	<i>27</i>
P22.2.1 Eingabe mit Maskenprogramm.....	27
P22.2.1.1 Maskenprogramm zur Eingabe von Individualdaten Prog22m	27
P22.2.1.2 Erläuterungen zu den Boxen	30
P22.2.1.3 Maskenprogramm zur Eingabe fertiger Tabellen mit gruppierten Daten	45
P22.2.1.4 Erläuterungen zu den Boxen	49
P22.2.3 Ergebnisse aus binärer Logitanalyse mit unabhängigen nominalen Variablen ...	51
P22.2.3.1 Modell-Prüfgrößen	52
P22.2.3.1 Interpretation der Regressionskoeffizienten	59
P22.2.3.2 Interpretation des Koeffizienten "Risiko" (beim Logit-Modell).....	61
P22.2.3.3 Interpretation der Kontraste	62
P22.2.3.4 Beobachtete und prognostizierte Häufigkeiten	63
P22.2.4 Ergebnisse aus binärer Logitanalyse mit unabhängigen nominalen und quantitativen Variablen	64
P22.2.5 Ergebnisse aus Logitanalyse mit polytomer abhängiger Variable	71
P22.2.6 Ergebnisse aus Logitanalyse mit ordinaler abhängiger Variable	78
<i>Literatur.....</i>	<i>81</i>

P22 Logit- und Probitanalyse

Die Merkmale des in Almo enthaltenen Logit- und Probit-Modells sind folgende:

1. Es wird die Maximum-Likelihood-Schätzung verwendet. Im Rahmen des allgemeinen linearen Modells wird mit Maskenprogramm Prog20mj die Logit- und Probit-Analyse auch als Kleinste-Quadrate-Schätzung gerechnet. Siehe dazu Almo-Dokument Nr. 13a "Allgemeines lineares Modell II", Abschnitt P20.9.3.3.
2. Beim Logit- und Probit-Modell können die unabhängigen Variablen nominal und/oder quantitativ sein. Ihre Anzahl ist beliebig.
3. Die abhängigen Variablen können nominal (dichotom oder polytom) oder ordinal sein - und zwar in folgender Weise

abhängige Variable	Logitmodell	Probitanalyse
dichotom (ordinal oder nominal)	binäres Logitmodell	binäres Probitmodell
polytom-nominal	multinomiales Logitmodell	nicht möglich
polytom-ordinal	ordinales Logitmodell	ordinales Probitmodell

Das Programm zur Probit- und Logitanalyse wurde von **Heinrich Potuschak** programmiert. **Johann Bacher** hat es an das ALMO-System adaptiert und Teile dieses Kapitels verfasst. **Kurt Holm** hat weitere Programmteile programmiert und mehrere Textteile hinzugefügt.

P22.0 Einleitung

Betrachten wir ein sehr einfaches Beispiel: Es soll untersucht werden, wie das Einkommen den Kauf einer Ware x bestimmt. Dieses Beispiel ist deswegen ein einfaches, weil nur eine unabhängige Variable und eine dichotome abhängige Variable verwendet wird.

Die unabhängige Variable, das Einkommen, ist quantitativ. Wir verwenden dafür in unseren Testdaten (".\Testdat\Testdat.fre") die Variable 5.

Die abhängige Variable, der Kauf, ist nominal. Kauf besitzt 2 Ausprägungen: ja und nein. Wir verwenden dafür in unseren Testdaten die Variable 10.

P22.0.1 Lineare Wahrscheinlichkeitsanalyse

Wir rechnen zuerst mit der Programm-Maske P20_Bspl.Alm ein allgemeines lineares Modell, genauer: eine lineare Diskriminanzanalyse bzw. eine lineare Wahrscheinlichkeitsanalyse. Das Programm findet man nach Klick auf den Knopf „alle Progs“ am Oberrand des Almo-Fensters.

Almo liefert uns folgende Ergebnisse (gekürzt):

Koeffizienten fuer quantitative Variable aus univariater Analyse

hinsichtlich der abhaengigen Variablen V10-0 Kauf: ja

Variable	Regr. koeff.	Standard fehler	95% Konfidenzbereich nach		erklärte Streuung	part. Korrel.	F-Wert	Signifikanz p	df1	df2	Teststärke
			oben	u. unten							
V5 Einkommen	0.0640	0.0330	0.0659	0.9092	0.245	3.767	0.057	94.31	1	59	0.4806

Koeffizienten fuer Konstante: 0.210491

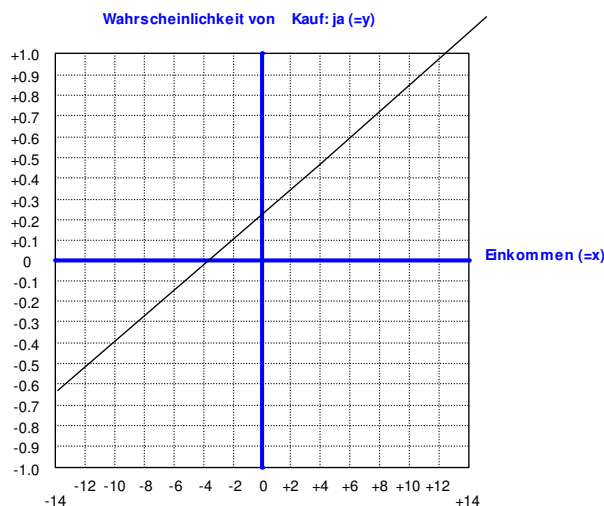
Die wesentlichen Ergebnisse sind also. Der Regressionskoeffizient beträgt 0.064. Er ist mit $(1-p)100 = 94.31\%$ signifikant. Die Konstante hat einen Wert von 0.2105

Wir können also die lineare Gleichung schreiben:

$$p = 0.064 * \text{Einkommen} + 0.2105$$

p = das ist die Wahrscheinlichkeit für "Kauf: ja"

Wir wollen die Gleichung als Gerade zeichnen:



Die Wahrscheinlichkeit eines Kaufs ist z.B. bei einem Einkommen

von 4 Einheiten: $p = 0.064 * 4 + 0.2105 = 0.4665$

von 8 Einheiten: $p = 0.064 * 8 + 0.2105 = 0.7225$

von 10 Einheiten: $p = 0.064 * 10 + 0.2105 = 0.8505$

Der höchste Einkommenswert in unseren Daten ist 10.

Nun wollen wir wissen, wie die Kaufwahrscheinlichkeit bei einem Einkommen von 14 ist.

14 Einheiten: $p = 0.064 * 14 + 0.2105 = 1.1950$

Es entsteht eine Wahrscheinlichkeit größer als 1.0. Das gibt es nicht. Das ist die Schwäche der linearen Wahrscheinlichkeitsanalyse. Es können Wahrscheinlichkeiten prognostiziert werden, die über 1.0 oder unter 0 liegen. Allerdings ist folgendes zur Verteidigung dieses Verfahrens zu sagen:

1. Wahrscheinlichkeiten größer 1.0 treten in der Regel nur auf, wenn man für die

unabhängige Variable Werte einsetzt, die weit außerhalb des Wertebereichs der empirisch gewonnen Daten liegen.

2. Man setzt Wahrscheinlichkeiten größer 1.0 einfach auf 1.0 und solche unter 0 auf 0.
3. Die Koeffizienten, die die lineare Wahrscheinlichkeitsanalyse liefert, sind einfach und klar zu interpretieren – so wie man es bei der Regressionsanalyse gewohnt ist. Dies gilt für die Koeffizienten der Logit- und Probitanalyse keinesfalls, wie wir noch sehen werden.

Die lineare Wahrscheinlichkeitsanalyse leidet noch an einem weiteren Problem, dem Problem der Heterokedastizität der Varianzen.

Es besteht modellbedingte Varianz-Heteroskedastizität mit der Folge, dass die Schätzer für die Parameter der ursächlichen Variablen zwar unverzerrt und konsistent, aber nicht mehr effizient sind. Das bedeutet, dass die Standardfehler der Effekte und Regressionskoeffizienten der ursächlichen Variablen nicht minimal sind, mit der Folge, dass die Signifikanzüberprüfung mit t- und F-Test nicht korrekt ist. Siehe dazu die ausführliche Darstellung bei Aldrich/Nelson (1984, S. 12ff) und Urban (1993, S. 17ff), sowie Urban (1982, Abschnitt 3.1 und 3.1.1).

P22.0.2 Logit-Analyse

Die Logit- und die Probit-Analyse besitzen diese Schwäche nicht. Wir wollen mit denselben Daten mit Programm-Maske P22_Bspl.Alm zuerst eine Logit- und dann eine Probit-Analyse rechnen. Der Benutzer findet das Programm durch Klick auf den Knopf „alle Progs“ am Oberrand des Almo-Fensters. Das Programm ist identisch mit der Standard-Programm-Maske Prog22m.Msk ist jedoch unserem Beispiel entsprechen anders ausgefüllt.

Almo liefert dieses Ergebnisse (gekürzt):

```

-----
Ergebnisse fuer 1. Auspraegung "ja"
der abhaengigen Variablen V10 Kauf
(als Referenz wird die letzte Auspraegung "nein" verwendet)

unabh.      Regress.      "Risiko"      Stand.- z-Wert  Signifik.  partielle
Variab.     Koeffiz.      exp(Regr.-   Fehler      (1-p)*100  Korrelat.
              koefiz.)
-----
Konstante  -1.21869      -             0.62553    1.95       94.86      -
Einkommen2 0.27058      1.31072      0.14594    1.85       93.63      0.13069
-----

```

Die Logit-Analyse verwendet die logistische Funktion. Deren Gleichung ist:

$$p = \frac{1}{1 + e^{-(c + \beta x)}}$$

wobei

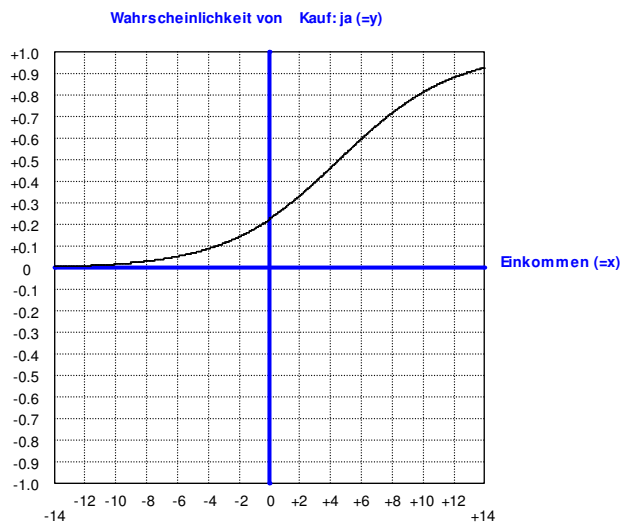
x = die unabhängige quantitative Variable
c = Regress.koeffizient der Konstanten,
β = Regressionskoeffizient von **x**

Für unser Beispiel lautet die Gleichung:

$$p = \frac{1}{1 + e^{-(-1.21869 + 0.27058 * \text{Einkommen})}}$$

Almo liefert folgende Grafik:

Logistische Funktion
 $Y = 1/(1+e^{-(1.2+0.27 \cdot X)})$



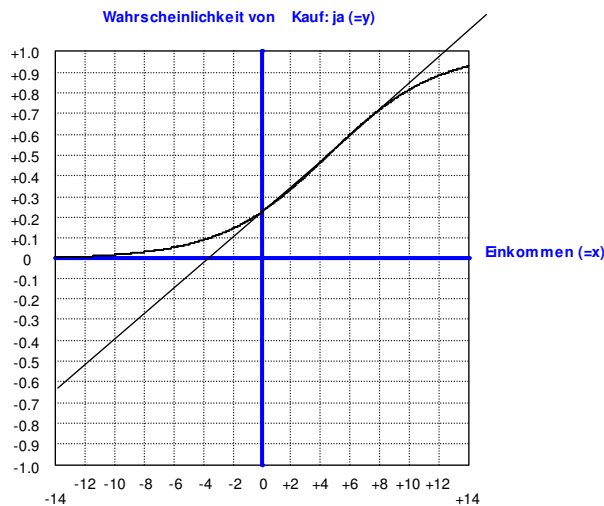
P22.0.2.1 Eigenschaften der logistischen Funktion:

1. Die logistische Funktion nähert sich asymptotisch den Werten $p = 0$ und $p = 1$
2. Die Konstante c bestimmt die horizontale Lage der Kurve. Je grösser c umso weiter links liegt die Kurve - bei positivem β . Ist β negativ dann umgekehrt.
3. Der Regressionskoeffizient β bestimmt die Steilheit. Je grösser β absolut ist umso steiler. Bei positivem Vorzeichen wächst die Kurve von links nach rechts, bei negativem Vorzeichen umgekehrt

Die Wahrscheinlichkeit eines Kaufs ist bei einem Einkommen

Einkommen	Logit-Analyse	lineare Wahrscheinlichkeitsanalyse
von 4 Einheiten:	$p = 0.4660$	0.4665
von 8 Einheiten:	$p = 0.7203$	0.7225
von 10 Einheiten:	$p = 0.8156$	0.8505
von 14 Einheiten:	$p = 0.9289$	1.1950

Nun wollen wir die Gerade und die logistische Funktion in einer gemeinsamen Grafik zeigen:



Man erkennt sehr deutlich, dass die Gerade und die logistische Funktion von Einkommen = 0 bis zu einem Einkommen von ca. 9 Einheiten sich decken. Erst dann gehen die beiden auseinander.

Die Kurve der logistischen Funktion

Mit der Programm-Maske "Nonlin1.Grff" kann die Kurve der logistischen Funktion gezeichnet werden. Die Maske wird gefunden durch Klick auf den Knopf "alle Progs" am Oberrand des Almo-Fensters. Die Maske wird folgender Weise ausgefüllt.

Nicht-lineare Kurve
für eine unabhängige und eine abhängige Variable

Grafik erzeugen --> Grafik

Hilfe zu Grafik --> Hilfe

Kurventyp

↑ ↓ 8

0 = Gerade	$Y=a \cdot X + \text{const}$	Hilfe
1 = Parabel oder Hyperbel	$Y=a \cdot X^2 + \text{const}$	Hilfe
2 = Exponentialfunktion	$Y=a \cdot e^{(b \cdot x)} + \text{const}$	Hilfe
3 = allgem. Exponentialfunktion	$Y=a \cdot b^{X} + \text{const}$	Hilfe
4 = Gompertz-Kurve	$Y=a \cdot b^{(c^{X})}$	Hilfe
5 = Polynom 2. Grades	$Y=a \cdot X^2 + b \cdot X + \text{const}$	Hilfe
6 = Polynom 3. Grades	$Y=a \cdot X^3 + b \cdot X^2 + c \cdot X + \text{const}$	Hilfe
7 = logarithmische Funktion	$Y=a \cdot \log_{10}(X) + \text{const}$	Hilfe
8 = logistische Funktion I	$Y=1 / (1/a + e^{-(b+c \cdot X)})$	Hilfe
9 = Ogive / kumulative Normalverteilung		Hilfe
10= logistische Funktion II	$Y=1 / (1/a + e^{-(c \cdot (b+X))})$	Hilfe
15= inverse kumulative Standard-Normalverteilung		Hilfe

Werte für die Parameter der Funktion

↔ 1	a Obergrenze, der sich die Kurve annähert
↔ -1.21869	b horizontale Lage der Kurve
↔ 0.27058	c Steilheit der Kurve
↔ 0	const

Koordinatensystem

↔ 14	x maximale Länge der x-Achse
↔ 1	y maximale Höhe der y-Achse

Durch Klick auf den großen Knopf "Grafik" oberhalb der 1. Eingabebox wird dann bewirkt, dass Almo den Grafik-Editor öffnet und die logistische Kurve zeichnet. Im Editor kann dann die Grafik in vielfältiger Weise bearbeitet werden. Siehe dazu die "AnleitungGrafik.pdf" im Wurzelverzeichnis von Almo.

Der Leser wird etwas dadurch verwirrt, dass die Notation die in der Grafik-Maske verwendet wird nicht übereinstimmt mit der Notation in obiger Formel der logistischen Funktion. Es gelten folgende Entsprechungen

Wert	Formel	Grafik-Maske	Bedeutung
1		a	Obergrenze, der sich die Kurve annähert
-1.21869	c Konstante	b	horizontale Lage der Kurve
0.27058	β Reg.koeff.	c	Steilheit der Kurve
14		x	maximale Länge der x-Achse
1		y	maximale Höhe der y-Achse

P22.0.3 Probit-Analyse

Wir rechnen nun mit Programm-Maske P22_Bspl.Alm eine Probit-Analyse. Dazu müssen wir in der Programm-Maske nur in einer Eingabe-Box das Wort "Logit" durch "Probit" ersetzen.

Almo liefert folgende Ergebnisse (gekürzt):

unabh. Variab.	Regress. Koeffiz.	Stand.- Fehler	z-Wert	Signifik. (1-p)*100
Konstante	-0.75435	0.37899	1.99	95.34
Einkommen	0.16831	0.08917	1.89	94.09

Die Probit-Analyse verwendet die Funktion der Ogive (=der kumulativen Normalverteilungsfunktion). Deren Gleichung ist:

$$p = \int \frac{\beta}{2.507} * e^{-(c+\beta x)^2 / 2}$$

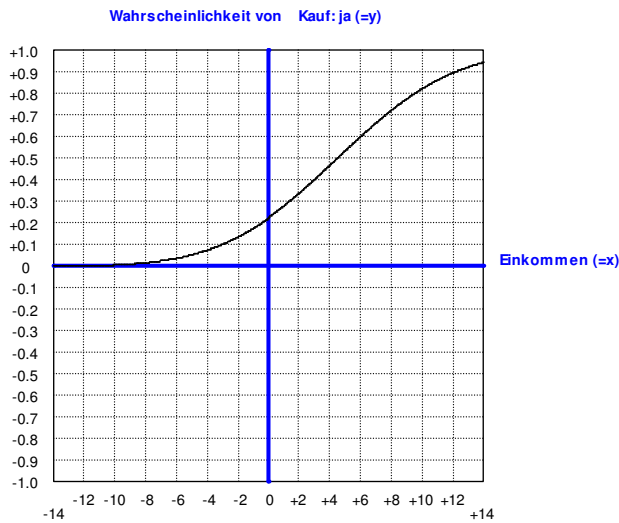
wobei c = Konstante und β = Regressionskoeffizient
2.507 ist Wurzel aus 2*pi

Für unser Beispiel lautet die Gleichung:

$$p = \int \frac{0.168}{2.507} * e^{-(-0.754+0.168x)^2 / 2}$$

Almo liefert folgende Grafik:

Ogive
 $Y = \text{Integral}(1 * 0.17 / 2.51 * e^{-(-0.75 + 0.17 * X) * (-0.75 + 0.17 * X) / 2})$



Man erkennt, dass Ogive und logistische Funktion nahezu denselben Kurvenverlauf besitzen.

Eigenschaften der Ogive:

1. Die Ogive nähert sich asymptotisch den Werten $p = 0$ und $p = 1$
2. Die Konstante bestimmt die horizontale Lage der Kurve. Je grösser umso weiter links liegt die Kurve - bei positivem β . Ist β negativ dann umgekehrt.
3. Der Regressionskoeffizient β bestimmt die Steilheit. Je grösser β (absolut) ist umso steiler. Ist das Vorzeichen positiv, dann wächst die Kurve von links nach rechts; ist das Vorzeichen negativ, dann umgekehrt.

Die Wahrscheinlichkeit eines Kaufs ist bei einem Einkommen

Einkommen	Probit-Analyse	Logit-Analyse	lineare Wahrscheinlichkeitsanalyse
von 4 Einheiten:	$p = 0.4677$	0.4660	0.4665
von 8 Einheiten:	$p = 0.7231$	0.7203	0.7225
von 10 Einheiten:	$p = 0.8235$	0.8156	0.8505
von 14 Einheiten:	$p = 0.9454$	0.9289	1.1950

Die Werte der Probit-Analyse in obiger Tabelle können mit Almo durch Klick auf den Knopf "stat. Tafelwert" am Oberrand des Almo-Fensters und Prog00t9.Msk berechnet werden. Wir blenden das kurz hier ein. Die ausgefüllte Maske für $x=4$ ist folgende:

gesucht: y-Wert der Ogive		
↔	4	x gewünschter x-Wert der Ogive
↔	1	a Obergrenze, der sich die Ogive annähert
↔	-0.75435	Konstante b horizontale Lage der Kurve
↔	1.68311	Reg.koeff. c Steilheit der Kurve
↔	0	konst verschiebt die Ogive nach rechts

Als Ergebnis wird ausgegeben:

Flaeche der allgemeinen Glockenkurve von minus pseudo-unendlich bis $x = 4$
 $= 0.467717$
 Berechnet wurde die Gleichung
 $Y = 1 * 0.1683 / 2.507 * \text{Integral}(e^{-(-0.7543 + 0.1683 * (X-0)) * (0.7543 + 0.1683 * (X-0)) / 2})$

Die Parameter in obiger Maske sind so zu verstehen:

- x = Wert der unabhang. Var., fur den der Probit-Wert p ermittelt werden soll
- b = Regress.wert der Konstanz
- c = Regress.wert der unabhangigen Variablen

Die Kurve der Ogive (=der kumulativen Normalverteilung) fur die Probit-Analyse

Mit der Programm-Maske "Nonlin1.Gr" kann die Kurve der kumulativen Normalverteilung fur die Probit-Analyse gezeichnet werden. Die Maske wird gefunden durch Klick auf den Knopf "alle Progs" am Oberrand des Almo-Fensters. Die Maske wird folgender Weise ausgefullt.

Nicht-lineare Kurve
fur eine unabhangige und eine abhangige Variable

Grafik erzeugen --> Grafik

Hilfe zu Grafik --> Hilfe

Kurventyp

↑ ↓	9		
		0 = Gerade	$Y=a \cdot X + \text{const}$
		1 = Parabel oder Hyperbel	$Y=a \cdot X^2 + \text{const}$
		2 = Exponentialfunktion	$Y=a \cdot e^{(b \cdot X)} + \text{const}$
		3 = allgem. Exponentialfunktion	$Y=a \cdot b^{X} + \text{const}$
		4 = Gompertz-Kurve	$Y=a \cdot b^{(c \cdot X)}$
		5 = Polynom 2. Grades	$Y=a \cdot X + b \cdot X^2 + \text{const}$
		6 = Polynom 3. Grades	$Y=a \cdot X + b \cdot X^2 + c \cdot X^3 + \text{const}$
		7 = logarithmische Funktion	$Y=a \cdot \log_{10}(X) + \text{const}$
		8 = logistische Funktion I	$Y=1 / (1/a + e^{-(b+c \cdot X)})$
		9 = Ogive / kumulative Normalverteilung	
		10= logistische Funktion II	$Y=1 / (1/a + e^{-(c \cdot (b+X))})$
		15= inverse kumulative Standard-Normalverteilung	

Werte fur die Parameter der Funktion

↔	1	a Obergrenze, der sich die Kurve annahert
↔	-0.75435	b horizontale Lage der Kurve
↔	0.16831	c Steilheit der Kurve
↔	0	const

Koordinatensystem

↔	14	x maximale Lange der x-Achse
↔	1	y maximale Hohe der y-Achse

Diese Grafik-Maske kann geladen werden durch Klick auf den Knopf „alle Progs“ am Oberrand des Almo-Fensters.

Auch hier entsprechen sich die Notation in der Grafik-Maske und der Notation in obiger Formel der Ogive nicht. Die richtige Entsprechung ist gleich wie oben bei der logistischen Funktion

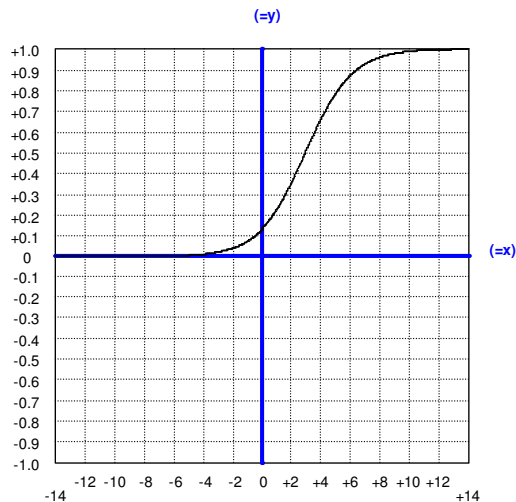
P22.0.4 Logit- Probit-Analyse als Kleinste-Quadrate-Schatzung.

Die Logit- und Probitanalyse kann auch als Kleinste-Quadrate-Schatzung (kurz: KQ) im Rahmen des allgemeinen linearen Modells mit der Programm-Maske Prog20mj.Msk gerechnet werden. Man findet dieses Programm durch Klick auf den Knopf "Verfahren", dann "Logit-Probitanalyse". Aus der KQ-Logitanalyse entstehen folgende Ergebnisse. Zum Vergleich geben wir auch die Ergebnisse der ML-

Schätzung aus

	Regressionskoeffizient		Signifikanz (1-p)100	
	KQ	ML	KQ	ML
Konstante	-1.8713	-1.21869	-	94.86 %
Einkommen	0.6347	0.27058	95.97	93.63 %

Grafisch dargestellt ergibt sich folgende logistische Kurve aus der Kleinste-Quadrate-Schätzung



Die Ergebnisse weichen erheblich von denen der oben dargestellten Logitanalyse als Maximum-Likelihood-Schätzung ab. Ob das immer so ist, das sei dahingestellt. Da das Kurvenbild der Maximum-Likelihood-Schätzung sich mit der Geraden aus der linearen Wahrscheinlichkeitsanalyse deckt, scheint die ML-Schätzung doch die vertrauenswürdiger zu sein. In der statistischen Literatur und der empirischen Forschung wird die ML-Schätzung bevorzugt.

Bei der Kleinste-Quadrate-Schätzung (in Programm-Maske Prog20mj) gibt Almo keine Funktionsgrafik aus. Die logistische Funktion muss der Benutzer selbst zeichnen.

Mit der Programm-Maske "Nonlin1.Grff" kann die Kurve gezeichnet werden. Die Maske wird gefunden durch Klick auf den Knopf "alle Progs" am Oberrand des Almo-Fensters. Die Maske wird folgender Weise ausgefüllt.

Nicht-lineare Kurve
für eine unabhängige und eine abhängige Variable

Grafik erzeugen --> **Grafik**
Hilfe zu Grafik --> **Hilfe**

Kurventyp

↑ ↓	8		
0 = Gerade	$Y=a \cdot X$	+ const	
1 = Parabel oder Hyperbel	$Y=a \cdot X^2$	+ const	Hilfe
2 = Exponentialfunktion	$Y=a \cdot e^{(b \cdot x)}$	+ const	Hilfe
3 = allgem. Exponentialfunktion	$Y=a \cdot b^x$	+ const	Hilfe
4 = Gompertz-Kurve	$Y=a \cdot b^{(c^x)}$		Hilfe
5 = Polynom 2. Grades	$Y=a \cdot X^2 + b \cdot X$	+ const	
6 = Polynom 3. Grades	$Y=a \cdot X^3 + b \cdot X^2 + c \cdot X$	+ const	
7 = logarithmische Funktion	$Y = a \cdot \log_{10}(X)$	+ const	
8 = logistische Funktion	$Y = 1 / (1/a + e^{-(b+c \cdot x)})$		Hilfe
9 = Ogive	$Y = \text{Integral}(a \cdot c / 2.507 \cdot e^{-(b+c \cdot (X-\text{const}))} \cdot (b+c \cdot (X-\text{const})) / 2)$		Hilfe

Werte für die Parameter der Funktion

↔	1	a	immer =1
↔	-1.8713	b	Wert der Konstanten
↔	0.6347	c	Regress.koeffiz. der unabhä. quant. Var.
↔	0	const	immer =0

Koordinatensystem

↔	14	x	Länge der x-Achse
↔	1	y	Länge der y-Achse immer =1

In der 1. Eingabebox wird als Kurventyp 8 eingesetzt. In der 2. Eingabebox werden die beiden Koeffizienten für die Konstante und die unabhängige quantitative Variable eingesetzt. In der 3. Eingabebox haben wir 14 eingegeben, d.h. wir lassen uns die Kurve bis zu einem Einkommen von 14 Einheiten zeichnen. Die anderen Werte sind fix.

Durch Klick auf den großen Knopf "Grafik" oberhalb der 1. Eingabebox wird dann bewirkt, dass Almo den Grafik-Editor öffnet und die logistische Kurve zeichnet. Im Editor kann dann die Grafik in vielfältiger Weise bearbeitet werden. Siehe dazu die "AnleitungGrafik.pdf" im Wurzelverzeichnis von Almo.

P22.1 Das Modell der Logit- und Probitanalyse

P22.1.1 Das Modell der binären und multinomialen Logitanalyse und der binären Probitanalyse

Der Modellansatz soll anhand eines Beispiels aus ARMINGER/KÜSTERS (1986: 35ff) dargestellt werden. (Zum Modellansatz der Logit- und Probitanalyse siehe auch die Ausführungen des Abschnitts P20.9.3.3 sowie ARMINGER/KÜSTERS 1986, GREEN 1990: 661-714, MADDALA 1990: 13-58).

Folgende Fragestellung soll untersucht werden: Welchen Einfluß haben das Geschlecht und das Alter auf die Wahrscheinlichkeit des Auftretens eines Unfall (Nichtfahrunfall oder Fahrunfall). Ausgangspunkt der Analyse ist folgende Häufigkeitstabelle:

Datensatz i	Alter (=V1)	Geschlecht (=V2)	Unfalltyp (=V3)		Gesamt
			Nichtfahrunfall (=1)	Fahrunfall (=2)	
1	1	1	10128	7468	17596
2	1	2	3847	2195	6042
3	2	1	15373	5167	20540
4	2	2	5770	2188	7958
5	3	1	3047	529	3576
6	3	2	517	85	602

Alter wird als nominale Variable mit 3 Ausprägungen behandelt.

Die Ausprägungen der Variablen bedeuten: "1" bei V1 = jung, "2" bei V1 = mittel, "3" bei V1 = alt; "1" bei V2 = männlich, "2" bei V2 = weiblich

In dem Beispiel sollen als abhängige Variablen die Auftrittswahrscheinlichkeiten von zwei Unfallarten untersucht werden, nämlich die Wahrscheinlichkeit des Auftretens eines Nichtfahrunfalls und jene des Auftretens eines Fahrunfalls. Diese beiden Variablen wollen wir mit P_1 und P_2 bezeichnen, die konkreten Ausprägungen mit p_{i1} und p_{i2} für einen bestimmten Datensatz i . Für den Datensatz 1 ($i=1$) ist $p_{11} = 10128/(10128 + 7468) = 0.58$ und $p_{12} = 7468/(10128 + 7468) = 0.42$. Die untersuchten abhängigen Variablen sind also die Zeilenanteilstwerte der Ausprägungen der abhängigen Variablen. Die Summe der Zeilenanteilstwerte ist gleich 1. Es gilt also $p_{i1} + p_{i2} = 1$ bzw. $p_{i1} = 1 - p_{i2}$ bzw. allgemein bei zwei Ausprägungen $p_{i1} = 1 - p_{i2}$.

Die Abhängigkeit von p_{i1} bzw. p_{i2} vom Geschlecht und Alter kann zunächst durch eine einfache lineare Regression mit

$$(1) \hat{p}_{i2} = b_0 + b_{11} \cdot x_{i11} + b_{12} \cdot x_{i12} + b_{21} \cdot x_{i21}$$

untersucht werden. Die Größen der Gleichung bedeuten:

\hat{p}_{i2} durch die Regressionsgleichung prognostizierte Wahrscheinlichkeit des Auftretens der Ausprägung 2 der abhängigen Variablen für den Datensatz i

Beachte: Werden Daten als fertige Tabellen eingelesen (wie oben gezeigt) und nicht als aufeinander folgende individuelle Datensätze und ist die abhängige Variable dichotom, dann analysiert ALMO deren 2. Ausprägung. Die Regressionskoeffizienten für die 1. Ausprägung ergeben sich sehr einfach durch Vorzeichenumkehr.

b_0 Regressionskonstante

Die x-Variablen werden in folgender Weise indiziert: x_{ijk} und die b-Koeffizienten: b_{jk} . Dabei bezeichnet i = den Datensatz; j = die unabhängige nominale Variable j ; k = die Ausprägung k der unabhängigen nominalen Variable.

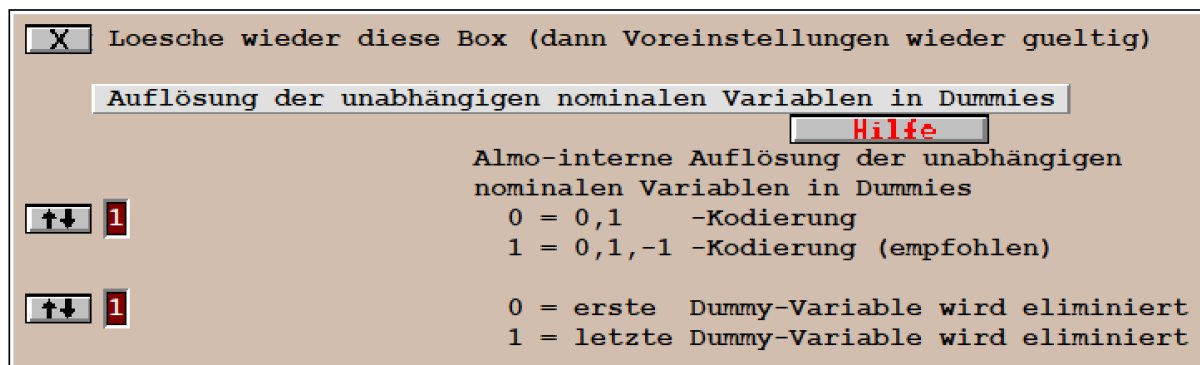
b_{11} Effekt der Ausprägung 1 der Variablen 1

x_{i11} Wert des Datensatzes in der Dummy-Variablen für die Ausprägung 1 der Variablen 1 ($x_{i11} = 1$, wenn der Datensatz i in der Variablen 1 die Ausprägung 1 besitzt, sonst ist $x_{i11} = 0$)

b_{12} Effekt der Ausprägung 2 der Variablen 1

- x_{i12} Wert des Datensatzes in der Dummy-Variablen für die Ausprägung 2 der Variablen 1 ($x_{i12} = 1$, wenn der Datensatz i in der Variablen 1 die Ausprägung 2 besitzt, sonst ist $x_{i12} = 0$)
- b_{21} Effekt der Ausprägung 1 der Variablen 2
- x_{i21} Wert des Datensatzes in der Dummy-Variablen für die Ausprägung 1 der Variablen 2 ($x_{i21} = 1$, wenn der Datensatz i in der Variablen 2 die Ausprägung 1 besitzt, sonst ist $x_{i21} = 0$)

Die Dummies werden in Prog22m.Msk in der Optionsbox "Auflösung der unabhäng. nominalen Variab. in Dummies" so gebildet:



Zur Erläuterung siehe Abschnitt P22.2.1.2, Box 8.

Binäre Logitanalyse

Im binären Logitmodell wird nicht die Wahrscheinlichkeit p_{i2} sondern der Logarithmus des Verhältnisses von der Auftretswahrscheinlichkeit der Ausprägung 1 zur Auftretswahrscheinlichkeit der Ausprägung 2 untersucht - oder alternativ umgekehrt der Ausprägung 2 zur Ausprägung 1.

Betrachten wir den zweiten Fall: Inhaltlich wird also untersucht, welche Faktoren dazu führen, dass die Ausprägung 2 (in unserem Beispiel "Fahrunfälle") häufiger auftritt als die Ausprägung 1 (in unserem Beispiel "Nichtfahrunfälle"). Die Modellgleichung für unserer Beispiel lautet:

$$(2) \ln\left(\frac{p_{i2}}{p_{i1}}\right) = b_0 + b_{11} \cdot x_{i11} + b_{12} \cdot x_{i12} + b_{21} \cdot x_{i21}.$$

In unserem Beispiel ist die Ausprägung 2 die Untersuchungsvariable. Die Ausprägung 1 wird als *Referenz-Ausprägung* bezeichnet.

In der Programm-Maske Prog22m.Msk (für individuelle Daten) kann der Benutzer wählen, welche Ausprägung er als Referenz festlegen möchte. Bei gruppierten Daten (mit fertig ausgezählter Tabelle) jedoch wird in Almo die 1. Variable als Referenz festgelegt. Der Benutzer kann das nicht ändern.

Die mit x bezeichneten Variablen sind in unserem Beispiel 0,1 -kodierte Dummies. Sie können aber auch *quantitative* unabhängige Variable bezeichnen. Die unabhängigen Variablen auf der rechten Gleichungsseite können also Dummies oder quantitative Variable oder beides zusammen sein.

Multinomiale Logitanalyse

Betrachten wir zuerst ein Beispiel, das in Abschnitt P22.2.5 ausgeführt wird. Die Daten dazu stammen aus Arminger/Küsters (1986, S.102). Das Programm ist unter dem Namen "Arm102k.Alm" in Almo enthalten. Man findet das Programm nach Klick auf den Knopf „alle Progs“ am Oberrand des Almo-Fensters.

Die abhängige polytome Variable besitzt 4 Ausprägungen:

Unfallart: (1)Sachschaden, (2)Leichtverletzt, (3)Schwerverletzte, (4) Tote

Die unabhängigen nominalen Variablen sind:

Geschlecht: maennlich, weiblich

Straßenzustand: trocken,nass,Eis

Die unabhängigen quantitative Variable ist:

Alter

Bei der multinomialen Logitanalyse wird eine Ausprägung als *Referenzausprägung* fixiert. In der Programm-Maske Prog22m.Msk (für individuelle Daten) kann der Benutzer wählen, welche Ausprägung er als Referenz festlegen möchte. Es kann die erste oder die letzte Variable sein.

Bei gruppierten Daten (mit fertig ausgezählter Tabelle) jedoch wird in Almo die 1. Variable als Referenz festgelegt. Der Benutzer kann das nicht ändern. Es muss somit die erste sein, also Sachschaden.

Es müssen nun 3 binäre Logitanalysen, wie oben in Gleichung (2) ausgedrückt, gerechnet werden. Dies sind:

$$\ln(p_2/p_1) \quad \ln(p_3/p_1) \quad \ln(p_4/p_1)$$

Die multinomiale Logitanalyse wird also in mehrere binäre Logitanalysen aufgelöst. Die inhaltliche Interpretation der Ergebnisse ist nun sehr viel komplizierter als bei der Analyse mit dichotomer abhängiger Variablen. Wir werden das in Abschnitt P22.2.5 ausführen.

Die *ordinale* Logitanalyse werden wir in Abschnitt P22.2.6 an Hand eines durchgerechneten Beispiel darstellen. Eine ordinale Logitanalyse kann sehr einfach mit der Programm-Maske Prog22m gerechnet werden.

Binäre Probitanalyse

Bei dieser wird unter Verwendung der Standardnormalverteilung die Wahrscheinlichkeit p_{i2} auf den Zahlenbereich von $-\infty$ bis $+\infty$ transformiert. Die Modellgleichung ist:

$$(3) \quad \Phi^{-1}(p_{i2}) = b_0 + b_{11} \cdot x_{i11} + b_{12} \cdot x_{i12} + b_{22} \cdot x_{i21},$$

wobei $\Phi^{-1}(\cdot)$ die Umkehrfunktion der Verteilungsfunktion der Standardnormalverteilung ist. Im Unterschied zum Logitmodell existiert für die Probitanalyse keine *multinomiale* Version jedoch eine *ordinale*. Für diese gelten dieselben Prinzipien wie sie in P22.2.6 für die ordinale Logitanalyse vorgestellt wurden. Eine ordinale Probitanalyse kann ebenfalls sehr einfach mit der Programm-Maske Prog22m gerechnet werden.

P22.1.2 Binäre Logit- und Probitanalyse mit unabhängig nominalen und quantitativen Variablen

Wir wollen ein einfaches Beispiel betrachten, bei dem nur 1 ursächliche nominale und 1 ursächliche quantitative Variable vorhanden ist: Personen kaufen ein

Produkt auf Kredit. Werden sie diesen Kredit zurückzahlen oder nicht?
 Das Beispiel ist als Syntax-Programm unter dem Namen "Prog22a.Msk" in Almo enthalten. Man findet es durch Klick auf den Knopf "alle Progs" am Oberrand des Almo-Fensters.

Die Variablen für unser vereinfachtes Beispiel sollen folgende sein:

Die Zielvariable ist Kredit-Rückzahlung: nein,
 ja

Die unabhängige nominale Variable ist Beruf: Arbeiter,
 Angestellter,
 Selbständiger

Die unabhängige quantitative Variable ist: Einkommen
 Sie wird in Einkommensklassen mit den Werten 1,2,3, ...,9 gemessen.

Almo liefert folgendes Ergebnis (gekürzt):

Ergebnisse für 2. Ausprägung "ja" der abhängigen Variablen "Rückzahlung"
 (die 1. Ausprägung "nein" wird als Referenzkategorie verwendet)

unabhängige Variable	Regress. Koeffiz.	"Risiko" exp(Regr.- koeffiz.)	relatives Risiko in %
c Konstante	1.88227	-	-
a1 Beruf:Arbeiter	1.37706	3.96324	296.32376
a2 Beruf:Angestellte	-0.92524	0.39644	-60.35623
a3 Beruf:Selbständige	-0.45182	0.63647	-36.35343
X Einkommen	-0.37586	0.68670	-31.33039

Die Logit-Modell-Gleichung ist folgende:

$$(0) \quad p_1 = \frac{1}{1 + e^{-(c+a(i)+\beta \cdot x)}}$$

Man beachte:

Als Referenzkategorie wurde die erste Ausprägung "Nein" angegeben. Da die Zielvariable binar ist, ist somit die 2. Ausprägung "ja" die zu analysierende. p1 ist also die Wahrscheinlichkeit für die 2. Ausprägung "ja" der Zielvariablen.

In Prog22m kann der Benutzer entscheiden, ob er die erste oder die letzte Kategorie der nominalen Zielvariablen als Referenz einsetzt.

Obige Gleichung kann so umgewandelt werden, dass auf der rechten Seite ein linearer Ausdruck steht.

$$(1) \quad \ln(p_1/p_2) = c + a(i) + \beta X$$

p1=Wahrscheinlichkeit für Rückzahlung: ja
 p2=Wahrscheinlichkeit für Rückzahlung: nein
 Natürlich gilt: p2 = 1-p1
 c =Konstante

a(i) bezeichnet die Regressionskoeffizient für die drei

Dummy-Variable des Berufs (die den 3 Ausprägungen entsprechen)

es ist also:

a1=Regressionskoeffizient für "Arbeiter"
a2=Regressionskoeffizient für "Angestellter"
a3=Regressionskoeffizient für "Selbständiger"

X =Einkommen

β =Regressionskoeffizient für Einkommen

Für einen Arbeiter in der Einkommensklasse X=4 lautet also die Gleichung

$$(1a) \quad \ln(p_1/p_2) = c + a_1 + \beta X \\ = 1.88 + 1.38 - 0.38 \cdot 4$$

Regressionskoeffizienten der nominalen Variablen

Der Regressionskoeffizienten a1=1.37706 für "Arbeiter" und a2=-0.92524 für "Angestellter" haben folgende Bedeutung:

1. Das negative Vorzeichen von a2 drückt aus, dass Angestellte im Vergleich zur "Durchschnittsperson in der Variablen Beruf" das logarithmierte Wahrscheinlichkeitsverhältnis $\ln(p_1/p_2)$ aus Gleichung 1 verringern. Vereinfacht: Angestellte haben eine geringere Wahrscheinlichkeit ihren Kredit zurückzuzahlen. Umgekehrt drückt das positive Vorzeichen von a1 aus, dass Arbeiter eine erhöhte Wahrscheinlichkeit haben ihren Kredit zurück zu zahlen. Auf den Begriff "Durchschnittsperson" werden wir im nachfolgenden Unterabschnitt **Risiko exp(β)** zurückkommen.
2. Je (absolut) größer der Regressionskoeffizient ist, umso stärker ist diese Tendenz.

Regressionskoeffizienten der quantitativen Variablen

Der Regressionskoeffizient β1=-0.37586 für "Einkommen" hat folgende Bedeutung: Wenn sich das Einkommen um 1 Einheit erhöht, dann verringert sich das logarithmierte Wahrscheinlichkeitsverhältnis $\ln(p_1/p_2)$. Vereinfacht: Wenn sich das Einkommen um 1 Einheit erhöht, dann nimmt die Wahrscheinlichkeit ab, den Kredit zurückzuzahlen. Ein negatives Vorzeichen bedeutet, dass sich die Wahrscheinlichkeit verringert, ein positives, dass sie sich erhöht. Je (absolut) größer der Regressionskoeffizient ist, umso stärker ist diese Tendenz.

Gewinn-zu-Verlust-Verhältnis ("odds")

Gleichung 1 bzw. 1a kann so transformiert werden, dass der auf der linken Gleichungsseite stehende Logarithmus verschwindet.

$$(2) \quad p_1/p_2 = \exp(c) * \exp(a(i)) * \exp(\beta * X)$$

exp (...) = Exponentialfunktion von ...

Für unseren Arbeiter mit Einkommen X=4

$$(2a) \quad p_1/p_2 = \exp(c) \quad * \exp(a_1) \quad * \exp(\beta * X) \\ = \exp(1.88) \quad * \exp(1.38) \quad * \exp(-0.38 * 4) \\ = 6.62 \quad * 3.96 \quad * 0.22 \\ = 5.7886$$

Zuerst ist festzuhalten, dass sich die Interpretation auf die 2. Ausprägung der

Zielvariablen also auf "Rückzahlung: Ja" bezieht.

p1 ist also die Wahrscheinlichkeit für Rückzahlung: ja
p2 ist also die Wahrscheinlichkeit für Rückzahlung: nein

Das Wahrscheinlichkeits-Verhältnis $p1/p2$ wird in der angelsächsischen Literatur "odds" genannt.

Wenn man $p1$ als Gewinn-Wahrscheinlichkeit und $p2$ als Verlust-Wahrscheinlichkeit interpretiert, dann könnte man $p1/p2$ als "Gewinn-zu-Verlust-Verhältnis" bezeichnen.

Ist die Zielvariable, wie in unserem Beispiel, dichotom, dann gilt

$$p2 = 1-p1$$

Ist $p1=0.5$ dann ist $p2$ auch $=0.5$. Dann ist $p1/p2=1$. Das "Gewinn-zu-Verlust-Verhältnis" ist also ausgeglichen.

Ist $p1=0.6666..$ dann ist $p2=0.333333...$ Dann ist $p1/p2 =2$. Die Gewinn-Chance ist 2 mal besser als die Verlust-Chance

In unserem Beispiel ist $p1/p2=5.7886$. Für unseren Arbeiter mit einem Einkommen von 4 gilt also, dass seine Wahrscheinlichkeit den Kredit zurückzuzahlen 5.7886 mal größer ist als ihn nicht zurückzuzahlen.

Wie groß ist dann $p1$?

Hier gilt die allgemeine Formel:

$$\begin{aligned} p1 &= f / (1+f) \\ &= 5.7886 / (1+5.7886) \\ &= 0.853 \end{aligned}$$

wobei $f=p1/p2$

Die Wahrscheinlichkeit unseres Arbeiters mit Einkommen 4 den Kredit zurückzuzahlen ist also $p1=0.853$.

Betrachten wir einige Werte von $p1$

p1	dann ist p2= 1-p1	"Gewinn-zu-Verlust-Verhältnis" p1/p2
0.1	0.9	0.111
0.2	0.8	0.250
0.3	0.7	0.429
0.4	0.6	0.667
0.5	0.5	1
0.6	0.4	1.500
0.7	0.3	2.333
0.8	0.2	4
0.9	0.1	9

Risiko $\exp(\beta)$

Betrachten wir nun wieder Gleichung 2 bzw. 2a. Alle Arbeiter haben - im Vergleich zum Durchschnitt aller Untersuchungspersonen - eine um den Faktor $\exp(a1)=3.96$ erhöhtes Wahrscheinlichkeits-Verhältnis $p1/p2$, d.h. ihre Wahrscheinlichkeit den Kredit zurückzuzahlen ist erhöht.

Dieser Faktor wird in der Literatur gelegentlich "Risiko" genannt. Auch der Begriff "Effekt-Koeffizient" wird gelegentlich gebraucht (so bei D. Urban: Logit-Analyse,

1993, s. 40).

Wäre $\exp(a1)=1$, dann würden sich die Arbeiter so verhalten wie der Durchschnitt.

Wir definieren nun als

$$\text{relatives Risiko} = (\exp(a(i)) - 1) * 100$$

Für die Arbeiter finden wir dann

$$\begin{aligned} \text{relatives Risiko} &= (\exp(a1) - 1) * 100 \\ &= (3.96 - 1) * 100 \\ &= 296 \end{aligned}$$

Wir können jetzt formulieren: Arbeiter haben ein um 296 % höheres Risiko einen Kredit zurückzuzahlen als die durchschnittliche Untersuchungsperson.

Das relative Risiko kann auch negativ sein. Wäre es z.B. -74 %, dann würde das bedeuten, dass die Arbeiter ein um 74% reduziertes Risiko besitzen als die Referenzgruppe

"Durchschnittsperson"

Zu beachten ist, dass die Bezugskategorie der *Durchschnitt aller Untersuchungspersonen in der betreffenden nominalen unabhängigen Variablen* ist. Dies ist in Almo der Fall, wenn die 0,1,-1 - Kodierung der Dummies der unabhängigen nominalen Variablen verwendet wird. Dies ist die Voreinstellung in Almo (siehe Abschnitt P45.16.1.2, Box 9).

Wird die 0,1 - Kodierung verwendet, dann wird (standardmäßig) die letzte Dummy, in unserem Beispiel die Selbständigen, auf 0 gesetzt. Sie erscheint dann auch nicht in der Ergebnis-Ausgabe.

Almo liefert folgendes Ergebnis (verkürzt):

Ergebnisse für 2. Auspräg. "ja" der abhäng. Variablen "Rückzahlung"
(die 1. Ausprägung "nein" wird als Referenzkategorie verwendet)

unabhängige Variable	Regress. Koeffiz.	"Risiko" exp(Regr.-koeffiz.)	relatives Risiko
c Konstante	1.43044	-	-
a1 Beruf:Arbeiter	1.82889	6.22695	522.69462
a2 Beruf:Angestellte	-0.47341	0.62287	-37.71264
X Einkommen	-0.37586	0.68670	-31.33039

Die Selbständigen sind jetzt die Bezugskategorie. Die Arbeiter haben im Vergleich zu den Selbständigen eine um 522 % erhöhte Wahrscheinlichkeit den Kredit zurückzuzahlen und die Angestellten eine um 37.7 % reduzierte Wahrscheinlichkeit.

In Almo ist es bei der 0,1 - Kodierung möglich, entweder die erste oder die letzte Dummy zu eliminieren.

Allgemein gilt:

- a. Bei der 0,1 - Kodierung ist die Bezugskategorie die

- eliminierte Dummy.
- b. Bei der 0,1,-1 - Kodierung ist die Bezugskategorie der Durchschnitt aller Untersuchungspersonen.

Risiko bei quantitativen Variablen

Betrachten wir nochmals obige Gleichung (2)

$$(2) \quad p1/p2 = \exp(c) * \exp(a(i)) * \exp(\beta * X)$$

Das Einkommen unseres Arbeiters ist $X=4$.

Der Ausdruck $\exp(\beta * X)$ ist also $\exp(-0.37586 * 4) = 0.22236$

Wenn sich das Einkommen dieser Person um 1 Einheit auf 5 erhöht, dann ist der Ausdruck $\exp(\beta * X) = \exp(-0.37586 * 5) = 0.15270$

Wenn wir für $X=5$ obige Gleichung (2) für unsere Person ausrechnen, dann erhalten wir

$$p1/p2 = 3.9750$$

Für $X=4$ haben wir oben errechnet

$$p1/p2 = 5.7886$$

So hat sich also $p1/p2$ um den multiplikativen Faktor

$$3.9750 / 5.7886 = 0.68670$$

verringert. Und das ist genau das in obiger Tabelle angegebene Risiko $\exp(\beta)$.

Risiko-Werte unter 1 führen zu einer Verringerung von $p1/p2$. D.h. $p1$ wird kleiner und $p2$ wird größer.

Risiko-Werte über 1 führen zu einer Erhöhung von $p1/p2$. D.h. $p1$ wird größer und $p2$ wird kleiner.

Wir können nun den Begriff "Risiko" ($=\exp(\beta)$) bei ursächlichen quantitativen Variablen allgemein definieren.

Nimmt die ursächliche quantitative Variable X um 1 Einheit zu, dann nimmt das Wahrscheinlichkeits-Verhältnis $p1/p2$ um den multiplikativen Faktor $\exp(\beta)$ zu.

Wir können diese Zunahme bzw. Abnahme auch in Prozentwerten ausdrücken. Sie beträgt dann $100(\exp(\beta)-1)$. Das ist das relative Risiko.

Betrachten wir für Arbeiter die Werte, die sich gemäß Gleichung 2 für Einkommenswerte X von 0 bis 6 ergeben.

X	$p1/p2$	Multiplikator
0	26.0326	
1	17.8765	0.6867
2	12.2758	0.6867
3	8.4298	0.6867
4	5.7886	0.6867
5	3.9750	0.6867
6	2.7297	0.6867

Das Wahrscheinlichkeits-Verhältnis p_1/p_2 einer nachfolgenden Einkommensstufe entsteht durch Multiplikation mit $\exp(\beta) = 0.6867$ des Wahrscheinlichkeits-Verhältnis p_1/p_2 der vorhergehenden Einkommensstufe.

P22.1.3 Die Logitanalyse mit nominal-polytomer abhängiger Variablen

Die Logitanalyse ermöglicht im Unterschied zur Probitanalyse auch eine Analyse von polytomen nominalskalierten abhängigen Variablen. Diese Analyse wird *polytome Logitanalyse* oder *multinomiale Logitanalyse* oder auch *multinomiale logistische Regression* genannt.

Betrachten wir ein Beispiel: Die abhängige nominale Variable sei die Freizeitbeschäftigung mit 3 Ausprägungen. Das Programm zu diesem Beispiel findet man nach Klick auf den Knopf "alle Progs" am Oberrand des Almo-Fensters unter dem Namen "PolyLogit.Alm" oder auch "Prog22m4.Msk". Letzteres ermöglicht es, verschiedene Optionen auszuprobieren. Die im Beispiel verwendeten Variablen sind

Abhängige nominale Variable

Freizeit: (1) Sport, (2) Lesen, (3) Fernsehen

Unabhängige nominale Variable

Geschlecht: männlich, weiblich

Beruf: Arbeiter, Angestellter, Selbständiger

Unabhängige quantitative Variable

Leistung

Alter

Wir verwenden die letzte Ausprägung der abhängigen Variablen **Freizeit** als Referenzkategorie und rechnen die 2 binäre Analysen $\ln(p_1/p_3)$ $\ln(p_2/p_3)$

Die multinomiale Logitanalyse wird also in mehrere binäre Logitanalysen aufgelöst.

In den Almo-Programm-Masken Prog22m, Prog22m3, Prog22m4, Prog22m5 kann der Benutzer wählen, ob er die erste oder die letzte Ausprägung als Referenz verwendet. Dies sind Programme, die aufeinander folgende individuelle Datensätze einlesen.

Anders bei der Programm-Maske Prog22mb, die eine schon ausgezählte, fertige Tabelle einliest, dort wird die erste Ausprägung zwangsweise als Referenz verwendet. Der Benutzer kann das nicht beeinflussen.

Die unabhängigen nominalen Variablen **Geschlecht** und **Beruf** werden mit der 0,1,-1 -Kodierung in Dummies aufgelöst. Almo erlaubt auch die 0,1-Kodierung.

Almo liefert folgende Ergebnisse (stark gekürzt)

Ergebnisse fuer 1. Auspraegung "Sport" und 2. Auspraegung "Lesen" der abhaengigen Variablen V5 Freizeit (als Referenz wird die letzte Auspraegung "Fernsehen" verwendet)

	1. Auspraegung "Sport"		2. Auspraegung "Lesen"	
unabhängige Variab.	Regress.	relatives	Regress.	relatives

	koeffiz. ß	Risiko in %	koeffiz. ß	Risiko in %
A1 Geschlec:männlich	0.22137	24.77798	0.51684	67.67140
A2 Geschlec:weiblich	-0.22137	-19.85765	-0.51684	-40.35954
B1 Beruf:Arbeiter	0.03988	4.06873	-0.52000	-40.54773
B2 Beruf:Angestellt	0.04174	4.26271	0.54253	72.03455
B3 Beruf:Selbständ.	-0.08162	-7.83825	-0.02253	-2.22769
V3 Leistung	0.44503	56.05349	0.43082	53.85167
V4 Alter	-0.15873	-14.67739	-0.25324	-22.37200
Konstante	-0.98402	-	0.04463	-

Da die letzte Ausprägung der abhängigen Variablen als Referenzkategorie verwendet wurde, kann für sie kein Ergebnis ausgegeben werden.

Die Interpretation der Ergebnisse bei polytomer Logitanalyse

Bei der 0,1,-1 -Kodierung der unabhängigen nominalen Variablen muss die Interpretation generell so lauten:

Bei Personen mit der Ausprägung k in der unabhängigen Variablen i tritt die Ausprägung j der abhängigen Variablen im Vergleich zur Referenzausprägung häufiger/geringer/gleich oft wie bei der "Durchschnittsperson" auf.

Bei Personen mit der Ausprägung "Arbeiter" in der unabhängigen Variablen "Beruf" tritt die Ausprägung "Sport" der abhängigen Variablen "Freizeit" im Vergleich zur Referenzausprägung "Fernsehen" häufiger/geringer/gleich oft wie bei der "Durchschnittsperson" in der Variablen "Beruf" auf.

Das hört sich sehr kompliziert an - und ist es auch. Die inhaltliche Interpretation der Regressionskoeffizienten und des "Risikokoeffizienten" bei der multinomialen Logitanalyse ist kompliziert und für den in der Logitanalyse nicht Geübten verwirrend.

Beobachtete und modellreproduzierte (prognostizierte) Wahrscheinlichkeiten

die unabhangigen nominalen Variablen sind

A = V1 Geschlecht

B = V2 Beruf

die unabhangigen quantitativen Variablen sind

quant1 = V3 Leistung

quant2 = V4 Alter

rep1 ... = reproduzierte (prognostizierte) Wahrscheinlichkeit in Prozent
fur das Auftreten der Auspragung .. in der abhang. Variablen

Daten satz Nr.	A	B	quant1	quant2	tatsachliche Gruppen zugehor.	rep1	rep2	rep3	prognostizierte Gruppenzugehorig
1	1	2	4.000	4.000	1	17.639	70.820	11.540	2
2	1	1	5.000	3.000	1	34.908	52.582	12.510	2
3	1	1	4.000	2.000	1	31.681	53.202	15.117	2
4	1	1	2.000	1.000	1	25.704	48.811	25.485	2
5	1	1	4.000	3.000	2	32.393	49.492	18.116	2
6	1	1	4.000	5.000	2	32.971	41.700	25.329	2
7	1	1	2.000	2.000	2	25.708	44.417	29.874	2

.

Die prognostizierten Wahrscheinlichkeiten entstehen durch folgenden Kalkül. Wir zeigen, wie die prognostizierte Wahrscheinlichkeit $rep1=17.639\%$ bzw. $p1=0.17639$ der Gruppe 1 anzugehören für den 1. Datensatz berechnet wird.

**prognostizierte Wahrscheinlichkeit p1 für Ausprägung 1 (Freizeit: Sport)
 im Datensatz Nr 1**

Variable	Variablen- Wert	Regress. Koeffiz.	Koeffiz*Wert
Konstante		-0.98402	-0.98402
A1 Geschlec:männlich	1	0.22137	0.221366
A2 Geschlec:weiblich	0	-0.22137	0
B1 Beruf:Arbeiter	0	0.03988	0
B2 Beruf:Angestel	1	0.04174	0.0417436
B3 Beruf:Selbstän	0	-0.08162	0
V3 Leistung	4	0.44503	1.78012
V4 Alter	4	-0.15873	-0.634924
sum1			0.424278

$exp(sum1) = exp(0.424278) = 1.52849$ (für 1. Ausprägung)

Entsprechend wird auch für die 2. Ausprägung (Freizeit: Lesen) $sum2$ und $exp(sum2)$ berechnet. Es entsteht:

$sum2 = 1.8143$
 $exp(sum2) = exp(1.814300) = 6.13680$ (für 2. Ausprägung)

Die $exp(sum..)$ aller Ausprägungen, außer der letzten (der Referenzkaragorie) werden summiert

$sumexp = exp(sum1) + exp(sum2) = 1.52849 + 6.13680 = 7.66529$

Nunmehr können die vom Modell prognostizierten Wahrscheinlichkeiten bestimmt werden gemäß

$p1 = exp(sum1) / (1 + sumexp) = 1.52849 / (1 + 7.66529) = 0.176392$
 $p2 = exp(sum2) / (1 + sumexp) = 6.13680 / (1 + 7.66529) = 0.708205$

Die 3. Ausprägung ist die Referenzkategorie, für die keine Koeffizienten berechnet wurden. Die Wahrscheinlichkeit für sie ergibt sich residual so dass die Summe der drei p-Werte gleich 1.0 ist.

$p3 = 1 - (p1 + p2) = 1 - (0.176392 + 0.708205) = 0.115403$
 oder
 $p3 = 1 / (1 + sumexp) = 1 / (1 + 7.66529) = 0.115403$

Im Verlauf des Kalküls kann folgendes Problem entstehen: Wenn die Regressionskoeffizienten große Werte annehmen, dann können in obigen Gleichungen die Ausdrücke $exp(sum..)$ und vor allem $sumexp$ so große Werte annehmen, dass im Computer ein "overflow" auftritt oder zumindest eine beträchtliche Rechenungenauigkeit auftritt.

Vergleich mit SPSS

Wir haben unsere Beispieldaten auch mit der SPSS-Funktion "Regression/"

multinomial logistisch" gerechnet. Die Daten und das Syntaxprogramm sind unter dem Namen "PolyLogit.sav" und "PolyLogit.sps" im Almo-Ordner TESTDAT enthalten. SPSS löst die unabhängigen Variablen *Geschlecht* und *Beruf* mit der 0,1-Kodierung in Dummies auf, während Almo standardmäßig die 0,1,-1 Kodierung verwendet. Almo erlaubt aber auch im Programm *PolyLogit.Alm* oder in der Programm-Maske *Prog22m* die 0,1-Kodierung. Die Regressionskoeffizienten für die Dummies im SPSS-Output sind

	Ergebnis aus SPSS		Ergebnis aus Almo	
	Regressionskoeffizienten		mit 0,1 -Kodierung	
	1.Ausprägung Sport	2.Ausprägung Lesen	Sport	Lesen
Geschlecht=1]	0,443	1,034	0.44273	1.03367
Geschlecht=2]	0	0	-	-
Beruf=1]	0,122	-,497	0.12151	-0.49747
Beruf=2]	0,123	,565	0.12337	0.56505
Beruf=3]	0	0	-	-

Almo liefert bei 0,1 -Kodierung exakt die gleichen Ergebnisse.

Wird beim Almo-Ergebnis (bei der 0,1,-1-Kodierung) der Wert der letzten Dummy von den 2 bzw. 3 Dummies hinzu addiert, dann entstehen dieselben Werte wie bei SPSS. Die prognostizierten Wahrscheinlichkeiten (im SPSS-Datenfenster *Est1*, *Est2*, *Est3* genannt) und alle anderen (vergleichbaren) Koeffizienten sind gleich.

P22.1.4 Die Logit- und Probitanalyse mit ordinaler abhängiger Variablen

Beim multinomialen Logitmodell in Almo werden standardmäßig für jede Ausprägung der abhängigen Variablen - mit Ausnahme der letzten - Effekte berechnet. Beim ordinalen Logitmodell und beim ordinalen Probitmodell dagegen wird - wie bei den binären Modellen - für jede unabhängige Variable bzw. für jede Dummy nur *ein einzelner* Effekt hinsichtlich der abhängigen ordinalen Variablen berechnet. Zusätzlich werden "Schwellenwerte" für die 2. und die folgenden Antwortkategorien berechnet - nicht jedoch für die letzte, die Referenzkategorie. Sie werden in Almo mit *alfa2*, *alfa3*, ... bezeichnet. Siehe dazu Abschnitt P22.2.6.

Insgesamt stehen somit folgende Modelle zur Verfügung:

abhängige Variable	Logitmodell	Probitmodell
dichotom	binäres Logitmodell	binäres Probitmodell
polytom, nominalskaliert	multinomiales Logitmodell	nicht definiert
polytom, ordinalskaliert	ordinales Logitmodell ("ordered-reponse logit model", "ordinal-level logit model")	ordinales Probitmodell ("ordered-reponse probit mode", "ordinal-level probit model")

P22.1.5 Die Maximum-Likelihood-Schätzung

Die Parameter b_0 , b_{11} , b_{12} usw. der Logit- oder Probitmodelle können mit Hilfe von zwei Methoden geschätzt werden: der gewichteten Kleinste-Quadrate-Methode (siehe Abschnitt P20.9.3.3) und der Maximum-Likelihood-Methode. Programm P22 enthält die Maximum-Likelihood-Methode. Die Parameter werden so geschätzt, dass

die Likelihoodfunktion

$$(4) \quad L = \prod_{i=1}^n \binom{n_i}{n_{i2}} \cdot \hat{p}_{i2}^{n_{i2}} \cdot \hat{p}_{i1}^{n_{i1}} = \prod_{i=1}^n \binom{n_i}{n_{i2}} \cdot \hat{p}_{i2}^{n_{i2}} \cdot (1 - \hat{p}_{i2})^{n_i - n_{i2}}$$

ein Maximum wird. Die Größen bedeuten:

n_i	Häufigkeit des Auftretens des Datensatzes i
n_{i2}	Häufigkeit des Auftretens der Ausprägung 2 der abhängigen Variablen im Datensatz i
n_{i1}	Häufigkeit des Auftretens der Ausprägung 1 der abhängigen Variablen im Datensatz i
\hat{p}_{i2}	durch die Logit- oder Probitanalyse prognostizierte Wahrscheinlichkeit des Auftretens der Ausprägung 2 der abhängigen Variablen für den Datensatz i
\hat{p}_{i1}	durch die Logit- oder Probitanalyse prognostizierte Wahrscheinlichkeit des Auftretens der Ausprägung 1 der abhängigen Variablen für den Datensatz i

Für $i=1$ ist $n_1 = 17596$, $n_{12} = 10128$ und $n_{11} = 7468$.

Die prognostizierten Wahrscheinlichkeiten sind:

(a) für das Logitmodell:

$$(5) \quad \hat{p}_{i2} = \frac{e^{b_0 + b_{11} \cdot x_{i11} + b_{12} \cdot x_{i12} + b_{21} \cdot x_{i21}}}{1 + e^{b_0 + b_{11} \cdot x_{i11} + b_{12} \cdot x_{i12} + b_{21} \cdot x_{i21}}}$$

und

$$(6) \quad \hat{p}_{i1} = \frac{1}{1 + e^{b_0 + b_{11} \cdot x_{i11} + b_{12} \cdot x_{i12} + b_{21} \cdot x_{i21}}},$$

wobei die geschätzten Parameter b_0 , b_{12} , b_{13} eingesetzt werden (siehe dazu ausführlicher Abschnitt P22.2).

(b) für das Probitmodell:

$$(7) \quad \hat{p}_{i2} = \Phi(1 + b_0 + b_{11} \cdot x_{i11} + b_{12} \cdot x_{i12} + b_{21} \cdot x_{i21})$$

und

$$(8) \quad \hat{p}_{i1} = 1 - \hat{p}_{i2} = 1 - \Phi(1 + b_0 + b_{11} \cdot x_{i11} + b_{12} \cdot x_{i12} + b_{21} \cdot x_{i21}).$$

Die Maximum-Likelihood-Schätzung - kann im Unterschied zur gewichteten Kleinste-Quadrate-Schätzung - auch für Individualdaten angewendet werden. In diesem Fall ist n_i gleich 1 und die zu maximierende Likelihoodfunktion vereinfacht sich zu

$$(9) \quad L = \prod_{i=1}^n \hat{p}_{i2}^{y_{i2}} \cdot \hat{p}_{i1}^{y_{i1}} = \prod_{i=1}^n \hat{p}_{i2}^{y_{i2}} \cdot (1 - \hat{p}_{i2})^{1 - y_{i2}},$$

wobei y_{i2} gleich 1 ist, wenn die Person i (Datensatz i) in der abhängigen Variablen die Ausprägung 2 besitzt. In den anderen Fällen ist y_{i2} gleich 0. y_{i1} ist analog

definiert: y_{i1} ist 1, wenn die Person i in der abhängigen Variablen die Ausprägung 1 besitzt, sonst ist $y_{i1}=0$.

Die beiden Schätzgleichungen (4) und (9) können sowohl für die Probit- als auch für die Logitanalyse verwendet werden. Die Schätzung der Parameter erfolgt iterativ über den sogenannten Newton-Raphson-Algorithmus.

P22.2 Eingabe und Ausgabe

P22.2.1 Eingabe mit Programm-Maske

Es sind 2 Programm-Masken vorhanden:

1. Eine Programm-Maske zur Eingabe von Individualdaten: Prog22m
2. Eine Programm-Maske zur Eingabe einer fertigen Tabelle mit gruppierten Daten: Prog22mb

Der Benutzer erreicht diese Programm-Masken durch Klick auf den Knopf „Verfahren“ (in der Knopfleiste unterhalb der Menüleiste), danach Klick auf „Logit-, Probitanalyse“

P22.2.1.1 Programm-Maske zur Eingabe von Individualdaten Prog22m

6 Wenn Dateiformat FIX oder Nicht-Standard-FREI

7 Analyse-Variable: Unabhängige Variable
Unabhängige quantitative Variable
 Verdienst
Unabhängige nominale Variable
 Geschlecht, Beruf

8 Option: Auflösung der unabhäng. nominalen Variab. in Dummies

9 Analyse-Variable: Abhängige Variable
Erlaubt ist:
1. Eine nominale Variable mit beliebig vielen Ausprägungen
oder (exklusiv)
2. Eine ordinale Variable
abhängige nominale Variable
 Kauf
abhängige ordinale Variable
BEACHTÉ: Die abhängige ordinale Variable muß ganzzahlig, mit Schrittweite 1 kodiert sein. Ist dies nicht der Fall, dann muß sie in der Umkodierungsbox umkodiert werden
 ■

10 Logit oder Probit

11 Option: Ein- und Ausschliessen von Untersuchungseinheiten

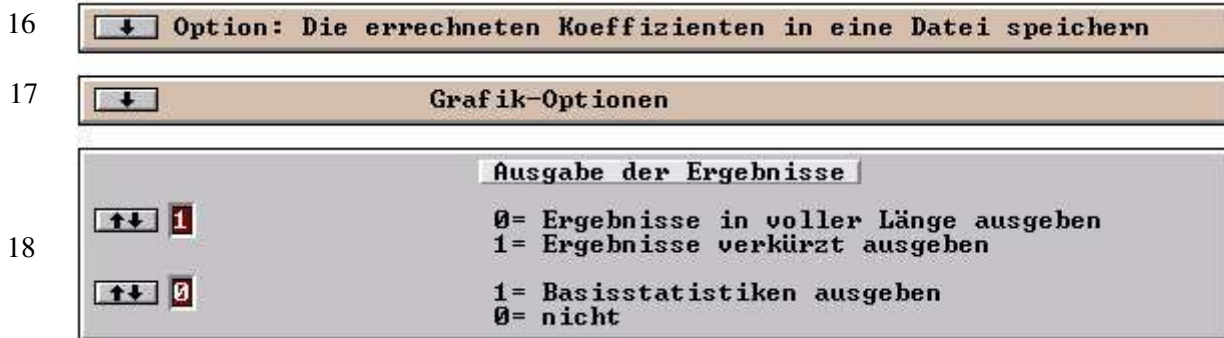
12 Option: Umkodierungen und Kein-Wert-Angaben

13 Option: Untersuchungseinheiten ganzzahlig gewichten

14 Option: Prognosewerte ermitteln

15 Option: Wertemuster

10a nachträglich eingeführt



P22.2.1.2 Erläuterungen zu den Eingabeboxen

Die meisten Boxen werden im Almo-Handbuch **PO Arbeiten mit Progs** ausführlich erläutert. Sie werden hier nicht nochmals behandelt.

Box 1: Vereinbare Variable

Siehe Almo-Dokument Nr. 0 "Arbeiten mit Almo", Abschnitt PO.1.

Box 2: Option: Weitere Vereinbarungen - nur wenn Almo dazu auffordert

Siehe Almo-Dokument Nr. 0 "Arbeiten mit Almo", Abschnitt PO.2.

Box 3: Datei der Variablennamen

Siehe Almo-Dokument Nr. 0 "Arbeiten mit Almo", Abschnitt PO.3.

Box 4: Freie Namensfelder

Siehe Almo-Dokument Nr. 0 "Arbeiten mit Almo", Abschnitt PO.3.

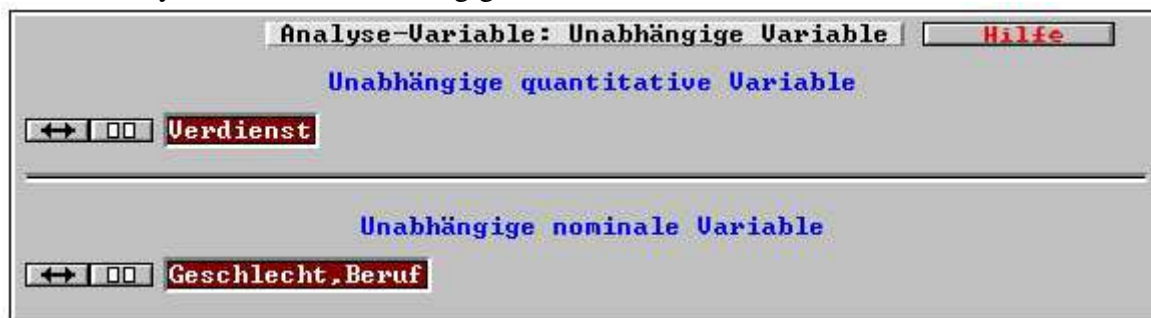
Box 5: Datei aus der gelesen wird

Siehe Almo-Dokument Nr. 0 "Arbeiten mit Almo", Abschnitt PO.4.

Box 6: Wenn Dateiformat FIX oder nicht Standard-FREI

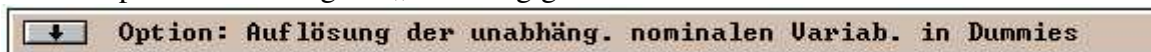
Siehe Almo-Dokument Nr. 0 "Arbeiten mit Almo", Abschnitt PO.4.

Box 7: Analyse-Variable: Unabhängige Variable

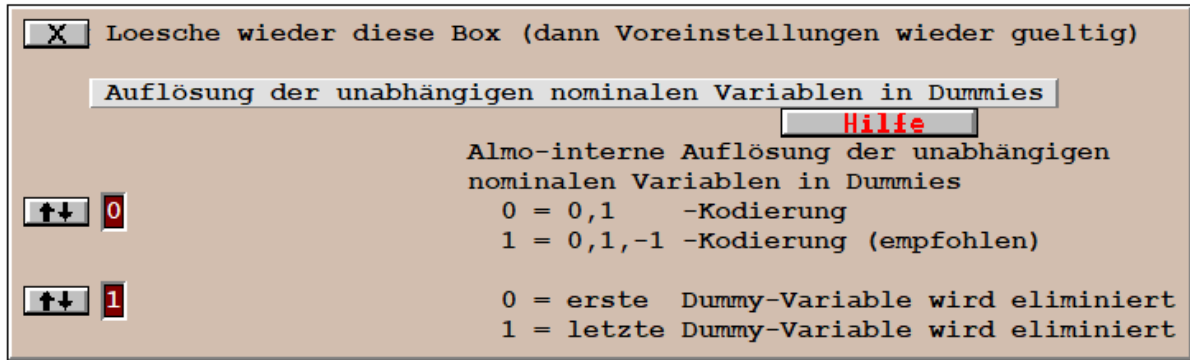


Die unabhängigen quantitativen und nominalen Variablen werden mitgeteilt. Wenn sie auf den Knopf mit den zwei kleinen Fenstern klicken, dann wird Ihnen eine Dialogbox zur Auswahl der Variablen präsentiert.

Box 8: Option: Auflösung der „unabhängigen“ nominale Variablen in Dummies



Optionsbox geöffnet:



Der Benutzer kann wählen,

1. ob er die 0,1 -Kodierung oder die 0,1,-1 -Kodierung verwenden möchte (siehe dazu Almo-Dokument 13a, Abschnitt P20.3 und P22)
2. ob er die erste oder letzte Dummy eliminieren möchte. Dies ist nur bei der 0,1 -Kodierung möglich. Bei der 0,1,-1 -Kodierung ist dies gleichgültig.
3. Ist die Optionsbox geschlossen, dann ist die 0,1,-1 -Kodierung voreingestellt.

Bei der 0,1 -Kodierung wird der Regressionskoeffizient der eliminierten Dummy auf .0 gesetzt. Bei der 0,1,-1 -Kodierung erhält jede Dummy einen Regressionskoeffizienten zugeordnet. Die Koeffizienten summieren sich dann zu .0. Wir empfehlen die 0,1,-1 -Kodierung, da hier für alle Dummies Koeffizienten (Beta, Standardfehler etc.) berechnet werden können.

Die 0,1 -Kodierung wird in der Literatur auch "Indikator-Kodierung" (bei SPSS: indicator-contrast) und die 0,1,-1 -Kodierung "Effekt-Kodierung" (bei SPSS: deviation-contrast) genannt.

SPSS verwendet in seiner „multinomialen Logit-Regression“ die 0,1 -Kodierung und eliminiert die letzte Dummy. Sollen die Ergebnisse aus Almo mit denen von SPSS verglichen werden, dann sollte der Benutzer die Optionen in der Box entsprechend einstellen: Im ersten Eingabefeld „0“ und im zweiten „1“.

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

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

Unterschiedliche Ergebnisse bei unterschiedlicher Kodierung.

Die Regressionskoeffizienten der Dummies und nicht nur diese, die aus diesen drei Kodierungsarten hervorgehen, sind verschieden.

Betrachten wir ein Beispiel. In unserem Bootstrap-Programm Prog22m5 ist die

abhängige nominale Variable

Wohnlage: (1) Land, 2) Stadtrand, (3) Stadt

Die unabhängigen nominalen Variablen sind

Geschlecht: (1) männlich, (2) weiblich

soziale Herkunft: (1) Unterschicht, (2) Mittelschicht, (3) Oberschicht

Die unabhängigen quantitativen Variablen (die hier im Zusammenhang mit den unterschiedlichen Kodierungsarten nicht interessieren) sind

Alter, Bildungsniveau

Wir erhalten folgende drei verschiedene Ergebnisse für die 1. Ausprägung "Land" der abhängigen Variablen "Wohnlage" aus den drei Kodierungsarten. Ausgabe gekürzt.

Ergebnisse fuer 1. Auspraegung der abaeng. Var. "Land"

unabhaengige Variab	Regress. koefiz.β	"Risiko" exp(β)	Stand. Fehler	Wald z*z	Signif. p	partielle Korrelat.

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

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

Hingegen sind die paarweisen Vergleich aus den 3 Analysen voll identisch

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

Vergleichs- paar	Differenz	"Risiko" exp(Differenz)	Stand.- Fehler	z-Wert	Signifikanz p	Signifikanz (1-p)*100
0,1 -Kodierung letzte Dummy-Variable wird eliminiert						
A1 - A2	-2.1250	0.1194	0.5241	4.055	0.000	100.00
B1 - B2	1.2611	3.5292	0.4990	2.527	0.012	98.84
B1 - B3	2.0471	7.7450	0.5632	3.635	0.000	99.97
B2 - B3	0.7860	2.1945	0.3710	2.119	0.034	96.58

Paarweise Vergleiche (Kontraste) der unabhaengigen nominalen Variablen fuer 2.Auspraegung "Stadttrand" der abhaeng. Var. V4 Wohnlage

0,1 -Kodierung erste Dummy-Variable wird eliminiert						
A1 - A2	-2.1250	0.1194	0.5241	4.055	0.000	100.00
B1 - B2	1.2611	3.5292	0.4990	2.527	0.012	98.84
B1 - B3	2.0471	7.7450	0.5632	3.635	0.000	99.97
B2 - B3	0.7860	2.1945	0.3710	2.119	0.034	96.58

Paarweise Vergleiche (Kontraste) der unabhaengigen nominalen Variablen fuer 3.Auspraegung "Stadt" der abhaeng. Var. V4 Wohnlage

0,1,-1 -Kodierung						
A1 - A2	-2.1250	0.1194	0.5241	4.055	0.000	100.00
B1 - B2	1.2611	3.5292	0.4990	2.527	0.012	98.84
B1 - B3	2.0471	7.7450	0.5632	3.635	0.000	99.97
B2 - B3	0.7860	2.1945	0.3710	2.119	0.034	96.58

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

Umrechnung der Regressionskoeffizienten.

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

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

0,1 -Kodierung letzte Dummy eliminiert		0,1 -Kodierung erste Dummy eliminiert	
unabhaengige Variab	Regress. koeffiz. ß	unabhaengige Variab	Regress. koeffiz. ß
Konstante	14.23315	Konstante	14.15518

A1 Geschlecht: männl	-2.12502	A2 Geschlecht: weibl	2.12502
B1 Herkunft: Untersch	2.04705	B2 Herkunft: Mittelsc	-1.26108
B2 Herkunft: Mittelsc	0.78598	B3 Herkunft: Oberschi	-2.04705
V11 Alter	-0.40542	V11 Alter	-0.40542
V1 Bildungsniveau	0.33416	V1 Bildungsniveau	0.33416

0,1,-1 -Kodierung
alle Dummies einbezogen

unabhaengige Variab	Regress. koeffiz. B

Konstante	14.11498
A1 Geschlecht: männl	-1.06251
A2 Geschlecht: weibl	1.06251
B1 Herkunft: Untersch	1.10271
B2 Herkunft: Mittelsc	-0.15837
B3 Herkunft: Oberschi	-0.94434
V11 Alter	-0.40542
V1 Bildungsniveau	0.33416

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

Notation:

Mit dem Begriff "Dummy-Wert" oder noch kürzer nur "Dummy" oder "Effekt" ist immer der Regressionskoeffizient der betreffenden Dummy gemeint

a_{letzt} = letzter Dummy-Wert aus der 0,1,-1 -Kodierung

a_{erst} = erster Dummy-Wert aus der 0,1,-1 -Kodierung

$a_1, a_2, \dots, a_i, \dots, a_{\text{letzt}-1}$ = die Dummies aus der 0,1,-1 -Kodierung

$b_1, b_2, \dots, b_i, \dots, b_{\text{letzt}-1}$ = die anderen Dummies aus der 0,1-Kodierung mit letzter eliminiertes Dummy

$c_2, c_3, \dots, c_i, \dots, c_{\text{letzt}}$ = die anderen Dummies aus der 0,1-Kodierung mit erster eliminiertes Dummy

Die 0,1-kodierten Dummies b_i bei letzter eliminiertes Dummy bzw c_i bei erster eliminiertes Variablen entstehen dann aus

0,1-kodiert, *letzte* Dummy eliminiert: $b_i = a_i - a_{\text{letzt}}$

0,1-kodiert, *erste* Dummy eliminiert: $c_i = a_i - a_{\text{erst}}$

Wird der *letzte* Dummy-Wert a_{letzt} subtrahiert von a_i , dann entsteht der Dummy-Wert b_i

Beispiel: Es soll der Wert der Dummy a_1 "Unterschicht" umgerechnet werden von der 0,1,-1-Kodierung auf die entsprechende Dummy b_1 aus der 0,1-Kodierung mit letzter eliminiertes Dummy. Wie entsteht $b_1=2.04705$ aus $a_1=1.10271$

$$a_1 - a_{\text{letzt}} = b_1$$

$$1.10271 - -0.94434 = 2.04705$$

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

$$\begin{array}{rcl} a_2 & - & a_{\text{erst}} = c_2 \\ -0.15837 & - & 1.10271 = -1.26108 \end{array}$$

Interpretation.

Wie müssen die Effekte a_i bzw. b_i bzw. c_i interpretiert werden ?

Soll beispielsweise der Effekt, den die „Mittelschicht“ auf die abhängige Variable „Wohnlage: Land“ ausübt, interpretiert werden, dann muss

1. bei der 0,1--Kodierung mit *letzter* eliminiertes Dummy wird die "Oberschicht" zur Referenzkategorie. Es muss formuliert werden: Im Vergleich zur Oberschicht ist der Effekt der Mittelschicht auf die „Wohnlage: Land“ stärker (positives Vorzeichen), aber weniger stark als bei der Unterschicht.
2. bei der bei der 0,1--Kodierung mit *erster* eliminiertes Dummy wird die "Unterschicht" zur Referenzgruppe. Im Vergleich zu ihr, ist nun der Effekt der "Mittelschicht" kleiner (negatives Vorzeichen) aber größer als der der Oberschicht.
3. bei der 0,1,-1 -Kodierung entstehen für alle 3 sozialen Schichten Regressionskoeffizienten. Ihre Summe ist 0. Das ist dann auch der *Durchschnitt* aus den 3 Regressionskoeffizienten. Betrachten wir die 3 β -Koeffizienten der Ausprägungen von "Herkunft"

B1 Herkunft:Untersch	1.10271
B2 Herkunft:Mittelsc	-0.15837
B3 Herkunft:Oberschi	-0.94434
Summe	0

Ist die Summe =0, dann ist selbstverständlich auch der Durchschnitt=0.

Der Effekt einer einzelnen sozialen Schicht wird auf 0 bzw. auf den Durchschnitt bezogen. Wir können formulieren: Der Effekt der "Mittelschicht" auf die abhängige Variable „Wohnlage: Land“ ist mit **-0.15837** geringer als der durchschnittliche Effekt der Variablen "Herkunft".

(Der Regressionskoeffizient von "Oberschicht" ist mit **-0.94434** noch weiter negativ von 0, bzw. vom Durchschnitt entfernt.)

Bei allen 3 Kodierungsmethoden sind die 3 Ausprägungen *innerhalb* ihrer Kodierungsmethode miteinander vergleichbar, aber nicht *zwischen* den Kodierungsmethoden, da ihre Regressionskoeffizienten auf verschiedene Referenzen bezogen sind. Die paarweisen Differenzen zwischen den Ausprägungen sind jedoch, wie wir oben gezeigt haben auch zwischen den Kodierungsmethoden gleich.

Box 9: Analyse-Variable: Abhängige Variable

Analyse-Variable: Abhängige Variable
Hilfe

BEACHTEN: Die abhängige nominale oder ordinale Variable muß ganzzahlig und mit Schrittweite 1 kodiert sein. Ist sie das nicht, dann muß sie in der Umkodierungsbox in spezieller Weise umkodiert werden Hilfe

abhängige nominale Variable

↔

□□

Kauf

↑↓

2

Referenz-Ausprägung

1= 1. Ausprägung wird Referenz

2= letzte Ausprägung wird Referenz

Hilfe

abhängige ordinale Variable

↔

□□

Die abhängige Variable wird mitgeteilt. Sie kann entweder eine nominale oder (exklusiv) eine ordinale Variable sein. Wenn sie auf den Knopf mit den zwei kleinen Fenstern klicken, dann wird Ihnen eine Dialogbox zur Auswahl der Variablen präsentiert.

Spezielle Umkodierung der abhängigen nominalen oder ordinalen Variablen

Gelegentlich besitzt die abhängige Variable sehr viele Ausprägungen, so dass es notwendig wird deren Zahl zu verringern. Das geschieht dadurch dass Ausprägungen zusammengefasst werden.

Betrachten wir ein Beispiel:

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

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

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

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

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

Folgende Änderungen müssen in der Programm-Maske vorgenommen werden

1. Die neue, freie Variable 14 erhält den Namen

Name 14 = ArbeitsbelastungII: (1)wenig, (2)viel;

Das geschieht in der Eingabebox "Freie Namensfelder"

Freie Namensfelder Hilfe

Leere alle Eingabefelder dieser Sub-Box

Name14 = ArbeitsbelastungII: (1)gering, (2)hoch;

;

erzeuge zusätzliche Namensfelder

2. Als abhängige nominale Variable wird nun eingesetzt: **ArbeitsbelastungII**

Das geschieht in der Eingabebox "Abhängige Variable" im Eingabefeld "nominale Variable"

abhängige nominale Variable

ArbeitsbelastungII

als Referenz wird die letzte Ausprägung der abhängigen nominalen Variablen verwendet Hilfe

3. Die Optionsbox "Umkodierungen und Kein-Wert-Angaben" wird geöffnet.

In eines der Eingabefelder wird eingetragen

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

Semikolon zum Schluss nicht vergessen !

Loesche wieder diese Sub-Box (Voreinstellungen wieder gueltig)

Eingabefelder für Umkodierungen und Kein-Wert-Angaben Hilfe

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

erzeuge zusätzliche Felder für Umkodierungen / Kein_Wert-Angaben

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

Box 10: Modell

Modell

Logit

Logit oder Probit

Sie entscheiden, ob Sie ein Logit- oder Probit-Modell rechnen wollen.

Abhängig vom Messniveau und der Zahl der Ausprägungen der abhängigen Variablen können folgende Modelle gerechnet werden:

abhängige Variable	Logitanalyse	Probitanalyse
dichotom (ordinal oder nominal)	binäres Logitmodell	binäres Probitmodell
polytom-nominal	multinomiales Logitmodell	nicht möglich
polytom-ordinal	ordinales Logitmodell	ordinales Probitmodell

Box 10a: Option: Grenzwerte für Modell

Option: Grenzwerte für Modell

Optionsbox geöffnet:

Loesche wieder diese Box (dann Voreinstellungen wieder gueltig)

Option: Grenzwerte für Modell
Hilfe

Grenzwert für Konvergenz
 eine 1 an der x-ten Dezimalstelle
 Voreinstellung: 4 (=0.0001)

Grenzwert für Verbesserung
 eine 1 an der x-ten Dezimalstelle
 Voreinstellung: 9 (=0.000000001)

Zahl der maximalen Iterationen
 Voreinstellung: 20

1= Iterationsprotokoll aus Newton-Raphson-Algorithmus ausgeben Hilfe
 0= nicht (Voreinstellung)

wenn keine Konvergenz erreicht wird
 und wenn letzte Iteration eine
 negative Verbesserung erbringt ...
 gültig nur für Analyse ohne Bootstrap

1 = ... dann die Lösung der vorletzten
 Iteration verwenden
 0 = ... dann letzte Lösung belassen
 (Voreinstellung)

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

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

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

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

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

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

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

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

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

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

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

1. Eingabefeld: Die "Konvergenz"

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

2. Eingabefeld: Die "Verbesserung"

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

Nicht selten tritt der Fall auf, dass die Verbesserung negativ ist. Ist dann der Konvergenzwert (genauer die 1. Ableitung der Likelihoodfunktion) noch weit vom angestrebten Grenzwert entfernt, dann kann dies ein Hinweis sein, dass die vorhandenen Daten nicht durch das ML-Modell der Logit- bzw. Probitanalyse analysierbar sind.

Möglich wäre es auch den "Likelihood-Ratio-Wert" zu verwenden um die Verbesserung zu bestimmen, wie dies in SPSS in der logistischen Regression getan wird. Dieser Koeffizient wird dort mit "-2 Log-Likelihood" bezeichnet. Er ist -2 mal dem Log-ML-Wert, mit dem Almo die Verbesserung berechnet. Almo gibt diesen Wert im Iterationsprotokoll aus, verwendet ihn jedoch nicht.

3. Eingabefeld: Die maximale Iterationszahl

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

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

4. Eingabefeld: Das Iterationsprotokoll

Almo gibt für das in Prog22m5 gerechnete Beispiel folgendes Iterationsprotokoll aus

Iterationsprotokoll

it	Konvergenz 1.Ableitung	Log-ML-Wert	Verbesserung	Likelihood-Ratio Testgrosse
0	1.48700e+003	-4.99869e+002	1.00000e+070	9.99737e+002
.....				
Reg.koeff.				
	(1)	0.00000	0.00000	0.00000
		0.00000	0.00000	
	(2)	0.00000	0.00000	0.00000
		0.00000	0.00000	
1	6.18239e+002	-4.22587e+002	7.72818e+001	8.45174e+002
.....				
Reg.koeff.				
	(1)	9.02359	-1.38361	1.22257
		-0.25303	0.18676	0.59236
	(2)	4.56202	-0.50697	0.43537
		-0.13226	0.14284	0.48854
2	9.44194e+001	-4.12828e+002	9.75903e+000	8.25656e+002
.....				
Reg.koeff.				
	(1)	13.13676	-1.96149	1.69262
		-0.37285	0.29915	0.75903
	(2)	7.67440	-0.96215	1.00683
		-0.21654	0.22341	0.46069
3	8.85301e+000	-4.12113e+002	7.14767e-001	8.24226e+002
.....				
Reg.koeff.				
	(1)	14.16195	-2.11357	1.99194
		-0.40326	0.33156	0.78491
	(2)	8.45768	-1.07338	1.26392
		-0.23940	0.24976	0.46884
4	2.15784e-001	-4.12101e+002	1.17675e-002	8.24202e+002
.....				
Reg.koeff.				
	(1)	14.23259	-2.12491	2.04564
		-0.40541	0.33413	0.78597
	(2)	8.51477	-1.08262	1.31732
		-0.24111	0.25190	0.46945
5	1.28461e-004	-4.12101e+002	6.62117e-006	8.24202e+002
.....				
Reg.koeff.				
	(1)	14.23315	-2.12502	2.04705
		-0.40542	0.33416	0.78598
	(2)	8.51525	-1.08272	1.31873
		-0.24112	0.25192	0.46945
6	5.38005e-011	-4.12101e+002	3.29692e-012	8.24202e+002
.....				
Reg.koeff.				
	(1)	14.23315	-2.12502	2.04705
		-0.40542	0.33416	0.78598
	(2)	8.51525	-1.08272	1.31873
		-0.24112	0.25192	0.46945

Betrachten wir Iterationsschritt 5

it	Konvergenz 1.Ableitung	Log-ML-Wert	Verbesserung	Likelihood-Ratio Testgrosse
----	---------------------------	-------------	--------------	--------------------------------

```

=====
.
.
-----
5    1.28461e-004  -4.12101e+002   6.62117e-006   8.24202e+002
.....
Reg.koeff.
(1)   14.23315   -2.12502    2.04705    0.78598
      -0.40542    0.33416
(2)   8.51525   -1.08272    1.31873    0.46945
      -0.24112    0.25192
-----
.
.
-----

```

Zuerst wird die Konvergenz ausgegeben. Mit 0.000128461 ist sie noch minimal größer als der angestrebte Grenzwert von 0.0001 . Die Verbesserung hat mit $6.62117e-006$ noch nicht den Grenzwert von $1.00000e-009$ unterschritten. Es muss also ein weiterer Iterationsschritt gerechnet werden. Beim 6. Schritt wird dann Konvergenz erzielt.

In der ersten Zahlenreihe wird noch der Log-Maximum-Likelihood-Wert und die Likelihood-Ratio-Testgröße ausgegeben. Sie ist gleich dem $-2 \cdot \text{Log-ML}$ -Wert und könnte somit auch dazu verwendet werden die Verbesserung zu ermitteln. Beides sind Maßzahlen für die Güte des Modells.

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

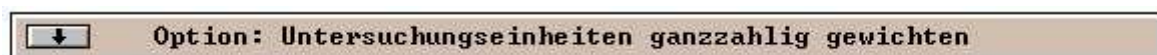
Vergleicht man die aufeinander folgenden 6 Werte für die Konvergenz und für die Verbesserung, dann erkennt man, dass sie sich allmählich den Grenzwerten nähern und im 6. Schritt diese unterschreiten.

Iteration	Konvergenz	
	1. Ableitung	Verbesserung
1	6.18239e+002	7.72818e+001
2	9.44194e+001	9.75903e+000
3	8.85301e+000	7.14767e-001
4	2.15784e-001	1.17675e-002
5	1.28461e-004	6.62117e-006
6	5.38005e-011	3.29692e-012
Grenzwert	1.00000e-004	1.00000e-009

Box 11: Option: Ein- und Ausschliessen von Untersuchungseinheiten
 Siehe Almo-Dokument Nr. 0 "Arbeiten mit Almo", Abschnitt P0.7.

Box 12: Umkodierung und Kein_Wert-Angabe
 Siehe Almo-Dokument Nr. 0 "Arbeiten mit Almo", Abschnitt P0.5.

Box 13: Option: Untersuchungseinheiten ganzzahlig gewichten



Optionsbox geöffnet:



Als Gewichtungsvariable wird die Almo-Variable

GEWICHT0

verwendet. Ihr muss ein ganzzahliger Wert zugewiesen werden. Eine Gewichtung mit einer Dezimalzahl ist nicht zulässig!

Wenn Sie z.B. in das Eingabefeld schreiben

```
Wenn V1 gleich 1 dann GEWICHT0 = 2; EndeWenn
Wenn V1 gleich 2 dann GEWICHT0 = 3; EndeWenn
```

dann wird eine Untersuchungseinheit, die in V1 den Wert 1 besitzt, so behandelt, wie wenn sie aus 2 Untersuchungseinheiten bestehen würde und eine Untersuchungseinheit, die in V1 den Wert 2 besitzt, so wie wenn sie aus 3 Untersuchungseinheiten bestehen würde.

Eine Gewichtung mit einer Dezimalzahl, z.B. 1.5 ist nicht zulässig. Folgende Anweisung wäre also falsch:

```
Wenn V1 gleich 1 dann GEWICHT0 = 1.5; EndeWenn Falsch !!!
```

In obigem korrekten Beispiel wurde (wenn V1=1) mit 2 und (wenn V1=2) mit 3 gewichtet. Das bedeutet, dass Untersuchungsobjekte mit V1=2 um den Faktor 1.5 höher gewichtet wurden als Untersuchungsobjekte mit V1=1. Auf diese indirekte Weise ist also doch eine Gewichtung mit einem Dezimalfaktor möglich.

BEACHTE:

Die Zahl der "eingeleseenen" Untersuchungseinheiten, die Almo mitteilt, ist entsprechend höher. Das wirkt sich auch bei Signifikanztests aus. Die Zahl der (auf n beruhenden) Freiheitsgrade ist dann höher - wodurch die Signifikanz überschätzt wird. Wenn der Benutzer diesen Effekt nicht wünscht, dann kann er - "von Hand" - den von Almo ausgegebenen t- bzw. F-Wert gegen die tatsächliche Zahl der Untersuchungseinheiten testen. Klicken Sie zu diesem Zweck auf das Menü

Statistik / Statistische Tafeln

und wählen Sie den zutreffenden Test.

BEACHTE: Wird (in unserem Beispiel) V1 umkodiert, so geschieht das im Rechenverlauf vor der Gewichtung. Wenn Almo die Gewichtungsbedingung einliest und verarbeitet, dann sind die Variablen bereits umkodiert.

Wenn Sie keine Gewichtung verwenden wollen, dann lassen Sie die Optionsbox geschlossen oder machen Sie das Eingabefeld leer.

Wollen Sie jedoch eine Gewichtung vornehmen, dann können Sie alle Fälle gleich gewichten, was nicht viel Sinn macht. Sie schreiben z.B. in das Eingabefeld

```
GEWICHT0 = 2;
```

Dadurch wird jede einzelne Untersuchungseinheit 2 mal in die Analyse

aufgenommen.

Gelegentlich befindet sich in der Datei eine Variable, die Gewichtungszahlen enthält - z.B. V48. In diesem Fall schreiben Sie in das Eingabefeld

```
GEWICHT0 = V48 (Runde 1);
```

Dann werden alle Analysevariable des Datensatzes mit der Zahl gewichtet, die in V48 gefunden wird (die von ALMO zuerst auf Ganzzahligkeit gerundet wird).

Sehr häufig wird man spezifische Untergruppen gewichten wollen. Sie können z.B. alle Männer doppelt zählen (!!). Sie schreiben dann

```
Wenn Geschlecht gleich 1 dann GEWICHT0 = 2; EndeWenn
```

(Geschlecht=1 sei der Code für die Männer)

Oder ein anderes Beispiel:

```
Wenn Beruf gleich 1 dann GEWICHT0 = 2; EndeWenn
```

```
Wenn Beruf gleich 2 dann GEWICHT0 = 3; EndeWenn
```

```
Wenn Beruf gleich 3 dann GEWICHT0 = 4; EndeWenn
```

(wobei: 1=Arbeiter, 2=Angestellte, 3=Selbständige)

Sie müssen also zwischen WENN und DANN einen logischen Ausdruck schreiben. Möglich sind z.B. folgende Anweisungen

```
WENN V5 gleich 7 DANN GEWICHT0 = 2; EndeWenn
```

```
WENN V6 groesser 8 DANN GEWICHT0 = 2; EndeWenn
```

```
WENN V7 kleiner 0.5 DANN GEWICHT0 = 2; EndeWenn
```

```
WENN V8 nichtgleich V9 DANN GEWICHT0 = 2; EndeWenn
```

```
WENN V10 groessergleich 2.5 DANN GEWICHT0 = 2; EndeWenn
```

```
WENN V11 kleinergleich 2 DANN GEWICHT0 = 2; EndeWenn
```

Anstelle der Worte "gleich", "groesser" usw. sind auch die üblichen mathematischen Symbole möglich, also:

```
= > < ~ = >= <=
      ↓
    nichtgleich
```

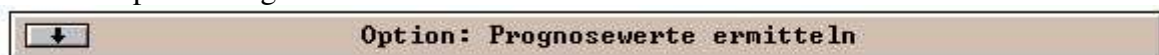
Auch UND sowie ODER sind möglich. Beispiel

```
WENN V5 gleich 7 UND V10 groesser V20 DANN GEWICHT0 = 2; EndeWenn
```

```
WENN V6 kleiner V5 ODER V12 nichtgleich 2 DANN GEWICHT0 = 2; EndeWenn
```

Zur WENN....DANN -Anweisung siehe Handbuch, Teil 2, Abschnitt 29 - 31.

Box 14: Option: Prognosewerte ermitteln



Optionsbox geöffnet:

↓ Loesche wieder diese Box

Prognosewerte ermitteln

↔ 61

für die ersten x Datensätze sollen die tatsächlichen und die durch das Modell prognostizierten Werte in der Zielvariablen ausgegeben werden

Wenn Sie hier zum Beispiel die Zahl 61 eingeben, dann werden für die ersten 61 Datensätze die beobachteten und die durch das Modell reproduzierten Werte in der abhängigen Variablen ausgegeben.

Box 15: Option: Wertemuster

↓ Option: Wertemuster

Optionsbox geöffnet:

↓ Loesche wieder diese Box

Option: Wertemuster Hilfe

↕ 2

Zahl der Wertemuster
=Spalten der nachfolgenden Wertematrix

	Wertemuster I	II	III	IV etc.
↔ □ □ Wohnort	1	2			
↔ □ □ Hausbesitz	1	1			
↔ □ □ Einkommen	4	3			
↔ □ □					
↔ □ □					
↔ □ □					
↔ □ □					
↔ □ □					
↔ □ □					

Betrachten wir ein Beispiel:

Die abhängige Variable sei:

Rückzahlung eines Kredits: nein, ja

Die unabhängigen nominalen Variablen seien:

Geschlecht: männlich (=1), weiblich (=2)

Wohnort: Stadt (=1), Land (=2)

Die unabhängigen quantitativen Variablen seien:

Einkommen

Alter

Sie wollen nun die Wahrscheinlichkeit der Rückzahlung prognostizieren für

1. Männer im Alter von 48

2. Frauen im Alter von 58

Geben Sie als Zahl der Wertemuster = 2 an und schreiben Sie in die Eingabefelder der Wertemustermatrix

	Wertemuster				
	I	II	III	IV etc.
[Geschlecht	, 1	, 2]		
[Alter	, 48	, 58]		

Zuerst wird also der Variablenname (oder -nummer) geschrieben, dann der Wert des 1. Wertemusters, dann der des 2. Es können beliebig viele Wertemuster angefordert werden.

WICHTIG: Als Trennzeichen innerhalb eines Eingabefeldes muss ein Beistrich geschrieben werden, auch hinter dem Variablennamen (bzw. Variablennummer). Am Zeilenende wird kein Beistrich geschrieben.

Almo setzt automatisch für die anderen unabhängigen Variablen, die der Benutzer nicht für die Wertemuster verwendet, deren Mittelwerte ein.

Das gilt auch für die nicht verwendeten nominalen Variablen. In unserem Beispiel wird die nominale Variable "Wohnort" nicht verwendet. Almo löst intern diese Variable in 2 Dummies auf und setzt für diese Dummies deren Mittelwert ein. Der Mittelwert einer Dummy-Variablen ist gleich dem Anteilswert der Probanden, die sich in der betreffenden Ausprägung befinden.

Möglich ist auch folgende Eingabe:

	Wertemuster				
	I	II	III	IV etc.
[Geschlecht	, 1	, 2]		
[Alter	, 48	, 58]		
[Einkommen	, 4	, kw]		

kw eingesetzt

Sie wollen beim 1. Wertemuster das Einkommen mit einer Höhe von 4 einbeziehen - beim 2. Wertemuster jedoch nicht. Dann schreiben Sie beim 2. Wertemuster

KeinWert oder kurz: kw

Almo setzt dann beim 2. Wertemuster für das Einkommen dessen Mittelwert ein.

Hinweis:

Wenn sie mehr Variable in das Wertemuster einbeziehen wollen als Zeilen vorhanden sind, dann gibt es folgende Möglichkeit, die wir an einem Beispiel illustrieren wollen.

	Wertemuster				
	I	II	III	IV etc.
[Geschlecht	, 1	, 2	, Alter	, 48, 58]
[Einkommen	, 7200	, 3500	, Bildung,	5, 3]

Sie schreiben in ein Eingabefeld 2 oder sogar mehrere Variable mit ihren Werten.

BEACHTTE: Alle Zahlenwerte und Variablennamen werden durch Beistrich getrennt. Am Schluß des Eingabefeldes wird kein Beistrich geschrieben. Die Überschrift und die Rahmen dienen nur der "Schönheit". Sie haben keine Bedeutung für Almo.

Box 16: Option: Die errechneten Koeffizienten in eine Datei speichern



Optionsbox geöffnet:



Die vom Programm errechneten Regressionskoeffizienten und Effekte werden in eine Datei gespeichert. Geben Sie einen Dateinamen an.

Box 17: Grafik-Optionen



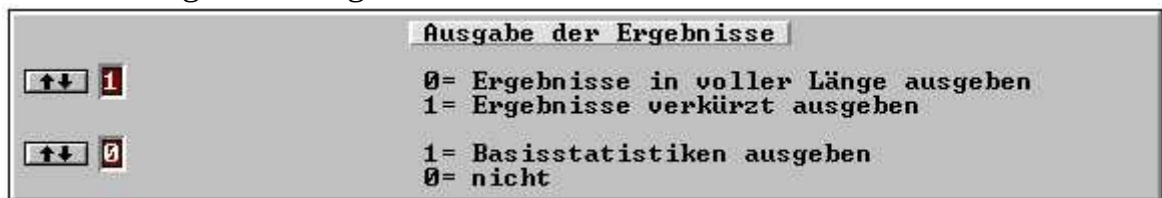
Optionsbox geöffnet:



Siehe Almo-Dokument Nr. 0 "Arbeiten mit Almo", Abschnitt P0.10.

Denn Begriff der "Gruppierungsvariablen" werden wir bei der Ausgabe in Abschnitt P22.2.4 erläutern.

Box 18: Ausgabe der Ergebnisse



Es können zusätzlich Basisstatitiken ausgegeben werden. Dies sind u.a.

- Mittelwerte
- Standardabweichungen
- Zahl der diversen Werte je Variable
- Zahl der fehlenden Werte je Variable

P22.2.1.3 Programm-Maske zur Eingabe fertiger Tabellen mit gruppierten Daten

Prog22mb.Msk
Logit- oder Probit-Analyse (Maximum-Likelihood-Schätzung)
mit Eingabe einer fertigen Tabelle

Beispiel: Der Einfluss der nominalen Variablen Geschlecht (U1) und Beruf (U3) sowie der quantitativen Variablen Einkommen (U5) auf die nominale Variable U10 (Kauf: Ja, Nein) soll ermittelt werden (siehe dazu auch PROG20mj)

Die Daten liegen als schon fertig ausgezählte Tabelle vor (=gruppierte Daten).

Beispiel:

unabhäng.nomin.U.		quantitat.U.	abhängige Var.	
Geschlecht	Beruf	Einkommensgruppe	Häufigkeit für Kauf	Nicht-Kauf
1	1	1	1	8
1	1	2	2	9
1	1	3	3	10
1	2	1	10	8
1	2	2	12	7
1	2	3	14	5
1	3	1	0	6
1	3	2	2	6
1	3	3	3	8
2	1	1	3	5
2	1	2	5	5
2	1	3	8	3
2	2	1	8	3
2	2	2	8	1
2	2	3	9	2
2	3	1	5	3
2	3	2	7	3
2	3	3	6	2

Die 1. Gruppe hat Geschlecht 1 (=männlich)
 Beruf 1 (=Arbeiter)
 Einkommensgruppe 1

Aus dieser 1.Gruppe hat 1 Person das Produkt gekauft und 8 haben es nicht gekauft

Die Einkommensgruppe wird als quantitative unabhängige Variable in das Modell eingeführt. Sie besitzt nur 3 Ausprägungen. Diese werden mit denen der nominalen Variablen "durchkombiniert".

Beim Logit- und Probit-Modell können die unabhängigen Variablen nominal und/oder quantitativ sein.

Abhängig vom Messniveau und der Zahl der Ausprägungen der abhängigen Variablen können folgende Modelle gerechnet werden:

abhängige Variable	Logitanalyse	Probitanalyse
dichotom (ordinal oder nominal)	binäres Logitmodell	binäres Probitmodell
polytom-nominal	multinomiales Logitmodell	nicht möglich
polytom-ordinal	ordinales Logitmodell	ordinales Probitmodell

Almo-Struktur --> Hilfe
 Bedienung --> Hilfe

1

Speicher fuer x Variable Hilfe

Vereinbare Variable= **20** ;

2

Freie Namensfelder Hilfe

↔	↔	↔	Name 1=Geschlecht:männlich,weiblich;
↔	↔	↔	Name 2=Beruf:Arbeiter,Angestellter,Selbständig;
↔	↔	↔	Name 3=Einkommen;
↔	↔	↔	Name 4=Kauf:ja,nein;

abhängige nominale Variable
Sie darf auch mehr als
2 Ausprägungen besitzen

[...] erzeuge zusätzliche Namensfelder

3

Analyse-Variable: Unabhängige nominale Variable Hilfe

↔	□□	↔	↔	Geschlecht, Beruf	Werte-Untergrenzen dieser Variablen Werte-Obergrenzen dieser Variablen
	↔			1, 1	
	↔			2, 3	

Hilfe

Almo-interne Auflösung der unabhängigen nominalen Variablen in Dummies
 0 = 0,1 -Kodierung
 -1 = 0,1,-1 -Kodierung (empfohlen)

↔ **-1**

↕ **1**

0 = erste Dummy-Variablen wird eliminiert
 1 = letzte Dummy-Variablen wird eliminiert

4

Analyse-Variable: Unabhängige quantitative Variable Hilfe

↔ □□ **Einkommen**

5

Analyse-Variable: Abhängige Variable Hilfe

↔ □□ **Kauf** abhängige nominale Variable

ODER (exklusiv)

↔ □□ **■** abhängige ordinale Variable

↔ **1** Werte-Untergrenzen der abhäng. Variablen

↔ **2** Werte-Obergrenzen der abhäng. Variablen

6

Modell Hilfe

[...] **Logit** Logit oder Probit

7

↓ Option: Prognosewerte ermitteln

8

↓ Option: Wertemuster

9

↓ Option: Die errechneten Koeffizienten in eine Datei speichern

10 **Grafik-Optionen**

11 **1** **Ausgabe der Ergebnisse**
 0= Ergebnisse in voller Länge ausgeben
 1= Ergebnisse verkürzt ausgeben

12 **18** **Zeilen und Spalten der einzulesenden Tabelle**
 Zahl der Zeilen der nachfolgend einzulesenden Tabelle
 5 Zahl der Spalten der nachfolgend einzulesenden Tabelle

13 **Schreiben der Tabellenwerte**

Schreiben Sie hier dahinter die Tabelle

In der 1. und den folgenden Spalten stehen die Ausprägungen der unabhängigen nominalen Variablen.

In den dann folgenden Spalten stehen die unabhängigen quantitativen Variablen

Dann folgen die Ausprägungshäufigkeiten der abhängigen Variablen

Schalten Sie dazu die Schreibsperre aus

Schreibsperre <--- EIN : rot
 AUS : grau

```

1 1 1 1 8
1 1 2 2 9
1 1 3 3 10
1 2 1 10 8
1 2 2 12 7
1 2 3 14 5
1 3 1 0 6
1 3 2 2 6
1 3 3 3 8
2 1 1 3 5
2 1 2 5 5
2 1 3 8 3
2 2 1 8 3
2 2 2 8 1
2 2 3 9 2
2 3 1 5 3
2 3 2 7 3
2 3 3 6 2

```

P22.2.1.4 Erläuterungen zu den Boxen

Die Boxen sind weitgehend dieselben wie bei der Programm-Maske Prog22m. Wir erläutern deswegen nur die neuen bzw. veränderten Boxen

Box: Freie Namensfelder



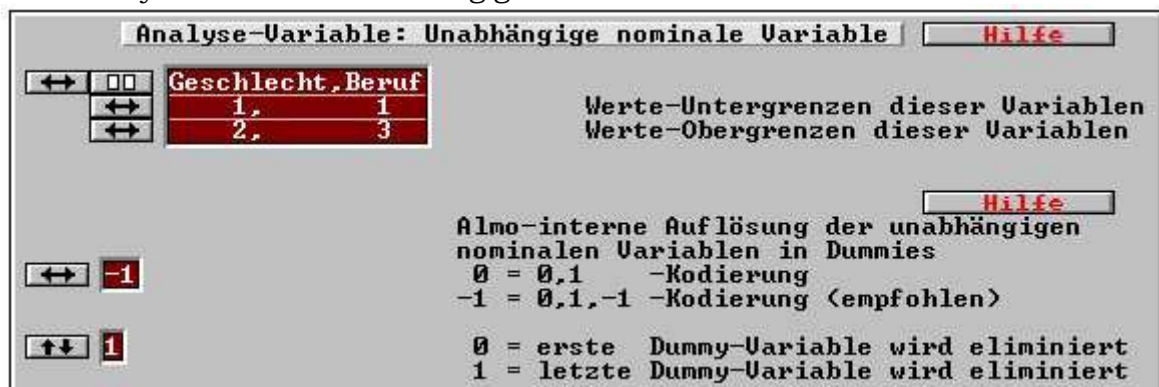
Siehe Almo-Dokument Nr. 0 "Arbeiten mit Almo", Abschnitt P0.3.

Geschlecht, Beruf und Einkommen werden in den nachfolgenden Boxen 3 und 4 als unabhängige Variable angegeben. Die Variablennummern in den Namensgebungen für diese 3 unabhängigen Variablen müssen der Reihenfolge der Spalten der einzugebenden Tabelle entsprechen. Unten, hinter der letzten Box, wird die Tabelle geschrieben. Siehe hier weiter unten im Text.

Ihre 1. Spalte ist das Geschlecht. Also erhält "Geschlecht" die Variablennummer 1.
Ihre 2. Spalte ist der Beruf. Also erhält "Beruf" die Variablennummer 2.
Ihre 3. Spalte ist das Einkommen. Also erhält "Einkommen" die Variablennummer 2.

Die 1. Ausprägung der abhängigen Variablen "Kauf" steht in der 4. Spalte der Tabelle. Also erhält sie die Variablennummer 4. Allgemein: Die Spaltennummer der 1. Ausprägung der abhängigen Variablen ist gleich der Variablennummer in der Namensgebungen für diese Variable.

Box: Analyse-Variable: Unabhängige nominale Variable



2. und 3. Eingabefeld: Hier müssen die Werte-Unter- und Obergrenzen der unabhängigen nominale Variable angegeben werden. Ansonsten ist die Box auszufüllen wie in P22.2.1.2 beschrieben.

Box: Zeilen und Spalten der einzulesenden Tabelle

Zeilen und Spalten der einzulesenden Tabelle	
<input type="text" value="18"/>	Zahl der Zeilen der nachfolgend einzulesenden Tabelle
<input type="text" value="5"/>	Zahl der Spalten der nachfolgend einzulesenden Tabelle

Box: Schreiben der einzulesenden Tabelle

Schreiben der Tabellenwerte	
Schreiben Sie hier dahinter die Tabelle	
In der 1. und den folgenden Spalten stehen die Ausprägungen der unabhängigen nominalen Variablen.	
In den dann folgenden Spalten stehen die unabhängigen quantitativen Variablen	
Dann folgen die Ausprägungshäufigkeiten der abhängigen Variablen	
Schalten Sie dazu die Schreibsperre aus	
<input type="checkbox"/> Schreibsperre	<--- EIN : rot AUS : grau

1	1	1	1	8
1	1	2	2	9
1	1	3	3	10
1	2	1	10	8
1	2	2	12	7
1	2	3	14	5
1	3	1	0	6
1	3	2	2	6
1	3	3	3	8
2	1	1	3	5
2	1	2	5	5
2	1	3	8	3
2	2	1	8	3
2	2	2	8	1
2	2	3	9	2
2	3	1	5	3
2	3	2	7	3
2	3	3	6	2

Die Tabelle wird in folgender Weise eingegeben

unabhängige nominale Variable		unabhängige quantitat.V.	abhängige Variable	
Geschlecht	Beruf	Einkommens gruppe	Häufigkeit für Kauf	Nicht-Kauf
V1	V2	V3	V4	V5
1	1	1	1	8
1	1	2	2	9
1	1	3	3	10
1	2	1	10	8
1	2	2	12	7
1	2	3	14	5
1	3	1	0	6
1	3	2	2	6
1	3	3	3	8
2	1	1	3	5
2	1	2	5	5
2	1	3	8	3
2	2	1	8	3
2	2	2	8	1
2	2	3	9	2
2	3	1	5	3
2	3	2	7	3
2	3	3	6	2

Die Einkommensgruppe wird als quantitative unabhängige Variable in das Modell eingeführt. Sie besitzt nur 3 Ausprägungen. Diese werden mit denen der nominalen Variablen "durchkombiniert". Das bedeutet, dass es nur Sinn macht - bei der Eingabe fertiger Tabellen - Variable als quantitative einzuführen, wenn sie nur wenige Ausprägungen besitzen. Andererseits ist es aber dann problematisch, sie als quantitative zu betrachten.

P22.2.3 Ergebnisse aus binärer Logitanalyse mit unabhängigen nominalen Variablen

Wir werden im folgenden die Ergebnisse aus dem Beispielprogramm ARM_35FM.Alm wiedergeben und erläutern. Das Programm kann aus dem „Kopf“ der Programm-Maske Prog22mb gestartet werden. ARM_35FM.Alm verwendet nur nominale Variable als unabhängige Variable. Wir werden später in Abschnitt P22.2.4 dann darstellen, wie die Ergebnisse bei quantitativen unabhängigen Variablen zu interpretieren sind.

Ergebnisse aus ALMO

Modellspezifikation:

binaeres Logit-Modell

2 unabh. nomin. Var. ... Auspraegungen: 3 2
0 unabh. quant. var.

Modellgleichung:

$p(i,2) = \exp(\text{nue}(i)) / (1 + \exp(\text{nue}(i)))$
 $\text{nue}(i) = \text{const} + \text{Summe } x(i,j) * b(j)$
 $x(6,4) = \text{Designmatrix}$

Es wird zunächst die Modellspezifikation und die entsprechende Modellgleichung ausgegeben. Im Programm ARM_35FFm wird ein binäres Logit-Modell gerechnet, da die abhängige Variable V3 zwei Ausprägungen besitzt (Untergrenze von V3 ist 1 und Obergrenze von V3 ist 2). Auf Seiten der unabhängigen Variablen werden zwei nominale Variable einbezogen, die erste nominale Variable hat 3 Ausprägungen (Dimensionen), die zweite 2 Ausprägungen.

Die Modellgleichung ist

$$(10) \hat{p}_{i2} = \frac{e^{\mu_i}}{1 + e^{\mu_i}}$$

mit

$$(11) \mu_i = b_0 + \sum_j x_{ij} \cdot b_j.$$

Nach der Ausgabe der Modellgleichung werden die in die Analyse einbezogenen Analysevariablen protokolliert:

Analysevariablen:

unabhaengige nominale Variablen:

V1	Alter	Werte-Untergrenze = 1 Obergrenze = 3
V2	Geschlecht	Werte-Untergrenze = 1 Obergrenze = 2

Beachte:

Es wird die 0,1,-1 Dummy-Kodierung verwendet.

abhaengige nominale Variable:

V3	Unfallart	Werte-Untergrenze = 1 Obergrenze = 2
----	-----------	--------------------------------------

Die Haeufigkeiten der abh. Variablen stehen in:

V3
V4

Als nächstes wird in der Ausgabe die Zahl der eingelesenen Datensätze und jene der in die Analyse einbezogenen Datensätze protokolliert:

Zahl der eingelesenen Datensätze = 6

Zahl der in Analyse einbezogenen Datensätze = 6

P22.2.3.0 Modell-Prüfgrößen

Konvergenz des Newton-Raphson-Algor. nach 4 Iterationen :

3.68e-05 = abs. groesste 1. Ableitung < 1.00e-03
5.22e-05 = letzte Verbess. der ln ml > 1.00e-09
2.35e-08 = abs. gr. Veraend. der Koeff.

Der Algorithmus zur Schätzung der Parameter hat nach vier Iterationen eine

konvergente Lösung gefunden, da die absolut größte Veränderung (=1. Ableitung) kleiner dem vorgegebenen Schwellwert von $1.00e-003 = 0.001$ ist.

Ergebnisse der Analyse:

Zahl der Zellen	12
Gesamtfallzahl	56314
durchschnittl. Fallzahl je Zelle	4692.83
Log-Maximum-Likelihood-Wert	-34002.09
Log-ML-Wert f. saturiertes Modell	-33963.44
Likelihood-Ratio-Testgroesse	77.30
Freiheitsgrade	2
Signifikanz 100*(1-p)	100.00
Pearson Chi-Quadrat-Wert	77.53
Freiheitsgrade	2
Signifikanz 100*(1-p)	100.00
Log-ML-Wert f. Nullmodell	-35002.52
LR-Wert (gesch. Modell - Nullmo.)	2000.87
Freiheitsgrade	3
Signifikanz 100*(1-p)	100.00
PRE-Koeffizient	0.963
erklarte Varianz (R**2)	0.999
Pseudo R**2	0.049
PED-Koeffizient	0.029

Die ausgegebenen Größen haben eine unterschiedliche Bedeutung. Folgende Größen stellen nur **Hilfsgrößen** zur Berechnung von Maßzahlen zur Modellbeurteilung dar:

- Zahl der Zellen
- Gesamtfallzahl
- Log-Maximum-Likelihood-Wert
- Log-ML-Wert für saturiertes Modell
- Log-ML-Wert für Nullmodell

"ML" steht für Maximum-Likelihood.

Für die **Modellbeurteilung** sind dagegen folgende Größen relevant:

- **Durchschnittliche Fallzahl je Zelle:** Ist diese Zahl klein, dann sollte nicht mit gruppierten Daten, sondern mit Individualdaten gerechnet werden. Ein genauer Schwellenwert läßt sich nicht angeben. Die durchschnittliche Fallzahl sollte auf jeden Fall größer $1/m$ (m =Zahl der Ausprägungen der abhängigen Variablen) sein.
- **Likelihood-Ratio-Testgroesse** (Likelihood-Ratio-Testgröße, Freiheitsgrade, Signifikanz): Mit dieser Testgröße wird folgende H_0 -Hypohtese geprüft: Das geschätzte Modell führt zu gleich guten Ergebnissen wie ein saturiertes Modell. Neben der Testgröße selbst wird die Zahl der Freiheitsgrade und die Signifikanz ausgegeben. "Signifikant" bedeutet, dass die H_0 -Hypothese verworfen werden muß. Die "Signifikanz" sollte also kleiner einem vordefinierten Schwellenwert von beispielsweise 1%, 5% oder 10% sein, da - im Unterschied zu dem üblichen

Vorgehen in der Statistik - nicht eine der untersuchten Hypothese entgegengesetzte H_0 -Hypothese formuliert wird, sondern die H_0 -Hypothese mit der untersuchten Hypothese identisch ist. Ein "Signifikanzniveau" von 50% beispielsweise bedeutet daher, dass dieselben oder bessere Ergebnisse erzielt werden, wenn zur Prognose der beobachteten Fälle reine Zufallsdaten verwendet werden.

- **Pearson Chi-Quadrat-Wert** (Pearson Chi-Quadrat-Wert, Freiheitsgrade, Signifikanz): Folgende H_0 -Hypothese wird untersucht: Die prognostizierten Häufigkeiten stimmen mit den empirischen Häufigkeiten perfekt überein. Diese H_0 -Hypothese ist strukturgleich mit jener des Likelihood-Ratio-Test. "Signifikant" bedeutet, dass die H_0 -Hypothese verworfen werden muß. Die "Signifikanz" sollte also ebenfalls kleiner einem vordefinierten Schwellenwert von beispielsweise 1%, 5% oder 10% sein.
- **LR-Testgröße (geschätztes Modell - Nullmodell)** (LR-Wert geschätzte Modell, Nullmodell, Freiheitsgrade, Signifikanz): Mit dieser Testgröße wird folgende H_0 -Hypothese geprüft: Das geschätzte Modell führt zu keinen besseren Ergebnissen wie das Nullmodell, das nur die Regressionskonstante enthält. Für diese Testgröße wird wiederum - neben der Testgröße selbst - die Zahl der Freiheitsgrade und die Signifikanz ausgegeben. In diesem Fall sind "signifikante" Ergebnisse erwünscht, da das untersuchte Modell besser sein sollte als das Nullmodell. Inhaltlich läßt sich das Nullmodell wie folgt beschreiben: Die Auftrittswahrscheinlichkeiten p_{ik} der Ausprägung k der abhängigen Variablen sind für alle Datensätze identisch.
- **PRE-Koeffizient:** Dieser gibt an, um wieviele Prozentpunkte sich die Prognose der beobachteten Häufigkeiten verbessert, wenn an Stelle des Nullmodells das untersuchte Modell verwendet wird. Bei gruppierten Daten ist dieser Wert immer hoch und nicht besonders aussagekräftig. Es sollten daher folgende Maßzahlen verwendet werden: pseudo- R^{*2} und der PED-Koeffizient.
- **erklärte Varianz (R^{*2}):** Diese Größe wird strukturgleich zur erklärten Varianz des allgemeinen linearen Modells berechnet. Das zugrundeliegende Nullmodell nimmt an, dass die Häufigkeit in jeder Zelle gleich der durchschnittlichen Fallzahl je Zelle ist. R^2 ist wie der PRE-Koeffizient bei gruppierten Daten relativ hoch und sollte daher nicht verwendet werden. Es sollte das "pseudo- R^{*2} " oder der PED-Koeffizient verwendet werden.
- **Pseudo R^{*2} :** Ist als erklärte Varianz definiert. Als Basis wird nicht die Streuung der beobachteten Häufigkeiten verwendet, sondern die Streuung der individuellen Daten. Diese Streuung wird von ARMINGER/KÜSTERS (1986: 32-33) als Basisdevianz bezeichnet. Die Werte des pseudo R^{*2} liegen zwischen 0 (=schlechte Modellanpassung) und 1 (=perfekte Modellanpassung).
- **PED-Koeffizient** (proportion of explained deviance): Diese Größe mißt den Anteil der erklärten Basisdevianz. Die Werte liegen zwischen 0 (=schlechte Modellanpassung) und 1 (=perfekte Modellanpassung).

Beachte: Wird mit Individualdaten gerechnet, ist die Verwendung der LR-Testgrößen und des Pearsonschen Chi-Quadratwertes äußerst problematisch. H. Potuschak hat diesebezüglich umfangreiche Simulationsstudien durchgeführt. In diesen hat sich gezeigt, dass die Testgrößen keine χ^2 -Verteilung besitzen. Bei Individualdaten sollte zur Beurteilung die berechnete Tabelle der Trefferhäufigkeiten verwendet werden (siehe dazu Abschnitt P22.3). Umkehrt ergeben sich bei großen Stichproben - wie sie für gruppierte Daten vorliegen - für χ^2 -verteilte Testgrößen fast immer signifikante Ergebnisse.

Insgesamt werden somit vier unterschiedliche Perspektiven zur Modellprüfung verwendet: (1) ein saturiertes Modell mit einer perfekten Modellanpassung, (2) ein Nullmodell mit einer Regressionskonstanten, (3) ein Nullmodell mit der durchschnittlichen Fallzahl je Zelle und (4) ein Basismodell für individuelle Daten. Die Modellprüfgrößen für diese drei Perspektiven sind:

Bezugspunkt 1: saturiertes Modell	Bezugspunkt 2: Nullmodell mit einer Regressionskonstante	Bezugspunkt 4: Basismodell
Likelihood-Ratio Test LR	LR-Wert(gesch.Modell - Nullmodell) (signifikante Ergebnisse erwünscht)	pseudo R**2
Pearson Chi-Quadrat (keine signifikanten Ergebnisse erwünscht)	darauf basierend: PRE-Koeffizient (möglichst hohe Werte erwünscht)	PED-Koeffizient (möglichst hohe Werte erwünscht)
	Bezugspunkt 3: Nullmodell mit durchschnittlicher Fallzahl je Zelle	
	erklärte Varianz (möglichst hohe Werte erwünscht)	

Die Definition der einzelnen Modellprüfgrößen und ihre Berechnung soll nachfolgend anhand unseres Beispiels dargestellt werden.

Die Zahl der Zellen ist gleich der Zahl der Datensätze mal der Zahl der Ausprägungen der abhängigen Variablen, in unserem Beispiel also gleich 6 mal 2 = 12. Die Gesamtfallzahl beträgt für unser Beispiel 56314 Fälle. Für die durchschnittliche Fallzahl je Zelle ergibt sich daher ein Wert von 56314 / 12 = 4692.83. Die durchschnittliche Fallzahl ist ausreichend groß, um mit gruppierten Daten zu rechnen.

Der ausgegebene Log-Maximum-Likelihood-Wert entspricht dem Logarithmus der Maximum-Likelihoodfunktion des untersuchten Modells. Tatsächlich wird nämlich nicht die Likelihoodfunktion maximiert, sondern ihr Logarithmus, da dies rechentechnisch einfacher ist. Für unser Beispiel ergibt sich ein Wert von:

$$\ln(L_M) = \ln\left(\prod_{i=1}^n \binom{n_i}{n_{i2}} \cdot \hat{p}_{i2}^{n_{i2}} \cdot \hat{p}_{i1}^{n_{i1}}\right) = -34002.09.$$

Da der Log-Maximum-Likelihood-Wert nur schwer zu interpretieren ist, wird er mit jenem eines saturierten Modells (=Perspektive 1) in Beziehung gesetzt. Als "saturiert" wird ein Modell mit einer perfekten Modellanpassung bezeichnet. Die prognostizierten Häufigkeiten stimmen mit den erhobenen perfekt übereinstimmen. Der entsprechende Log-Maximum-Likelihood-Wert für das saturierte Modell ist in unserem Beispiel gleich -33963.44. Er soll mit $\ln(L_S)$ bezeichnet werden.

Aus dem Log-ML-Wert des geschätzten und des saturierten Modells läßt sich die sogenannte Likelihood-Ratio-Teststatistik berechnen mit

$$LR_M = -2 \cdot (\ln(L_M) - \ln(L_S)),$$

wobei $\ln(L_M)$ der Log-Maximum-Likelihood-Funktionswert des untersuchten Modells ist, $\ln(L_S)$ ist der Log-Maximum-Likelihood-Wert des saturierten Modells. In unserem Beispiel ergibt sich eine Likelihood-Ratio-Teststatistik von

$$\begin{aligned} LR_M &= -2 \cdot (\ln(L_M) - \ln(L_S)) \\ &= -2 \cdot (-34002.09 - (-33963.44)) \\ &= 77.30. \end{aligned}$$

Mit dieser Teststatistik kann - wie bereits erwähnt - die H_0 -Hypothese, dass das untersuchte Modell zu gleich guten Ergebnissen wie das (perfekte) saturierte Modell führt, geprüft werden. Die LR-Statistik ist χ^2 -verteilt mit $df(L_M) - df(L_S)$ Freiheitsgraden. $df(L_M)$ sind die Freiheitsgrade des geschätzten Modells. Diese berechnen sich allgemein wie folgt: $n \cdot (m-1) - k - o$, wobei n die Zahl der in die Analyse einbezogenen Datensätze ist. m ist die Zahl der Ausprägungen der abhängigen Variablen, k die Zahl der geschätzten Parameter und o die Zahl der leeren Zellen (=Zellen mit einer Häufigkeit von 0). $df(L_S)$ sind die Freiheitsgrade des saturierten Modells. Diese sind immer gleich 0. In dem Beispiel ist $n=6$, $m=2$ und $k=4$, $df(L_M)$ ist daher gleich 2 und die Freiheitsgrade der LR-Statistik sind gleich $2 - 0 = 2$.

Für die LR-Statistik ergibt sich in unserem Beispiel eine Signifikanz von 100.00%. Das untersuchte Modell führt also zu signifikant schlechteren Ergebnissen als ein saturiertes Modell. Dieses Ergebnis ist nicht erwünscht.

Der Pearsonsche χ^2 -Test prüft, ob die prognostizierten Häufigkeiten mit den empirischen Häufigkeiten übereinstimmen. Die untersuchte H_0 -Hypothese ist: "Die durch das untersuchte Modell prognostizierten Häufigkeiten stimmen mit den empirischen Häufigkeiten überein." Da das saturierte Modell zu einer perfekten Modellanpassung (prognostizierte Häufigkeiten = empirische Häufigkeiten) führt, ist die H_0 -Hypothese identisch mit der von der LR-Testgröße untersuchten H_0 -Hypothese: "Das untersuchte Modell führt zu gleich guten Ergebnisse wie das saturierte Modell". Die Pearsonsche Teststatistik sollte daher auch ungefähr gleich groß wie die LR-Testgröße sein. Ist dies nicht der Fall, sollten beide Teststatistiken nicht verwendet werden, da die Anwendungsvoraussetzungen mit großer Wahrscheinlichkeit nicht erfüllt sind. In unserem Beispiel sind beide Teststatistiken annähernd gleich groß. Der Pearsonsche χ^2 ist gleich 77.30. Die Signifikanz von 100% bedeutet wiederum, dass die H_0 -Hypothese mit einem Fehler von 0% verworfen werden muß. Wie bei der LR-Statistik wäre aber ein "nicht signifikantes" Ergebnis wünschenswert.

Zusammenfassend kann somit hinsichtlich der ersten Perspektive der Modellprüfung gesagt werden, dass das untersuchte Modell signifikant schlechter ist als ein saturiertes Modell.

Das untersuchte Modell kann aber nicht nur mit einem saturierten Modell in Beziehung gesetzt werden, sondern auch mit dem sogenannten Nullmodell. Das Nullmodell ist jenes Modell, das nur die Regressionskonstante enthält. Inhaltlich wird angenommen, dass die Auftrittswahrscheinlichkeiten in einer Ausprägung der

abhängigen Variablen für alle Datensätze identisch sind. Der Log-ML-Wert für das Nullmodell ist in unserem Beispiel gleich -35002.52. Für den Vergleich dieses Modells mit dem untersuchten Modell kann wiederum eine LR-Statistik mit

$$\begin{aligned} LR_{M-0} &= -2 \cdot (\ln(L_0) - \ln(L_M)) \\ &= -2 \cdot (-35002.52 - (-34002.09)) \\ &= 2000.87 \end{aligned}$$

berechnet werden, wobei $\ln(L_0)$ der Logarithmus der ML-Funktion des Nullmodells ist. Die berechnete Teststatistik besitzt eine χ^2 -Verteilung mit $df(L_0) - df(L_M)$ Freiheitsgraden. Die Freiheitsgrade $df(L_0)$ des Nullmodells berechnen sich allgemein wie folgt: $n \cdot (m-1) - 1 - o$, wobei n die Zahl der in die Analyse einbezogenen Datensätze ist, m ist die Zahl der Ausprägungen der abhängigen Variablen, o die Zahl der leeren Zellen (=Zellen mit einer Häufigkeit von 0). In dem Beispiel ist $n=6$, $m=2$ und $o=0$, $df(L_0)$ ist daher gleich 5. Die Freiheitsgrade der LR-Teststatistik sind daher gleich 3 (=5 - 2), da das untersuchte Modell 2 Freiheitsgrade ($df(L_M)=2$) hat (siehe oben). Die Teststatistik ist mit 100% signifikant. Dieses Ergebnis ist erwünscht. Unser Modell ist signifikant besser als das Nullmodell, das nur die Regressionskonstante enthält. Um wieviele Prozentpunkte unser Modell besser ist als das Nullmodell bringt der PRE-Koeffizient und die erklärte Varianz zum Ausdruck.

Der PRE-Koeffizient ist definiert als

$$PRE = \frac{LR_0 - LR_M}{LR_0},$$

wobei LR_0 der LR-Wert des Nullmodells ist:

$$LR_0 = -2 \cdot (\ln(L_0) - \ln(L_S)) = -2 \cdot ((-35002.52) - (-33963.44)) = 2078.16$$

Für unser Beispiel ergibt sich ein Wert von

$$PRE = \frac{LR_0 - LR_M}{LR_0} = \frac{2078.16 - 70.30}{2078.16} = 0.963.$$

Das untersuchte Modell führt zu einer deutlichen Fehlerreduktion (um 96,3%) gegenüber dem Nullmodell.

Die erklärte Varianz wird berechnet mit

$$R^2 = \frac{\sum_i \sum_j (n_{ij} - \hat{n}_{ij})^2}{\sum_i \sum_j (n_{ij} - \bar{n})^2},$$

wobei n_{ij} die Häufigkeit der j -ten Ausprägung der abhängigen Variablen für den Datensatz i ist. \hat{n}_{ij} sind die entsprechenden prognostizierte Häufigkeiten und \bar{n} die durchschnittliche Fallzahl je Zelle.

Der PRE-Koeffizient und die erklärte Varianz sind bei gruppierten Daten i.d.R. immer sehr groß und unterscheiden sich für unterschiedliche Modelle kaum.

ARMINGER/KÜSTERS (1986) und andere Autoren schlagen deshalb vor, als Bewertungsmaßstab nicht die gruppierten Daten zu verwenden, sondern die Basisdevianz in den Individualdaten. Die Basisdevianz in den Individualdaten ist gleich dem LR-Wert des Nullmodells mit einer Regressionskonstante bezogen auf das saturierte Modell bei Individualdaten. Bezeichnen wir die Basisdevianz mit LR_B , so ist diese wie folgt definiert:

$$LR_B = -2 \cdot (\ln(L_0) - \ln(L_{S(ind)})) = -2 \cdot \ln(L_0),$$

da der Logarithmus des saturierten Modells für Individualdaten gleich 0 ist. In unserem Beispiel ergibt sich für LR_B ein Werte von $(-2) \cdot (-35002.52) = 70005.04$.

Mit Hilfe von LR_B lässt sich nun der Anteil der erklärten Devianz (proportion of explained deviance) berechnen mit

$$PED = \frac{LR_0 - LR_M}{LR_B}$$

Für unser Beispiel ergibt sich ein Wert von:

$$PED = \frac{LR_0 - LR_M}{LR_B} = \frac{2078.16 - 77.30}{-2 \cdot (-35002.52)} = 0.029.$$

Der Anteil der erklärten Devianz ist somit sehr gering. Er beträgt 2.9%.

Beachte: Der PED-Koeffizient von ARMINGER/KÜSTERS (1986: 32-34) ist identisch mit dem Likelihood-Ratio-Index von GREEN (1990: 682).

Das "pseudo R^{*2} " von MADDALA (1990: 40) ist wie folgt definiert:

$$pseudo - R^2 = \frac{(e^{\ln(L_0)})^{2/n} - (e^{\ln(L_M)})^{2/n}}{(e^{\ln(L_0)})^{2/n}} = \frac{(e^{2 \cdot \ln(L_0)})^{1/n} - (e^{2 \cdot \ln(L_M)})^{1/n}}{(e^{2 \cdot \ln(L_0)})^{1/n}} = \frac{(e^{LR_B})^{1/n} - (e^{2 \cdot \ln(L_M)})^{1/n}}{(e^{LR_B})^{1/n}}$$

In unserem Beispiel ergibt sich ein Wert von

$$pseudo - R^2 = \frac{(e^{-35002.52})^{2/n=6} - (e^{-34002.09})^{2/n=6}}{(e^{-35002.52})^{2/n=6}} = 0.049.$$

Die mit dem "pseudo- R^{*2} " gemessene erklärte Basisdevianz ist somit ebenfalls relativ gering. Im Unterschied zum PED-Koeffizient wird die Basisdevianz auf die Likelihood-Funktion zurückgerechnet. Im Nenner steht die durchschnittliche Basisdevianz je Individuum.

Die Modellprüfung kann wie folgt zusammengefasst werden: Das untersuchte Modell ist signifikant schlechter als ein saturiertes Modell. Es erbringt aber auf der anderen Seite signifikant besser Ergebnisse als ein Nullmodell, das nur die Regressionskonstante enthält. In dieser Hinsicht ist es zur Prognose von (gruppierten) Häufigkeiten besser geeignet als das Nullmodell. Dies trifft auf die Prognose individueller Beobachtungen nicht zu. Hier reduziert das Modell nur geringfügig die

Devianz in den Individualdaten. Abhängig von dem Analyseziel wird man das untersuchte Modell verwerfen oder vorläufig akzeptieren. Ist man nur daran interessiert, die Prognose von Häufigkeiten gegenüber dem Nullmodell zu verbessern, wird man das Modell beibehalten. Fordert man dagegen, dass das Modell die beobachteten Häufigkeiten relativ gut reproduziert und/oder die Basisdevianz deutlich reduziert, wird man das Modell verwerfen.

P22.2.3.1 Interpretation der Regressionskoeffizienten

Neben der allgemeinen Bewertung des Modells ist man an den Effekten der untersuchten unabhängigen Variablen interessiert. Hierzu wird für jede untersuchte Variable bzw. deren Dummies der Regressionskoeffizient, dessen Standardfehler, der entsprechende z-Wert und dessen Signifikanz ausgegeben.

Ergebnisse fuer 2. Auspraegung "Fahrunfall" der abhaengigen Variablen V3 Unfallart
(als Referenz wird die 1. Auspraegung "Nichtfahrunf" verwendet)

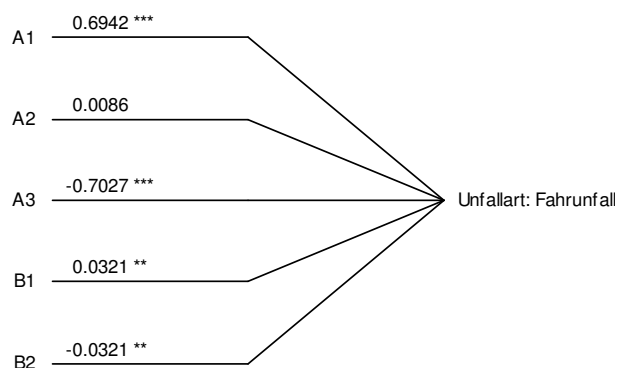
unabhaengige Variab	Regress. koeffiz.	"Risiko" exp (Regr.- koeffiz.)	relatives Risiko	Stand.- Fehler	z-Wert	Signifik. (1-p)*100	partielle Korrelat.
Konstante	-1.07897	-	-	0.01694	63.712	100.00	-
A1 Alter: jung	0.69418	2.00207	100.20667	0.01763	39.377	100.00	0.14873
A2 Alter: mittel	0.00861	1.00865	0.86496	0.01773	0.486	37.29	0.00502
A3 Alter: alt	-0.70279	0.49520	-50.47994	0.02986	23.537	100.00	-0.08880
B1 Geschlec:maennlic	0.03223	1.03276	3.27554	0.01060	3.041	99.74	0.01017
B2 Geschlec:weiblich	-0.03223	0.96828	-3.17165	0.01060	3.041	99.74	-0.01017

ALMO zeichnet noch folgendes Flußdiagramm der Regressionskoeffizienten.

Effekte

A Alter: A1=jung A2=mittel A3=alt

B Geschlecht: B1=maennlich B2=weiblich



Auf den Strichen stehen die Regressionskoeffizienten. Die Sterne hinter den Koeffizienten symbolisieren die Signifikanz 1 (p-100).

3 Sterne = ist mit 99.9% signifikant
 2 Sterne = ist mit 99.0% signifikant
 1 Stern = ist mit 95.0% signifikant
 Kein Stern = Signifikanz unter 95.

Die den Sternen zugeordneten Signifikanzwerte können im Grafikeditor auf der

rechten Seite beliebig gesetzt werden.

Die Interpretation hängt davon ab, ob die 0/1 -Kodierung oder die 0|1|-1 -Kodierung verwendet wurde und ob die erste oder die letzte Dummy eliminiert wurde.

In unserem Beispiel haben wir die 0|1|-1 -Kodierung verwendet. Für die Ausprägung 1 (=jung) der nominalen Variablen 1 (=Alter) ergab sich ein Koeffizient von $A1=0.6974$. Dieser ist wie folgt zu interpretieren: Bei Personen im jungen Alter (=Ausprägung 1 der Variablen V1) treten mit einer überdurchschnittlichen Wahrscheinlichkeit Fahrurfälle auf.

Die 3 Regressionskoeffizienten $A1$, $A2$, $A3$ summieren sich bei der 0|1|-1 -Kodierung zu $.0$. Der Wert $.0$ kann also als "durchschnittlicher Effekt" begriffen werden.

Für jeden berechneten Effekt kann die **Signifikanz** berechnet werden. Dazu wird zunächst für jeden Effekt ein **z-Wert** berechnet. Dieser ist gleich: "Effekt dividiert durch Standardfehler". Für den Effekt $A1$ (=junges Alter) beträgt der Standardfehler gleich 0.01763 . Es ergibt sich ein z-Wert von $(0.694/0.01763)=-39.38$. Die Signifikanz $(1-p)$ 100 dieses z-Wertes ist gleich 99.99% . ALMO rundet diesen Wert auf 100% . Das bedeutet: Personen im jungen Alter haben eine signifikant (zum Niveau von 99.99%) höhere Fahrurfallwahrscheinlichkeit als der Durchschnitt.

Bei der 0/1/-1-Kodierung gibt der Regressionskoeffizient an, ob er größer (positives Vorzeichen) oder kleiner (negatives Vorzeichen) ist als die durchschnittliche Gesamtwirkung der untersuchten Variablen. Nachfolgende Abbildung fasst die Interpretation der Effekte für die unterschiedlichen Kodierungen zusammen.

Dummy-Kodierung	Interpretation der Effekte
Version 1: 0/1-Kodierung, die erste Ausprägung wird gestrichen. Es werden Effekte für die zweite, dritte, vierte, usw. Ausprägung berechnet.	Die erste Ausprägung ist die Referenzgruppe. Die Effekte geben an, ob die Wirkung der anderen Ausprägungen größer/kleiner als jene der Referenzgruppe (erste Ausprägung) ist.
Version 2: 0/1-Kodierung, die letzte Ausprägung wird gestrichen. Es werden Effekte für die erste, zweite, usw. bis zur vorletzten Ausprägung berechnet.	Die letzte Ausprägung ist die Referenzgruppe. Die Effekte geben an, ob die Wirkung der anderen Ausprägungen größer/kleiner als jene der Referenzgruppe (letzte Ausprägung) ist.
Version 3: 0/1/-1-Kodierung. Es werden die Effekte von allen Ausprägungen berechnet. Im Programm wird automatisch die letzte Ausprägung gestrichen.	Die Effekte der Ausprägungen geben an, ob die Wirkung einer Ausprägung größer/kleiner der durchschnittlichen Wirkung von allen Ausprägungen ist. Die durchschnittliche Wirkung ist $.0$

Die 3 Versionen werden bei den Programm-Masken Prog22m und Prog22mb durch entsprechende Eingabe in die Box „Analyse-Variable: Unabhängige nominale Variable“ erzeugt.

Beim „selbst geschriebenen“ Almo-Syntax-Programm werden sie durch folgende Optionen erzeugt.

Version 1: Option 11 = 0; Option 12 = 0;

Version 2: Option 11 = 1; Option 12 = 0;
 Version 3: - Option 12 = 1;

P22.2.3.2 Interpretation des Koeffizienten "Risiko" (beim Logit-Modell)

Das Logit-Modell für unser Beispiel kann durch folgende Gleichungen ausgedrückt werden.

$$(1) \ln\left(\frac{p(\text{Unfall})}{p(\text{keinUnfall})}\right) = \text{const} + A_i + B_k$$

wobei A_i die Regressionkoeffizienten für das Alter sind (mit $i = 1$ ist jung, $i = 2$ ist mittelalt, $i = 3$ ist alt) und B_k die Regressionskoeffizienten für das Geschlecht (mit $k = 1$ ist männlich und $k = 2$ ist weiblich). Entscheidend für die Interpretation ist, dass auf der linken Gleichungsseite ein Bruch steht, der noch dazu logarithmiert ist. Die unabhängigen Variablen auf der rechten Gleichungsseite prognostizieren also das logarithmierte Verhältnis von Unfallwahrscheinlichkeit zu "Kein-Unfallwahrscheinlichkeit". Wir können also folgendes sagen: Je größer der von der rechten Gleichungsseite prognostizierte Wert ist, umso mehr überwiegt die Unfallwahrscheinlichkeit die "Kein-Unfallwahrscheinlichkeit".

Betrachten wir A_i für $i = 1$ (jung). A_i ist 0.694. Das bedeutet: Bei jungen Menschen ist die Unfallwahrscheinlichkeit höher als beim Durchschnitt aus allen Altersgruppen. Etwas genauer: Bei jungen Menschen ist der Quotient aus Unfallwahrscheinlichkeit zu "Kein-Unfallwahrscheinlichkeit" größer als beim Durchschnitt der 3 Altersgruppen. Noch genauer: Der Logarithmus dieses Quotienten ist um 0.694 höher. Den Durchschnitt erhalten wir, wenn wir in obige Gleichung für $A_i = 0$ einsetzen. Nun ist es wenig anschaulich in Quotienten und Logarithmen zu denken

Wenn man den Logarithmus auf der linken Gleichungsseite weg haben will, dann kann man schreiben:

$$(2) \frac{p(\text{Unfall})}{p(\text{keinUnfall})} = e^{\text{const} + A_i + B_k}$$

oder

$$(3) \frac{p(\text{Unfall})}{p(\text{keinUnfall})} = e^{\text{const}} * e^{A_i} * e^{B_k}$$

Auf der linken Gleichungsseite steht nun nur noch der Quotient von Unfallwahrscheinlichkeit zu "Kein-Unfallwahrscheinlichkeit". Auf der rechten Seite stehen multiplikativ verknüpfte Glieder.

Betrachten wir jetzt noch einmal den Fall A_i für $i = 1$ (jung). Wäre $e^{A_i} = 1.0$ dann hätte dieses Glied der rechten Gleichungsseite keinen Einfluß auf den Quotienten auf der linken Seite. Die rechte Gleichungsseite wird mit 1 multipliziert. Dadurch ändert sich nichts. e^{A_i} ist jedoch $e^{0.694} = 2.002$. Es erhöht als den Quotienten um den Faktor 2.002, d.h. das Verhältnis Unfallwahrscheinlichkeit zu "Kein-Unfallwahrscheinlichkeit" verändert sich bei jungen Menschen im Vergleich zum Durchschnitt aller Altersgruppen um den Faktor 2.002 zugunsten der Unfallwahrscheinlichkeit. Der Ausdruck e^{A_i} wird gelegentlich „Risikofaktor“ genannt. Dieser Begriff wird aber nicht allgemein verwendet. Urban (1993, S. 40)

verwendet den Begriff „Effekt-Koeffizient“.

Ist der Regressionskoeffizient $A_i > 0$ dann wird der Risikofaktor e^{A_i} größer 1. Ist A_i kleiner 0 dann wird e^{A_i} kleiner 1 (aber größer 0). Für die Altersgruppe $k = 3$ (alt) ist $A_3 = -0.703$ und der Risikofaktor $e^{A_3} = 0.495$, gerundet = 0.5. Wir können interpretieren: Das Verhältnis (der Quotient) von Unfallwahrscheinlichkeit zu "Kein-Unfallwahrscheinlichkeit" ist bei alten Menschen nur halb so groß wie beim Durchschnitt aller Altersgruppen. Diesen Durchschnitt erhalten wir, wenn wir $A_i = 0$ setzen.

Almo gibt noch das „relative Risiko“ aus. Dies ist sehr einfach.

$$\text{relatives Risiko} = (\text{Risiko} - 1.0) * 100$$

Für die Altersgruppe A3 erhalten wir

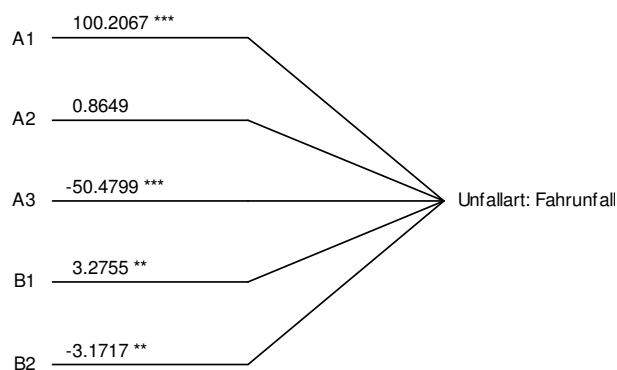
$$\text{relatives Risiko} = (0.495 - 1) * 100 = - 50.5$$

Wir interpretieren: Die alten Menschen haben ein um 50.5% verringertes Risiko einen Fahrnunfall zu erleiden – im Vergleich zur „Durchschnittsperson“.

Auch der Begriff des „relativen Risikos“ ist kein allgemein verwendeter.

Almo zeichnet ein Flußdiagramm des relativen Risikos.

relative Risikokoeffizienten
 fuer unabhangige Variable
 A Alter: A1=jung A2=mittel A3=alt
 B Geschlecht: B1=maennlich B2=weiblich



P22.2.3.3 Interpretation der Kontraste

Mitunter werden bei der Interpretation der Ergebnisse die Kontraste zwischen den einzelnen Auspragungen einer nominalen Variablen benotigt. Das Programm berechnet dazu folgende paarweisen Differenzen:

Kontraste der nominalen Variablen fuer
 2.Auspraegung Unfall der abhaeng. Var. V3 Unfallart

Kontrast- paar	Kontrast	"Risiko" exp (Kontrast)	Stand.- Fehler	z-Wert	Signifik. (1-p) *100
A1 - A2	0.6856	1.9849	0.0189	36.206	100.000
A1 - A3	1.3970	4.0429	0.0457	30.555	100.000

A2 - A3	0.7114	2.0369	0.0458	15.521	100.000
B1 - B2	0.0645	1.0666	0.0212	3.041	99.745

Wir wollen beispielsweise prüfen, ob sich Personen im mittleren Alter (=A2) von jenen im hohen Alter (=A3) signifikant unterscheiden. Die Differenz beträgt: $0.00861 - (-0.70279) = 0.7114$. Diese Differenz besitzt einen Standardfehler von 0.0458. Der z-Wert der Differenz ist daher gleich $0.7114 / 0.0458 = 15.5213$ und besitzt eine Signifikanz von $\approx 100\%$. Personen im hohen Alter haben also eine signifikant geringere Fahrnunfallhäufigkeit als Personen im mittleren Alter, da die Differenz negativ und signifikant ist.

P22.2.3.4 Beobachtete und prognostizierte Häufigkeiten

Das Programm gibt für die ersten sechs Datensätze die beobachteten und die durch das Modell prognostizierten bzw. reproduzierten Häufigkeiten aus: (in unserem Beispiel gibt es nur 6 Datensätze)

Der Benutzer erreicht dies beim „selbst geschriebenen“ Almo-Programm, indem er „Option 7=6;“ setzt und bei den Programm-Masken Prog22m und Prog22mb, indem er in der Box „Optionen“ die Zahl 6 einträgt.

Beobachtete und durch das Modell reproduzierte Häufigkeiten

```

die unabhangigen nominalen Variablen sind
A = V1 Alter
B = V2 Geschlecht

beo1 ... = beobachtete Haeufigkeiten in
          der Auspraegung 1 der abhaeng. Variablen
repl ... = reproduzierte Haeufigkeiten in
          der Auspraegung 1 der abhaeng. Variablen

Nr.  A B  beo1  beo2   repl   rep2
-----
1  1  1  10128  7468 10333.1  7262.9
2  1  2   3847  2195  3641.9  2400.1
3  2  1  15373  5167 15168.6  5371.4
4  2  2   5770  2188  5974.4  1983.6
5  3  1   3047   529  3046.4   529.6
6  3  2    517    85   517.6    84.4

```

In der Spalte "beo1" stehen die beobachteten Häufigkeiten der 1. Ausprägung der abhängigen nominalen Variablen, in der Spalte "beo2" die beobachteten Häufigkeiten der 2. Ausprägung. Die prognostizierten bzw. reproduzierten Häufigkeiten werden mit "rep1" bzw. "rep2" berechnet. Sie können aufgrund der Gleichungen (5) und (6) aus Abschnitt P22.1 berechnet werden. Wir wollen dies für den Datensatz $i=1$ zeigen:

$$\begin{aligned}
\hat{p}_{i2} &= \frac{e^{b_0(=-1.079)+b_{11}(=0.694) \cdot x_{i11}(=1)+b_{12}(=0.009) \cdot x_{i12}(=0)+b_{21}(=0.032) \cdot x_{i21}(=1)}}{1 + e^{b_0(=-1.079)+b_{11}(=0.694) \cdot x_{i11}(=1)+b_{12}(=0.009) \cdot x_{i12}(=0)+b_{21}(=0.032) \cdot x_{i21}(=1)}} \\
&= \frac{e^{-0.35256}}{1 + e^{-0.35256}} = 0.413
\end{aligned}$$

und

$$\hat{p}_{i1} = \frac{1}{1 + e^{b_0(=-1.079) + b_{11}(=0.694) \cdot x_{i11}(=1) + b_{12}(=0.009) \cdot x_{i12}(=0) + b_{21}(=0.032) \cdot x_{i21}(=1)}}$$

$$= \frac{1}{1 + e^{-0.35256}} = 0.587$$

p_{i1} und p_{i2} ist die Wahrscheinlichkeit das für die abhängige Variable die Ausprägung 1 bzw. 2 auftritt. i ist der Index für die Datensatznummer. In unserem Beispiel ist $i=1$.

Da der 1. Datensatz eine Auftrittshäufigkeit von 17596 Fällen besitzt, ergibt sich für "rep1" ein Wert von $17596 \cdot 0.587 = 10328.9$. Für "rep2" ergibt sich analog ein Wert von $17596 \cdot 0.413 = 7267.1$. (Die Abweichungen zur Ausgabe entstehen dadurch, dass wir nur mit einer Genauigkeit von 3 Kommastellen gerechnet haben, während ALMO intern mit einer Genauigkeit von bis zu 16 Kommastellen rechnet.)

P22.2.4 Ergebnisse aus binärer Logitanalyse mit unabhängigen nominalen und quantitativen Variablen

Es ist möglich eine Analyse mit nur unabhängigen quantitativen Variablen zu rechnen, aber auch eine bei der diese mit unabhängigen nominalen Variablen gemischt sind. In Abschnitt P22.2.1.1 haben wir die Programm-Maske Prog22m.Msk dargestellt.

Das Programm liefert folgendes Ergebnis (gekürzt)

Ergebnisse fuer 2. Auspraegung "ja" der abhaengigen Variablen V10 Kauf
(als Referenz wird die 1. Auspraegung "nein" verwendet)

unabhaengige Variab	Regress. koeffiz.	"Risiko" exp(Regr.-koeffiz.)	relatives Risiko	Stand.- Fehler	z-Wert	Signifik. (1-p)*100	partielle Korrelat.
Konstante	1.48708	-	-	0.77648	1.915	94.45	-
A1 Geschlec:männlich	0.54538	1.72527	72.52719	0.32991	1.653	90.14	0.09332
A2 Geschlec:weiblich	-0.54538	0.57962	-42.03812	0.32991	1.653	90.14	-0.09332
B1 Beruf:Arbeiter	1.42062	4.13969	313.96933	0.55006	2.583	99.03	0.23558
B2 Beruf:Angestel	-1.08293	0.33860	-66.13972	0.42484	2.549	98.92	-0.23118
B3 Beruf:Selbstän	-0.33769	0.71341	-28.65865	0.48047	0.703	51.77	-0.13378
V5 Verdienst	-0.27493	0.75963	-24.03724	0.17588	1.563	88.18	-0.07259

Paarweise Vergleiche (Kontraste) der nominalen Variablen fuer
2. Auspraegung ja der abhaeng. Var. V10 Kauf

Vergleichs- paar	Differenz	"Risiko" exp(Differenz)	Stand.- Fehler	z-Wert	Signifik. (1-p)*100
A1 - A2	1.0908	2.9766	0.6598	1.653	90.138
B1 - B2	2.5035	12.2258	0.8575	2.920	99.636
B1 - B3	1.7583	5.8027	0.9415	1.868	93.820
B2 - B3	-0.7452	0.4746	0.7212	1.033	69.845

Beobachtete und durch das Modell reproduzierte (prognostizierte) Wahrscheinlichkeiten (in %)

die unabhaengigen nominalen Variablen sind
A = V1 Geschlecht
B = V3 Beruf
ihre Auspraegungen werden mit 1,2,3,... durchnummeriert

die unabhaengigen quantitativen Variablen sind

quant1 = V5 Verdienst

beo1 ... = Ausprägung 1 der abhaengigen Variablen
1=aufgetreten 0=nicht aufgetreten

rep1 ... = reproduzierte (prognostizierte) Wahrscheinlichkeit fuer das
Auftreten der Auspraegung 1 in der abhaeng. Variablen

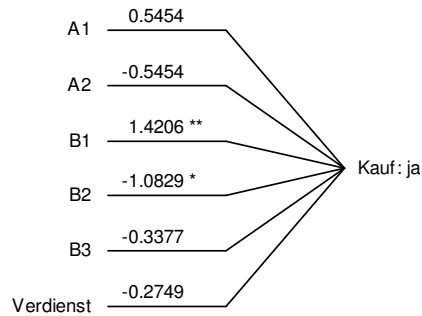
Nr.	A	B	quant1	beo1	beo2	rep1	rep2
1	1	1	4.000	0	1	8.7	91.3
2	1	1	5.000	0	1	11.1	88.9
3	1	1	4.000	0	1	8.7	91.3
4	1	1	2.000	0	1	5.2	94.8
5	1	2	4.000	0	1	53.7	46.3
6	1	2	4.000	0	1	53.7	46.3
.
.
.
.

Dabei liefert Almo auch noch folgendes Flußdiagramm für die Zusammenhänge zwischen unabhängigen Variablen und abhängiger Variabler.

Effekte und Regressionskoeffizienten

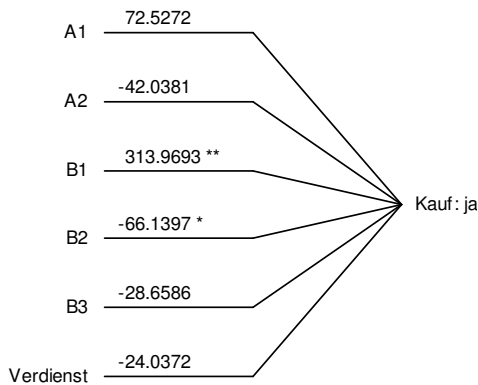
A Geschlecht: A1=männlich A2=w weiblich

B Beruf: B1=Arbeiter B2=Angestellter B3=Selbständiger



Auch ein Flußdiagramm der „relativen Risikokoeffizienten“ wird gezeichnet.

relative Risikoeffizienten
 fuer unabhaengige Variable
 A Geschlecht: A1=männlich A2=weiblich
 B Beruf: B1=Arbeiter B2=Angestellter B3=Selbständiger



Die Regressionskoeffizienten und die „Risikokoeffizienten“ der quantitativen Variablen sind analog denen in P22.2.3.1 und P22.2.3.2 beschriebenen Koeffizienten der nominalen Variablen zu interpretieren.

Regressionskoeffizient: Der Wert von -0.1559 für die unabhängige Variable Alter bedeutet, dass wenn sich das Alter um 1 Maßeinheit erhöht, die Wahrscheinlichkeit für einen Kauf sich verringert; genauer: dass der Logarithmus des Quotienten Kaufwahrscheinlichkeit zu Nichtkaufwahrscheinlichkeit sich um -0.1559 verringert.

Risikokoeffizient: Der Wert von 0.8556 bedeutet, dass wenn sich das Alter um 1 Maßeinheit erhöht, das „Risiko“ für einen Kauf um den Faktor 0.8556 sich verringert; genauer: dass der Quotient vom Kauf zum Nichtkauf sich um den Faktor 0.8556 verringert. Wäre das Risiko höher als 1.0, dann müsste von einer Erhöhung des Risikos gesprochen werden.

Almo gibt noch die Trefferhäufigkeit aus

Trefferhaeufigkeiten bei Individualdaten
 fuer abhaengige Variable V10 Kauf

		tatsaechlich		prognostiziert	
		1 nein	2 ja	1 nein	2 ja
nein	1	28	0	18	10
	ja	0	33	10	23

		prognostiziert		erwartet	
		relativ		Zufall	
		1 nein	2 ja	1 nein	2 ja
nein	1	16.5	11.5	12.9	15.1
	ja	11.5	21.5	15.1	17.9

absolut: Chi-Quadrat (1) = 7.044 Signifikanz 100*(1-p) = 99.197
 relativ: Chi-Quadrat (1) = 3.465 Signifikanz 100*(1-p) = 93.732

Im Verlauf der Logit-Analyse wird für jede Person die Wahrscheinlichkeit prognostiziert, dass sie der Gruppe der Nicht-Käufer bzw. der Gruppe der Käufer angehört. Ist die Wahrscheinlichkeit den Nicht-Käufern anzugehören größer als die für die Käufer, dann wird sie der Gruppe der Nicht-Käufer zugerechnet und entsprechend umgekehrt.

In der 1. Teil-Tabelle, überschrieben mit "tatsächlich", erkennen wir, dass 28 Personen Nicht-Käufer sind und 33 Käufer. In der 2. Teil-Tabelle, überschrieben mit "prognostiziert absolut", sehen wir, dass 18 Nicht-Käufer vom Logit-Modell richtig identifiziert und 10 falsch identifiziert wurden. Von den 33 Käufern werden 23 richtig und 10 falsch identifiziert.

In der 4. Teil-Tabelle, überschrieben mit "erwartet Zufall" wird uns gezeigt, wie die Prognose wäre, wenn wir zufällig aus den 61 Personen 33 Käufer auswählen würden. Dann wären davon nur 17.9 (gerundet 18) Personen richtig getroffen worden. Die Trefferquote wäre $100 \cdot 18 / 33 = 54.5\%$. Das Logit-Modell hat eine Trefferquote von $100 \cdot 23 / 33 = 69.7\%$. Das ist allerdings nicht sehr viel besser. Unsere Testdaten sind simulierte Daten mit schwachen Zusammenhängen.

Almo gibt uns nun auch noch aus, ob die Logit-Prognose im Vergleich zur "Zufalls-Prognose" signifikant verschieden ist. Es wird ein Chi-Quadrat-Wert von 7.044 gefunden, der mit 99.197 % signifikant ist.

Die 3. Tabelle, überschrieben mit "prognostiziert relativ" und der dazu gehörende Chi-Quadrat-Wert, bezeichnet mit "relativ: Chi-Quadrat(1) =" wird hier nicht erläutert. Siehe dazu Abschnitt P22.3.

Almo zeichnet abschließend noch die logistische Funktion für die unabhängige Variable des Einkommens. Da wir in Prog22m in der Box "Grafik" als "Gruppierungsvariable" das Geschlecht angegeben haben werden 2 Kurven gezeichnet:

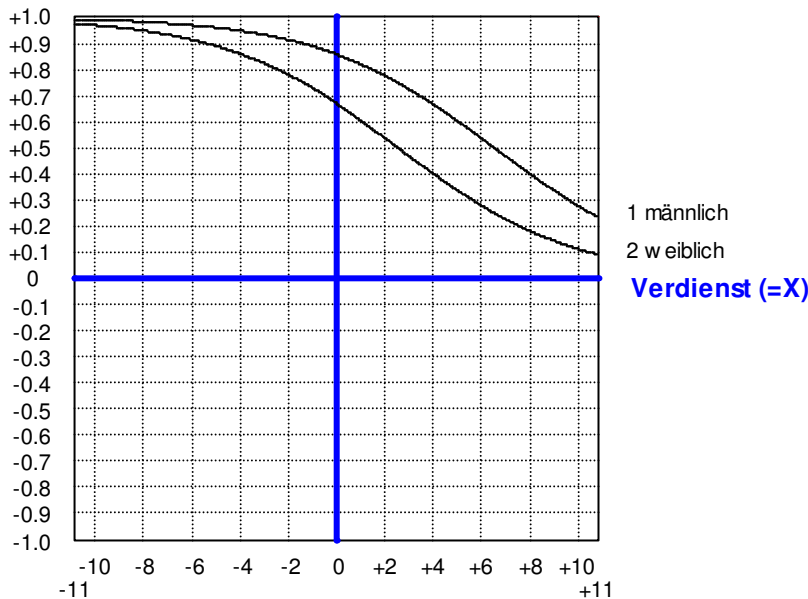
Logistische Funktion fuer
 abhaengige Variable: V10 Kauf: 2.Auspraegung: ja
 unabhaengige Variable: V5 Verdienst

Gruppierungsvariable:
 1: gruene Linie: V1 Geschlecht 1.Auspraegung: männlich
 2: blaue Linie: V1 Geschlecht 2.Auspraegung: weiblich

Logistische Funktion

1: gruene Linie: männlich $Y = 1/(1+e^{-(1.8017-0.27493*X)})$
 2: blaue Linie: weiblich $Y = 1/(1+e^{-(0.71091-0.27493*X)})$

Wahrscheinlichkeit von Kauf: ja (=Y)



Um unsere Erläuterung dieser Kurven allgemein halten zu können, wollen wir von unserem Beispiel etwas abweichen und folgende Analyse-Variablen unterstellen:

Die abhängige Variable ist "Kauf eines Produktes" mit den Ausprägungen "Kauf_Nein" und "Kauf_Ja". Sie wird bestimmt durch

unabhängige nominale Variable
 Geschlecht: männlich, weiblich
 Beruf: Angestellter, Beamter

unabhängige quantitative Variable
 Verdienst
 Schulden

In Abweichung von unserem Beispiel haben wir also unterstellt, wir hätten noch eine 2. quantitative unabhängige Variable (die Schulden) und die nominale unabhängige Variable Beruf hätte nur 2 und nicht 3 Ausprägungen.

Almo zeichnet nun je eine logistische Funktion für die 2 unabhängigen quantitativen Variablen. Dabei wird die unabhängige quantitative Variable an die x-Achse geschrieben und die abhängige Variable "Wahrscheinlichkeit für Kauf_Ja" an die y-Achse.

Zuerst wird die logistische Funktion für "Verdienst" (x-Achse) und "Wahrscheinlichkeit für Kauf_Ja" (y-Achse) gezeichnet.

Dann wird die logistische Funktion für "Schulden" (x-Achse) und "Wahrscheinlichkeit für Kauf_Ja" (y-Achse) gezeichnet.

Die jeweils anderen unabhängigen quantitativen Variablen werden dabei auf ihren Mittelwert gesetzt. Auch die Dummies der unabhängigen nominalen Variablen werden auf ihre Mittelwert gesetzt. Dieser entspricht dem Anteilswert der Ausprägungen.

Die Logit-Modell-Gleichung für unser Beispiel ist nachfolgend in (1) angegeben.

$$(1) \quad p=1/(1+e^{**-(\beta_1 * E + \beta_2 * S + \beta_3 * G_m + \beta_4 * G_w + \beta_5 * B_a + \beta_6 * B_b + \text{const}))})$$

Die Gleichung, die Almo zeichnet, in steht in (2)

$$(2) \quad p=1/(1+e^{**-(\beta_1 * E + \beta_2 * MS + \beta_3 * MG_m + \beta_4 * MG_w + \beta_5 * MB_a + \beta_6 * MB_b + \text{const}))})$$

e =das ist die Zahl e (=2.718...)
****** =Symbol fuer Exponentiation
p ="Wahrscheinlichkeit für Kauf_Ja"
E =Verdienst
S =Schulden
MS =Mittelwert aus Schulden
MG_m, MG_w =Mittelwert für Geschlecht: männlich bzw. weiblich
MB_a, MB_b =Mittelwert für Beruf: Angestellter bzw. Beamter
β₁ =Regressionskoeffizient für Verdienst
β₂ =Regressionskoeffizient für Schulden
β₃ =Regressionskoeffizient für Geschlecht: männlich
β₄ =Regressionskoeffizient für Geschlecht: weiblich
β₅ =Regressionskoeffizient für Beruf: Angestellter
β₆ =Regressionskoeffizient für Beruf: Beamter
const =Konstante

Für die unabhängige quantitative Variable "Schulden" ist in (2) deren Mittelwert eingesetzt worden. Ebenso für die Dummies der unabhängigen nominalen Variablen. Das entspricht der Einsetzung einer "Durchschnittsperson".

Wir können also etwas verkürzt formulieren:

In der Almo-Grafik wird für die "Durchschnittsperson" der logistische Zusammenhang zwischen Verdienst und "Wahrscheinlichkeit für Kauf_Ja" gezeichnet.

Im Titel der Almo-Graphik wird Gleichung (2) angegeben. Dabei wird der Gleichungsteil

$$\beta_2 * MS + \beta_3 * MG_m + \beta_4 * MG_w + \beta_5 * MB_a + \beta_6 * MB_b + \text{const}$$

aus obiger Gleichung in einem Zahlenwert zusammengefasst.

Gruppierungsvariable

Nun besteht die Möglichkeit eine oder mehrere Gruppierungsvariable anzugeben.

Beachte: Als Gruppierungsvariable können nur Variable verwendet werden, die als unabhängige nominale Variable angegeben wurden.

Es wird beispielsweise das "Geschlecht" als Gruppierungsvariable angegeben. Almo zeichnet dann die logistischen Funktionen (so wie oben beschrieben) für die beiden Ausprägungen des Geschlechts. Es werden also folgende Kurven gezeichnet:

1. Verdienst (x-Achse) mit "Wahrscheinlichkeit für KaufJa" (y-Achse) für die Männer.
2. Verdienst (x-Achse) mit "Wahrscheinlichkeit für KaufJa" (y-Achse) für die Frauen.
3. Schulden (x-Achse) mit "Wahrscheinlichkeit für KaufJa" (y-Achse) für die Männer.
4. Schulden (x-Achse) mit "Wahrscheinlichkeit für KaufJa" (y-Achse) für die Frauen.

Die jeweils anderen ursächlichen Variablen sind dabei auf ihren Mittelwert gesetzt. Bei der Kurve 1 wird also

Schulden Beruf:Angestellter Beruf:Arbeiter

auf den Mittelwert bzw. Anteilswert gesetzt.

Kombinierte Gruppierungsvariable

Zwei oder mehrere Gruppierungsvariable können auch kombiniert werden. Dazu wird die MIT-Anweisung aus der Almo-Programmiersprache verwendet. Siehe dazu Handbuch Teil 2. Betrachten wir ein Beispiel:

In die Box "Grafik-Optionen" schreiben Sie in das Eingabefeld für die Gruppierungsvariable

Geschlecht MIT Beruf

Beachte: Es können maximal 4 Variable durch MIT kombiniert werden.

Almo erzeugt dann folgende Kombinationen in folgender Reihenfolge

männlich mit Angestellter
männlich mit Beamter
weiblich mit Angestellter
weiblich mit Beamter

Die jeweils hintere Variable "läuft" über ihre Ausprägungen. Für jede Kombination wird eine Funktionsgrafik gezeichnet

Mehrere Gruppierungsvariable

Es können mehrere Gruppierungsvariable (durch Beistrich getrennt) angegeben werden. Beispiel:

Geschlecht, Beruf

Almo zeichnet dann beispielsweise für den Verdienst 4 Kurven, eine für die Männer, eine für die Frauen, eine für die Angestellten und eine für die Beamten. Ebenso

werden 4 Kurven für die Variable "Schulden" gezeichnet.

Mehrere einzelne Gruppierungsvariable und mehrere durch MIT kombinierte Gruppierungsvariable können angegeben werden.

Beispiel:

Geschlecht, Beruf, Geschlecht MIT Beruf

P22.2.5 Ergebnisse aus Logitanalyse mit polytomer abhängiger Variablen

Beim multinominalen Logitmodell werden nicht nur für die zweite Ausprägung der abhängigen Variablen Effekte berechnet, sondern auch für die dritte Ausprägung, für die vierte Ausprägung usw.

Wir verwenden ein Beispiel aus Arminger/Küsters (1986, S.102). Das Programm ist unter dem Namen "Arm102k.Alm" in Almo enthalten. Man findet das Programm nach Klick auf den Knopf „alle Progs“ am Oberrand des Almo-Fensters. Die Daten liegen als schon fertig ausgezählte Tabelle vor (=gruppierte Daten). Ein Beispiel mit Individualdaten ist PolyLogit.Alm, das in der gleichen Weise zu finden ist.

Die abhängige polytome Variable ist:

V4 Unfallart: Sachschaden, Leichtverletzt, Schwerverletzte, Tote

Als *Referenzkategorie* verwendet Alm die erste Ausprägung (im Beispiel: Sachschaden). Bei Daten in der Form fertig ausgezählten Tabellen (wir nennen sie "gruppierte Daten") verwendet Almo immer die erste Ausprägung. Der Benutzer kann das nicht ändern.

Die unabhängigen nominalen Variablen sind:

V3 Geschlecht: maennlich, weiblich

V1 Straßenzustand: trocken,nass,Eis

Die unabhängige quantitative Variable ist:

V2 Alter: jung,mittel,alt

(Diese Variable besitzt also nur 3 Ausprägungen)

Die Daten liegen als schon fertig ausgezählte Tabelle vor (=gruppierte Daten).

V1	V2	V3	Unfallart (Fallzahl)			
			V4	V5	V6	V7
Straße	Alter	Geschlecht	Sachschad	Leichtverl	Schwerverl	Tote
1	1	1	4037	2510	2042	212
1	1	2	1043	912	805	37
1	2	1	4981	2923	1833	258
1	2	2	1530	1097	769	76
1	3	1	956	591	424	67
1	3	2	144	110	82	12
2	1	1	3131	1819	1492	150
2	1	2	899	738	601	37
2	2	1	4012	2157	1239	161
2	2	2	1415	938	621	57
2	3	1	608	356	252	35

2	3	2	89	61	46	4
3	1	1	863	712	579	49
3	1	2	302	319	336	13
3	2	1	1331	922	657	66
3	2	2	496	540	394	25
3	3	1	108	91	77	11
3	3	2	15	23	13	3

BEACHTEN dass ungewöhnlicherweise die quantitative Variable "Alter" als 2. Variable, zwischen den beiden nominalen Variablen "Straße" und "Geschlecht", steht. Bei Küsters wird die Tabelle jedoch in dieser Form angegeben, da dort das Alter als nominale Variable behandelt wird. Wir wollen diese Variable hier jedoch als quantitative behandeln.

Da eine fertige, schon ausgezählte Tabelle vorliegt verwenden wir unsere Programm-Maske Prog22mb (siehe Abschnitt P22.2.1.3). Die vollständig ausgefüllte Programm-Maske ist als Beispielprogramm ARM102K.ALM in Almo enthalten. Der Benutzer findet das Programm nach Klick auf den Knopf „alle Progs“ am Oberrand des Almo-Fensters.

Wir wollen hier nur die wesentlichen Boxen dieses Programm abbilden und erläutern.

Box 2: Freie Namensfelder



Die Variablennummern in den Namensgebungen für die 3 unabhängigen Variablen müssen der Reihenfolge der Spalten der einzugebenden Tabelle entsprechen.

In der 1. Spalte der Tabelle steht der Straßenzustand.
Also erhält "Strasse" die Variablennummer 1.

In der 2. Spalte der Tabelle steht das Alter
Also erhält "Alter" die Variablennummer 2.

In der 3. Spalte der Tabelle steht das Geschlecht
Also erhält "Geschlecht" die Variablennummer 3.

Die 1. Ausprägung der abhängigen Variablen "Unfallart" steht in der 4. Spalte der Tabelle. Also erhält sie die Variablennummer 4. Allgemein: Die Spaltennummer der 1. Ausprägung der abhängigen Variablen ist gleich der Variablennummer in der Namensgebungen für diese Variable.

Box 3 und 4: Unabhängige nominale und quantitative Variable

Analyse-Variable: Unabhängige nominale Variable Hilfe

↔ **Strasse, Geschlecht**

↔ Werte-Untergrenzen dieser Variablen

↔ Werte-Obergrenzen dieser Variablen

Hilfe

Almo-interne Auflösung der unabhängigen nominalen Variablen in Dummies

↔ 0 = 0,1 -Kodierung
-1 = 0,1,-1 -Kodierung

↕ 0 = erste Dummy-Variable wird eliminiert
1 = letzte Dummy-Variable wird eliminiert

Analyse-Variable: Unabhängige quantitative Variable Hilfe

↔ **Alter**

Box 5: Abhängige polytome Variable

Analyse-Variable: Abhängige Variable Hilfe

↔ **Unfallart**

abhängige nominale Variable

↔ ODER (exklusiv)

abhängige ordinale Variable

↔ Werte-Untergrenzen der abhäng. Variablen

↔ Werte-Obergrenzen der abhäng. Variablen

Almo liefert folgende Ergebnisse, die wir hier gekürzt wiedergeben.

Ergebnisse fuer 2. Auspraegung "Leichtverlet" der abhaengigen Variablen V4 Unfallart (als Referenz wird die 1. Auspraegung "Sachschaden" verwendet)

unabhaengige Variab	Regress. koefiz.	"Risiko" exp(Regr.- koefiz.)	relatives Risiko	Stand.- Fehler	z-Wert	Signifik. (1-p)*100	partielle Korrelat.
Konstante	-0.22762	-	-	0.02958	7.694	100.00	-
A1 Strasse: trocken	-0.05616	0.94539	-5.46082	0.01408	3.989	100.00	-0.01043
A2 Strasse: nass	-0.13453	0.87412	-12.58765	0.01478	9.105	100.00	-0.02515
A3 Strasse: Eis	0.19069	1.21008	21.00835	0.01914	9.961	100.00	0.02757
B1 Geschlec:maennlic	-0.13551	0.87327	-12.67284	0.01135	11.942	100.00	-0.03316
B2 Geschlec:weiblich	0.13551	1.14512	14.51191	0.01135	11.942	100.00	0.03316
V2 Alter	-0.05281	0.94856	-5.14360	0.01619	3.262	99.87	-0.00822

Ergebnisse fuer 3. Auspraegung "Schwerverlet" der abhaengigen Variablen V4 Unfallart
(als Referenz wird die 1. Auspraegung "Sachschaden" verwendet)

unabhaengige Variab	Regress. koeffiz.	"Risiko" exp(Regr.- koeffiz.)	relatives Risiko	Stand.- Fehler	z-Wert	Signifik. (1-p)*100	partielle Korrelat.
Konstante	-0.24326	-	-	0.03247	7.491	100.00	-
A1 Strasse: trocken	-0.06292	0.93902	-6.09804	0.01544	4.074	100.00	-0.01068
A2 Strasse: nass	-0.18927	0.82756	-17.24392	0.01634	11.580	100.00	-0.03214
A3 Strasse: Eis	0.25219	1.28684	28.68426	0.02066	12.207	100.00	0.03391
B1 Geschlec:maennlic	-0.17427	0.84007	-15.99318	0.01238	14.081	100.00	-0.03918
B2 Geschlec:weiblich	0.17427	1.19038	19.03795	0.01238	14.081	100.00	0.03918
V2 Alter	-0.21974	0.80273	-19.72717	0.01826	12.032	100.00	-0.03341

Ergebnisse fuer 4. Auspraegung "Tote" der abhaengigen Variablen V4 Unfallart
(als Referenz wird die 1. Auspraegung "Sachschaden" verwendet)

unabhaengige Variab	Regress. koeffiz.	"Risiko" exp(Regr.- koeffiz.)	relatives Risiko	Stand.- Fehler	z-Wert	Signifik. (1-p)*100	partielle Korrelat.
Konstante	-3.17914	-	-	0.08717	36.472	100.00	-
A1 Strasse: trocken	0.04501	1.04604	4.60418	0.04090	1.101	72.88	0.00248
A2 Strasse: nass	-0.12686	0.88086	-11.91419	0.04385	2.893	99.61	-0.00706
A3 Strasse: Eis	0.08185	1.08529	8.52880	0.05699	1.436	84.89	0.00070
B1 Geschlec:maennlic	0.05989	1.06172	6.17217	0.03540	1.692	90.91	0.00260
B2 Geschlec:weiblich	-0.05989	0.94187	-5.81336	0.03540	1.692	90.91	-0.00260
V2 Alter	0.08499	1.08871	8.87090	0.04600	1.848	93.54	0.00332

----- raus

Die Auspraegungen der untersuchten abhaengigen Variablen sind in dem Beispiel: Auspraegung 1=Sachschaden, 2=Leichtverletzte, 3=Schwerverletzte und 4=Tote. Die Effekte der Auspraegung k einer nominalen Variablen i hinsichtlich der Auspraegung j einer abhaengigen Variablen sind wie folgt zu interpretieren:

Bei Personen mit der Auspraegung k in der unabhaengigen Variablen i tritt die Auspraegung j der abhaengigen Variablen im Vergleich zur 1. Auspraegung der abhaengigen Variablen signifikant hoeufiger/geringer/gleich hoeufig auf wie:

- im Durchschnitt der Variablen i. Dies ist der Fall bei der 0|1|-1 -Kodierung der unabhaengigen nominalen Variablen
- im Vergleich zur letzten Auspraegung

Das hoeert sich sehr kompliziert an - und ist es auch. Die inhaltliche Interpretation der Regressionskoeffizienten und des "Risikokoeffizienten" bei der multinomialen Logitanalyse ist kompliziert und fuer den in der Logitanalyse nicht Geuebten verwirrend.

Betrachten wir beispielhaft den Effekt des Geschlechts. Dabei wollen wir den Risikokoeffizienten erlaeuern.

----- Ende raus

Risiko bei polytomer Zielvariabler

Der Begriff "Risiko", wie er im Almo verwendet wird, ist nicht durchgaengig in der Literatur anzutreffen. Gelegentlich wird auch von "Effekt" gesprochen (so bei Urban, 1993, S. 40) oder von "Chance" (so bei Tutz, 2000, S.60) oder einfach von "exp(β)".

Zuerst ist festzuhalten, dass sich die von Almo gelieferten Ergebnisse auf die 2., 3. und 4. Auspraegung der abhaengigen Variablen "Unfallart", also auf "Leichtverletzt" und "Schwerverletzt" und "Tote" beziehen. Die 1. Auspraegung "Sachschaden" ist die Bezugskategorie.

Risiko bei ursächlichen nominalen Variablen

Betrachten wir die beiden obersten Zeilen

unabhaengige Variable	Risiko exp(β)	relatives Risiko	Signifikanz (1-p)*100	partielle Korrelation
B1 Geschlec: Männer	0.87327	-12.67284	100.00	-0.03316
B2 Geschlec: Frauen	1.14512	14.51191	100.00	0.03316

Das "Risiko" ist

$$\exp(\beta)$$

Das "relative Risiko" ist

$$(\text{Risiko}-1) * 100$$

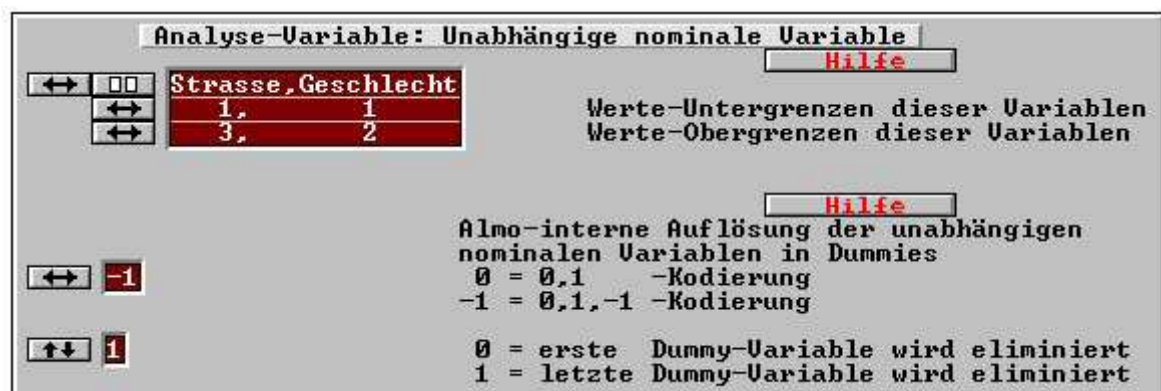
Die Männer haben - im Vergleich zum Durchschnitt aller Personen - eine um 12.67284 % verringerte Wahrscheinlichkeit einen Unfall mit Leichtverletzten zu erleiden, die Frauen eine um 14.51191 % erhöhte Wahrscheinlichkeit. Ein negatives Vorzeichen beim relativen Risiko bedeutet also - im Vergleich zum Durchschnitt - eine reduzierte Wahrscheinlichkeit, ein positives eine erhöhte Wahrscheinlichkeit.

Die Interpretation ist nicht vollständig. Sie vergisst zu erwähnen, dass sie sich auf die 1. Ausprägung "Sachschaden" der abhängigen Variablen "Unfallart" als Referenzgruppe bezieht. Wir werden darauf zurückkommen.

Entsprechend sind auch die relativen Risikowerte für den Straßenzustand zu interpretieren.

Diese Interpretation gilt im Falle der 0,1,-1 - Kodierung der Dummies der ursächlichen nominalen Variablen. Dies ist die Voreinstellung in Almo. Die Bezugsgruppe ist dabei die "Durchschnitts-Person".

Wird die 0,1 - Kodierung verwendet, dann wird (standardmäßig) die letzte Dummy, beim Geschlecht beispielsweise "weiblich", auf 0 gesetzt. Sie ist dann die Bezugsgruppe, mit der die Männer verglichen werden. In der 3. Box in Programm-Maske Prog22mb "Analyse-Variable: Unabhängige nominale Variable" wäre im 4. Eingabefeld eine "0" und im 5. Eingabefeld eine "1" eingetragen.



Wir erhalten in diesem Fall folgendes Ergebnis (gekürzt):

Ergebnisse für 2. Ausprägung "Leichtverlet" der abhängigen Variablen "V4 Unfallart" (als Referenz wird die 1. Ausprägung "Sachschaden" verwendet)

unabhaengige Variab	Regress. koeffiz.	"Risiko" exp(Regr.-koeffiz.)	relatives Risiko	Signifik. (1-p)*100
Konstante	0.09858	-	-	98.45
A1 Strasse: trocken	-0.24685	0.78126	-21.87383	100.00
A2 Strasse: nass	-0.32522	0.72237	-27.76337	100.00
B1 Geschlec:maennlic	-0.27102	0.76260	-23.73967	100.00
V2 Alter	-0.05281	0.94856	-5.14360	99.87

Die Bezugsgruppe erscheint nicht in der Ergebnis-Ausgabe. Die Männer haben - im Vergleich zu den Frauen ein um 23.73967 % verringerte Wahrscheinlichkeit einen Unfall mit Leichtverletzten zu erleiden.

Nun tritt ein Interpretationsproblem auf, das nur im Falle der polytomen abhängigen Variablen erkennbar wird. Betrachten wir nochmals die Männer im Vergleich zum Durchschnitt aller Personen. Unsere Interpretation muß, wenn sie vollständig und korrekt sein soll, folgendermaßen lauten:

Die Männer haben - im Vergleich zum Durchschnitt aller Personen - eine um 12.67284 % verringerte Wahrscheinlichkeit eher einen Unfall mit Leichtverletzten als einen Unfall mit Sachschaden zu erleiden

Die Frauen haben - im Vergleich zum Durchschnitt aller Personen - eine um 14.51191 % erhöhte Wahrscheinlichkeit eher einen Unfall mit Leichtverletzten als einen Unfall mit Sachschaden zu erleiden.

Wir haben zwei Bezugsgruppen:

- (1) Je eine auf Seiten der unabhängigen nominalen Variablen
- (2) und eine auf Seiten der abhängigen polytomen Variablen.

Die erstere ist in unserem Beispiel der Durchschnitt aller Personen (bzw. bei der 0,1 - Kodierung die letzte Ausprägung). Die zweite ist die 1. Ausprägung der abhängigen Variablen, in unserem Beispiel also der Unfall mit Sachschaden.

Entsprechend sind auch die Ergebnisse für die unabhängige Variable des Straßenzustands und für die anderen Ausprägungen der abhängigen Variablen "Unfallart" zu interpretieren.

Almo verwendet in unserem Datenbeispiel die erste Ausprägung als Bezugsgruppe der abhängigen Variablen. Die Ergebnisse für diese Bezugsgruppe, in unserem Beispiel der Unfall mit Sachschaden werden nicht ausgegeben. Natürlich interessieren auch diese. Wir müssen um diese Ergebnisse zu bekommen eine 2. Analyse rechnen, bei der wir beispielsweise den Unfall mit Leichtverletzten zur ersten Ausprägung und damit zur Bezugsgruppe machen.

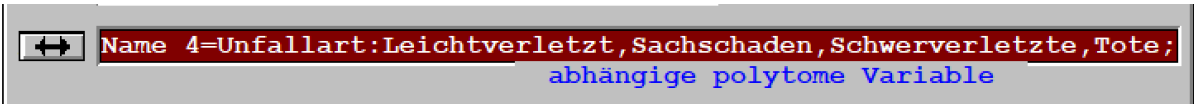
Vertauschen der 1. und 2. Ausprägung bei *tabellierten Daten*

In unserem Beispiel liegen die Daten als (bereits ausgezählte) Tabelle vor. Das macht erhebliche Probleme. Man muss die beiden Spalten "physisch" vertauschen. Aus der originalen Tabelle würde dann die folgende "vertauschte" Tabelle entstehen

originale Tabelle	Tabelle mit vertauschten Spalte 1 und 2
-----	-----

1 1 1	4037	2510	2042	212	1 1 1	2510	4037	2042	212
1 1 2	1043	912	805	37	1 1 2	912	1043	805	37
1 2 1	4981	2923	1833	258	1 2 1	2923	4981	1833	258
1 2 2	1530	1097	769	76	1 2 2	1097	1530	769	76
...
...
Sachschaden		Leichtverl.			Leichtverl.		Sachschaden		

Damit die den Codeziffern zugeordneten Ausprägungsnamen wieder stimmen, müssen auch die für die abhängige Variable "Unfallart vergebenen Ausprägungsnamen vertauscht werden. Es entsteht



Vertauschen der 1. und 2. Ausprägung bei *individuellen Daten*

Würden die Daten als aufeinander folgende individuelle Datensätze vorliegen, dann erreichen wir die Vertauschung sehr einfach dadurch, dass wir die abhängige Variable "Unfallart" in der Umkodierungsbox umkodieren

Die Ausprägung 2 (also Unfall mit Leichtverletzten) wird zu 1 und die seitherige Ausprägung 1 (also Unfall mit Sachschaden) wird zu 2. Die Ausprägungen 3 und 4 bleiben unverändert.

Der Eintrag lautet:

Unfallart (2=1; 1=2)

Damit die den Codeziffern zugeordneten Ausprägungsnamen wieder stimmen, schreiben wir in die Box "Freie Namensfelder", wie bereits oben gezeigt, folgende veränderte Namensgebung.

Name 4=Unfallart:Leichtverletzt,Sachschaden,Schwerverletzte,Tote;

Eine 2. Analyse ist eigentlich nicht notwendig. Denn selbstverständlich besteht ein eindeutiger Zusammenhang. Das bedeutet, dass die Ergebnisse der 2. Analyse aus denen der 1. Analyse leicht errechnet werden können. Siehe dazu auch Handbuch zu P45 "Almo Data-Mining", Abschnitt P45.16.2.1. bzw. (identisch) Almo-Dokument Nr. 25. Statistische Datenanalyse Teil II, Abschnitt P45.16.2.1.

Wir verwenden folgende Notation

R1= Risiko $ep_x(\beta)$ für unabh.Var. i hinsichtlich abh. Var. X2 bezüglich abh. Var. X1 aus 1. Analyse. Es ist 0.94539

Im Beispiel ist unabhäg. Var i= Strasse: trocken
 Zielvariable X2= Leichtverletzt
 Bezugsvariable X1= Sachschaden

R2= Risiko $ep_x(\beta)$ für unabh.Var. i hinsichtlich abh. Var. X1 bezüglich abh. Var. X2 aus 2. Analyse.

Im Beispiel ist unabhäg. Var i= Strasse: trocken
 Zielvariable X1= Sachschaden
 Bezugsvariable X2= Leichtverletzt

R2 kann nun leicht aus R1 errechnet werden. Es ist der Kehrwert von R1

$$\begin{aligned}
R2 &= 1 / R1 \\
&= 1 / 0.94539 \\
&= 1.05776
\end{aligned}$$

Risiko bei den ursächlichen quantitativen Variablen

Bei den unabhängigen quantitativen Variablen fällt die Interpretation leichter.

Betrachten wir das Alter.

Nimmt das Alter um 1 Einheit zu, dann verringert sich die Wahrscheinlichkeit, eher einen Unfall mit Leichtverletzten als einen Unfall mit Sachschaden zu erleiden um 5.14360 %. Wir haben hier also nur eine Bezugsgruppe auf Seiten der nominalen Zielvariablen.

Dabei ist es nun natürlich ausschlaggebend, in welchen Maßeinheiten das Alter gemessen wurde. In unserem Beispiel besitzt die Variable des Alters nur die 3 Ausprägungen "jung", "mittel" und "alt".

P22.2.6 Ergebnisse aus Logitanalyse ordinaler abhängiger Variable

Wir verwenden wieder das Beispielpogramm ARM102K.ALM, das wir bereits in Abschnitt P22.2.5 vorgestellt haben. Man findet das Programm nach Klick auf den Knopf „alle Progs“ am Oberrand des Almo-Fensters. In der Box "Analyse-Variable: Abhängige Variable" geben wir die abhängige Variable "Unfallart" nunmehr als ordinale Variable an.

Analyse-Variable: Abhängige Variable Hilfe

abhängige nominale Variable

ODER (exklusiv)

Unfallart

abhängige ordinale Variable

1

4

Werte-Untergrenzen der abhäng. Variablen
Werte-Obergrenzen der abhäng. Variablen

Als unabhängige Variable werden eingesetzt:

Analyse-Variable: Unabhängige nominale Variable Hilfe

↔	□□	Strasse, Geschlecht
↔	↔	1, 1
↔	↔	3, 2

Werte-Untergrenzen dieser Variablen
Werte-Obergrenzen dieser Variablen

Hilfe

Almo-interne Auflösung der unabhängigen nominalen Variablen in Dummies

0 = 0,1 -Kodierung
-1 = 0,1,-1 -Kodierung

↔ -1

↑↓ 1

0 = erste Dummy-Variable wird eliminiert
1 = letzte Dummy-Variable wird eliminiert

Analyse-Variable: Unabhängige quantitative Variable Hilfe

↔ □□ Alter

Almo liefert folgende Ausgabe (gekürzt):

Nr	unabhaengige Variable	Regress. Koeffiz.	Standard Fehler	z-Wert	Signifik. (1-p)*100	exp(Regr.-koeffiz.)	Risiko in %
1	Konstante	0.47028	0.02366	19.873	100.00	-	-
2	A1 Strasse: trocken	-0.04109	0.01105	3.719	99.98	0.95974	-4.02604
3	A2 Strasse: nass	-0.14291	0.01166	12.259	100.00	0.86683	-13.31681
	A3 Strasse: Eis	0.18400	0.01483	12.410	100.00	1.20202	20.20200
4	B1 Geschlec:maennlic	-0.12121	0.00888	13.652	100.00	0.88584	-11.41553
	B2 Geschlec:weiblich	0.12121	0.00888	13.652	100.00	1.12887	12.88660
5	V2 Alter	-0.12587	0.01294	9.731	100.00	0.88173	-11.82698
6	alfa2 (Schwellenwert α_2)	1.31699	0.00913	144.265	100.00	-	-
7	alfa3 (Schwellenwert α_3)	3.93903	0.02846	138.395	100.00	-	-

Beobachtete und durch das Modell reproduzierte (prognostizierte) Haeufigkeiten

die unabhaengigen nominalen Variablen sind
A = V1 Strasse
B = V3 Geschlecht
ihre Auspraegungen werden mit 1,2,3,... durchnummeriert

die unabhaengigen quantitativen Variablen sind
quant1 = V2 Alter

beo1 ... = beobachtete Haeufigkeiten in der Auspraegung 1 der abhaeng. Variablen
rep1 ... = reproduzierte (prognostizierte) Haeufigkeiten in der Auspraegung 1 der abhaeng. Variablen

Nr.	A	B	quant1	beo1	beo2	beo3	beo4	rep1	rep2	rep3	rep4
1	1	1	1.000	4037	2510	2042	212	4000.9	2659.1	1940.1	200.9
2	1	2	1.000	1043	912	805	37	1106.0	878.2	732.0	80.8
3	1	1	2.000	4981	2923	1833	258	4857.0	2930.6	2005.6	201.7
4	1	2	2.000	1530	1097	769	76	1478.7	1072.0	832.5	88.8
5	1	1	3.000	956	591	424	67	1054.5	576.0	371.1	36.3
.
.
.
.

Der Konstanteneffekt und die Effekte A1 etc. können wie beim binären Logit- oder Probitmodell interpretiert werden. Zusätzlich werden *Schwellenwerte* für die ordinalen Antwortkategorien berechnet. In obiger Tabelle der Regressionskoeffizienten wurden sie mit α_2 und α_3 bezeichnet. In den nachfolgenden Formeln bezeichnen wir sie kurz mit α . Dann ist α_2 der Schwellenwert zwischen der Ausprägung 2 und 3, α_3 jener zwischen der Ausprägung 3 und 4. Der Schwellenwert α_1 zwischen der Ausprägung 1 und 2 wird gleich 0 gesetzt. Also gibt ihn deswegen gar nicht aus. Da die Antwortkategorien geordnet sind, muss α_2 größer α_1 sein, also größer 0, und α_3 größer α_2 . Ist diese Bedingung nicht erfüllt, kann die abhängige Variable nicht als ordinalskaliert betrachtet werden.

Die Schwellenwerte sind bedeutungslos, wenn es darum geht, festzustellen, wie stark die unabhängigen Variablen die abhängige polytome Variable determinieren. Sollen jedoch die Wahrscheinlichkeiten einer Person (in unserem Beispiel) in die 1. oder 2. oder 3. oder 4. Ausprägung der Variablen "Unfallart" zu fallen prognostiziert bzw. reproduziert werden, dann benötigen wir die Schwellenwerte.

Wir wollen für die 1. Personengruppe die 4 Wahrscheinlichkeiten berechnen. Diese Personengruppe besitzt folgende Werte in den unabhängigen Variablen

Konstante
Strassenzustand: 1
Geschlecht: 1
Alter: 1

Wir erhalten also, ohne die Schwellenwerte zu berücksichtigen, folgenden Wert V :

$$V = 0.47028 - 0.04109 * 1 - 0.12121 * 1 - 0.12587 * 1 = 0.18211$$

Konstante	A1	B1	Alter

Wir können indem wir die Schwellenwert berücksichtigen, die Wahrscheinlichkeiten berechnen, dass diese Personengruppe in die 1. oder 2. oder 3. oder 4. Ausprägung der Variablen "Unfallart" fallen.

Die allgemeine Formel lautet:

$$(0) \quad p_j = \frac{e^{(\alpha_j - V)}}{1 + e^{(\alpha_j - V)}} - \frac{e^{(\alpha_{j-1} - V)}}{1 + e^{(\alpha_{j-1} - V)}}$$

Für p_1 , die Wahrscheinlichkeit die Unfallart „Sachschaden“ zu erleiden ist der Schwellenwert $\alpha_1 = 0$. Also gibt ihn schon gar nicht aus. α_{1-1} existiert nicht. Die Formel für die 1. Ausprägung vereinfacht sich also zu

$$(1) \quad p_j = \frac{e^{(0 - V)}}{1 + e^{(0 - V)}} = \frac{e^{-0.18211}}{1 + e^{-0.18211}} = 0.4546$$

Die obige allgemeine Formel (0) wird nun für die 2. und 3. Ausprägung (Leichtverletzte und Schwerverletzte) verwendet.

$$(2) \quad p_2 = \frac{e^{(1.31699-0.18211)}}{1+e^{(1.31699-0.18211)}} - \frac{e^{(0-0.18211)}}{1+e^{(0-0.18211)}} = 0.30214$$

$$(3) \quad p_3 = \frac{e^{(3.93903-0.18211)}}{1+e^{(3.93903-0.18211)}} - \frac{e^{(1.31699-0.18211)}}{1+e^{(1.31699-0.18211)}} = 0.22044$$

Die letzte, in unserem Fall die 4. Ausprägung (Tote), ist dann

$$(4) \quad p_4 = 1 - \frac{e^{(3.93903-0.18211)}}{1+e^{(3.93903-0.18211)}} = 1 - 0.97718 = 0.0228$$

In der 1. Personengruppe befinden sich insgesamt

$$4037 + 2510 + 2042 + 212 + 212 = 8801$$

Personen. Diese Zahl muß mit den 4 Wahrscheinlichkeiten multipliziert werden. Wir erhalten damit die reproduzierten Häufigkeiten für die 4 Ausprägungen, die Almo ausgibt, also

$$4000.9 \quad 2659.1 \quad 1940.1 \quad 200.9$$

Zur Interpretation der Schwellenwerte siehe die ausführliche Darstellung bei D. Urban (1993, S. 88ff).

P22.2.6.1 Schwellenwert bei SPSS

Betrachten wir nochmals den Ausdruck $(\alpha - \mathbf{V})$ im Zähler der Formel (0). V enthält neben den (gewichteten) Regressionskoeffizienten die *Konstante* K . In unserem Beispiel hat sie den Wert 0.47028 . Mit \mathbf{V}^* bezeichnen wir den verbleibenden V -Wert, nachdem die Konstante herausgenommen wurde. Obiger Ausdruck kann dann auch so geschrieben werden:

$$(\alpha - \text{Konstante} - \mathbf{V}^*)$$

Bei SPSS (und auch teilweise in der Literatur zur ordinalen Logitanalyse) wird zusammengefasst

$$S = \alpha - \text{Konstante}$$

Ausgegeben wird dann:

$$\begin{aligned} \text{Schwelle Unfallart 1 Leichtverletzt: } S_1 &= \alpha_1 - K = 0 - 0.47028 = 0.47028 \\ \text{Schwelle Unfallart 2 Sachschaden: } S_2 &= \alpha_2 - K = 1.31699 - 0.47028 = 0.84671 \\ \text{Schwelle Unfallart 3 Schwerverletzt: } S_3 &= \alpha_3 - K = 3.93903 - 0.47028 = 3.46875 \end{aligned}$$

Damit wird eine sehr sinnvolle Interpretation der Ergebnisse der ordinalen Logitanalyse nahe gelegt:

Im Unterschied zur multinomialen Logitanalyse, die für jede Kategorie der abhängigen Variablen einen anderen Satz von Regressionskoeffizienten ausgibt, wird bei der ordinalen Logitanalyse für alle Kategorien ein gemeinsamer Satz von Regressionskoeffizienten geliefert. Spezifisch für die Kategorien der abhängigen Variablen sind dann nur die beschriebenen Schwellenwerte S . Diese können als „modifizierte Konstante“ begriffen werden.

Literatur

- ARMINGER, G.:** Statistische Verfahren zur Analyse qualitativer Variablen.
KÜSTERS, U.: Bergisch Gladbach, 1986
GREENE, W.H.: Econometric Analysis. New-York, London, S. 661-714.

- 1990
- MADDALA, G.S.:** Limited-Dependent and Qualitative Variables in Econometrics. Cambridge, 1990
- TUTZ, G.:** Die Analyse kategorialer Daten, Oldenbourg Verlag, München, Wien, 2000
- URBAN, D.:** Logit-Analyse, Gustav Fischer Verlag, Stuttgart, 1993
- Michael Windzio** Regressionsmodelle für Zustände und Ereignisse, Kap.9, 2013, VS Verlag für Sozialwissenschaften, Springer VS