Koeffizienten der Logitanalyse

P22.5

Kurt Holm

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

2014

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

Siehe auch das Almo-Dokument Nr. 9 "Logitanalyse".
Im Text wird häufig auf das Dokument **P0** Bezug genommen. Dabei handelt es sich um das Almo-Dokument "Arbeiten mit Almo.PDF" (Dokument 0).

Weitere Almo-Dokumente

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

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

Kurt Holm

P22.5 Koeffizienten der Logitanalyse

Wie sind die im Rahmen der Logit-Analyse errechneten Regressionskoeffizienten und Risiko-Koeffizienten zu interpretieren ?

Im 1. Teil der folgenden Darstellung wird diese Frage für die binäre (dichotome) Logitanalyse und im 2. Teil für die multinomiale (polytome) Logitanalyse beantwortet. Große Teile des Textes sind, etwas erweitert und verändert, aus dem Dokument Nr.9 "Logitanalyse" entnommen.

P22.5.1 Die binäre Logitanalyse

Betrachten wir ein Beispiel:
Kunden kaufen auf Kredit. Werden sie ihren Kredit zurückzahlen ?
Die Variablen für unser Beispiel sollen folgende sein:

```
Die Zielvariable ist     Kredit-Rückzahlung: nein,
                                             ja
```

Die Zielvariable soll durch folgende Variablen determiniert sein:

```
Die unabhängigen nominalen Variablen sind
                                      Wohnort: Stadt
                                               Land

                                   Hausbesitz: kein Haus
                                               hat Haus

Die unabhängigen quantitativen Variablen sind
                                               Einkommen
                                               Rückzahlungsrate
                                               Kredit-Laufzeit
```

Die nominalen unabhängigen Variablen werden mit 0, 1, -1 kodiert. Bei dieser Kodierungsart werden *alle* Dummies verwendet. Ihre Regressionskoeffizienten summieren sich zu 0. Bei der zweiten Kodierungsweise, der 0, 1 –Kodierung wird die letzte Dummy auf 0 gesetzt. Wir werden diesen Fall weiter unten darstellen.

Almo liefert folgende Ausgabe (verkürzt).

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

    unabhaengige Variable   Regress.     Risiko    relatives  Signifikanz   partielle
                            koeff.ß      epx(ß)    Risiko      (1-p)*100    Korrelation
-------------------------------------------------------------------------------------
    A1    Wohnort:   Stadt  -0.43493    0.64731    -35.26902      100.00    -0.13168
    A2    Wohnort:    Land   0.43493    1.54486     54.48553      100.00     0.13168

    B1   Hausbesi:kein Hau  -0.74569    0.47440    -52.55955      100.00    -0.15825
    B2   Hausbesi:hat Haus   0.74569    2.10791    110.79059      100.00     0.15825
```

```
   V4   Einkommen          0.68943    1.99257     99.25744      100.00     0.25486
   V7   Rueckrate         -0.00077    0.99923     -0.07689      100.00    -0.25619
   V8   Laufzeit           0.04562    1.04667      4.66727       99.22     0.06526
-------------------------------------------------------------------------------------
```

Betrachten wir die beiden Regressionskoeffizienten für den Wohnort

```
A1  Wohnort:   Stadt -0.43493
A2  Wohnort:    Land  0.43493
```

Das Logit-Modell lautet

$$p_1 = \frac{1}{1+e^{-(c+a(i)+b(j)+\beta_1 \cdot E+\beta_2 \cdot R+\beta_3 \cdot L)}}$$

Diese Gleichung kann so umgewandelt werden, daß auf der rechten Seite ein linearer Ausdruck steht

(1) ln(p1/p2) = c + a(i) + b(j) + ß1*E + ß2*R + ß3*L

p1=Wahrscheinlichkeit für Kreditkauf: ja
p2=Wahrscheinlichkeit für Kreditkauf: nein (p2=1-p1)
Natürlich gilt: p2 = 1-p1
e =e-Zahl 2.718
c =Konstante

a(i) bezeichnet die Regressionskoeffizienten für die Dummy-Variablen des Wohnorts
b(j) bezeichnet die Regressionskoeffizienten für die Dummy-Variablen des Hausbesitz

es ist also:

a1=Regressionskoeffizient für "Stadt"
a2=Regressionskoeffizient für "Land"

E =Einkommen
ß1=Regressionskoeffizient für Einkommen

R =Rueckrate
ß2=Regressionskoeffizient für Rueckrate

L =Laufzeit
ß3=Regressionskoeffizient für Laufzeit

P22.5.1.1 Regressionskoeffizienten der nominalen Variablen

Der Regressionskoeffizienten a1=-0.43493 für "Stadt" und a2=0.43493 für "Land" haben folgende Bedeutung:

1. Das negative Vorzeichen von a1 drückt aus, daß Städter im Vergleich zur "Durchschnittsperson" in der Variablen "Wohnlage" das logarithmierte Wahrscheinlichkeitsverhältnis ln(p1/p2) aus Gleichung 1 verringern. Vereinfacht: Städter haben eine geringere Wahrscheinlichkeit ihren Kredit zurückzuzahlen. Umgekehrt drückt das positive Vorzeichen von a2 aus, daß Leute vom Land eine erhöhte Wahrscheinlichkeit haben ihren Kredit zurückzuzahlen.

2. Je (absolut) größer der Regressionskoeffizient ist, umso stärker ist diese Tendenz.

P22.5.1.2 Regressionskoeffizienten der quantitativen Variablen

Der Regressionskoeffizient ß1=0.68943 für "Einkommen" hat folgende Bedeutung: Wenn sich das Einkommen um 1 Einheit erhöht, dann erhöht sich das logarithmierte Wahrscheinlichkeitsverhältnis ln(p1/p2). Vereinfacht: Wenn sich das Einkommen um 1 Einheit erhöht, dann nimmt die Wahrscheinlichkeit zu, den Kredit zurückzuzahlen. Ein negatives Vorzeichen würde bedeuten, dass sich die Wahrscheinlichkeit verringert. Je (absolut) größer der Regressionskoeffizient ist, umso stärker ist diese Tendenz.

P22.5.1.3 Der Risiko-Koeffizient exp (ß)

Anstelle des Begriffs "Risiko-Koeffizient", den wir hier verwenden, wird in der Literatur auch der Begriff "Effekt-Koeffizient" gebraucht (so bei D. Urban, 1993).

Unser Beispiel ist relativ komplex. Wir haben 2 ursächliche nominale Variable und 3 ursächliche quantitative Variable.

Um unsere Erläuterung übersichtlich gestalten zu können, wollen wir ein anderes, einfacheres Beispiel betrachten, bei dem nur 1 ursächliche nominale und 1 ursächliche quantitative Variable vorhanden ist.

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:

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

 unabhängige Variable          Regress.    "Risiko"    relatives
                               Koeffiz.    exp(Regr.-  Risiko
                                           koeffiz.)   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:

$$p1 = \frac{1}{1 + e^{-(c + a(i) + \beta \cdot x)}}$$

Man beachte:
p1 ist die Wahrscheinlichkeit für die Ausprägung "ja" der Zielvariablen "Rückzahlung". Mit p2 werden wir die Wahrscheinlichkeit für die Referenzkategorie "nein" bezeichnen

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

(1) ln(p1/p2) = c + a(i) + ßX

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

a(i) bezeichnet die Regressionskoeffizient für die 3
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
ß =Regressionskoeffizient für Einkommen

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

```
(1a)  ln(p1/p2) =   c  +   a1 +   ßX
                = 1.88 + 1.38 - 0.38*4
```

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

(2) p1/p2 = exp(c) * exp(a(i)) * exp(ß*X)

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

Für unseren Arbeiter mit Einkommen X=4

```
(2a) p1/p2 = exp(c)    * exp(a1)   * exp(ß*X)
           = exp(1.88) * exp(1.38) * exp(-0.38*4)
           =   6.62    *   3.96    *  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

Gewinn-zu-Verlust-Verhältnis ("odds")

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. Daraus folgt: Ist p1/p2 positiv, dann ist der Gewinn größer. Ist p1/p2 negativ, dann ist der Verlust größer

Ist p1=0.6666.. dann ist p2=0.33333... 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:

```
                    p1 = f / (1+f)
                       = 5.7886 / (1+5.7886)
                       = 0.853
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

Betrachten wir nun wieder Gleichung 2 bzw. 2a. Alle Arbeiter haben - im Vergleich zum Durchschnitt aller "Berufstätigen", genauer: der Durchschnitt aller Untersuchungspersonen in der Variablen "Beruf" - 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, Gustav Fischer, Stuttgart, 1993).

Wäre exp(a1)=1, dann würden sich die Arbeiter so verhalten wie der Durchschnitt.aus allen "Berufstätigen"

Wir definieren nun als

```
relatives Risiko" = (exp(a(i)) - 1) * 100
```

Für die Arbeiter finden wir dann

```
relatives Risiko = (exp(a1) - 1) * 100
                 = (3.96    - 1) * 100
                 = 296
```

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

Zu beachten ist, dass die Bezugskategorie der Durchschnitt aller Untersuchungspersonen in der Variablen "Beruf" 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.

Die 0,1 -Kodierung der unabhängigen nominalen Variablen

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

Almo liefert folgendes Ergebnis (verkürzt):

Ergebnisse für 2. Auspräg. "ja" der abhängigen Variablen "Rückzahlung"

	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 in der betrffenden unabhängigen Variablen**

Risiko bei quantitativen Variablen

Betrachten wir nochmals obige Gleichung (2)

(2) p1/p2 = exp(c) * exp(a(i)) * exp(ß*X)

Das Einkommen unseres Arbeiters ist X=4.
Der Ausdruck **exp(ß*X)** ist also **exp(–0.37586 * 4) = 0.22236**

Wenn sich das Einkommen dieser Person um 1 Einheit erhöht, dann
ist der Ausdruck **exp(ß*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(ß).

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" 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(ß) zu.

Wir können diese Zunahme bzw. Abnahme auch in Prozentwerten ausdrücken. Sie beträgt dann **100*(exp(ß)–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 p1/p2 einer nachfolgenden Einkommensstufe entsteht durch Multiplikation mit exp(ß)=0.6867 des Wahrscheinlichkeits-Verhältnis p1/p2 der vorhergehenden Einkommensstufe.

P22.5.2 Die multinomiale Logitanalyse

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 und dessen Ergebnisse und Koeffizienten ebenfalls in gleicher Weise zu interpretieren sind.

.

Die abhängige polytome Variable ist:

```
V4 Unfallart:(1)Sachschaden,(2)Leichtverletzt,(3)Schwerverletzte,
             (4)Tote
```

Die unabhängigen nominalen Variablen sind:

```
V3 Geschlecht: maennlich, weiblich
V1 Straßenzustand: trocken,nass,Eis
```

Die unabhängigen quantitative Variable ist:

```
V2 Alter: jung,mittel,alt
```

(Diese Variable besitzt also nur 3 Ausprägungen)

Bei der multinomialen Logitanalyse *gruppierter* Daten wird von Almo zwangsweise die erste Ausprägung als Referenzausprägung fixiert. 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 Individualdaten einlesen. Anders bei der Programm-Maske Prog22mb, die eine schon ausgezählte, fertige Tabelle *gruppierter Daten* einliest, dort wird die erste Ausprägung zwangsweise als Referenz verwendet. Das hat Almo-interne Gründe.

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

```
ln(p2/p1)    ln(p3/p1)    ln(p4/p1)
```

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 P22.5.2.2

ausführen.

P22.5.2.1 Eingabe mit gruppierten Daten

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

			Unfallart (Fallzahl)			
V1	V2	V3	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

BEACHTE 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. Das vollständige Programm ist als Beispielprogramm ARM102K.ALM in Almo enthalten. Der Benutzer findet das Programm durch Klicken auf das Menü "alle Progs".

Wir wollem 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.

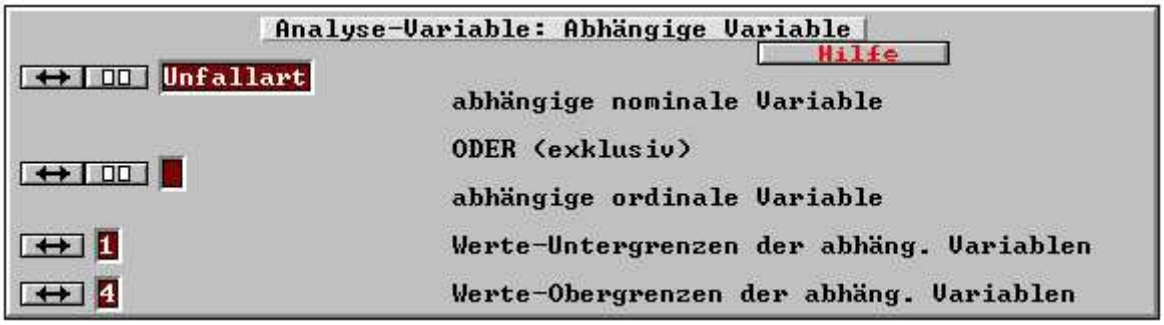
In der 1. Spalte der Tabelle steht der Straßenzustand.
Also erhält "Strasse" die Variablennummer 1.

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

Box 5: Abhängige polytome Variable



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.     "Risiko"     relatives      Stand.- z-Wert  Signifik.  partielle
                      koeffiz.     exp(Regr.-   Risiko         Fehler          (1-p)*100  Korrelat.
                                   koeffiz.)
-------------------------------------------------------------------------------------------------
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

Interpretation bei polytomer Logitanalyse Teil 1

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

Bei der 0|1|-1 -Kodierung der unabhängigen nominalen Variablen die wir verwendet haben:
Bei Personen mit der Ausprägung k in der unabhängigen Variablen i tritt die Ausprägung j der abhängigen Variablen im Vergleich zur 1. Ausprägung der abhängigen Variablen signifikant häufiger/geringer oder gleich häufig wie im Durchschnitt 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.

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

P22.5.2.2 Risiko bei polytomer Zielvariabler

Der Begriff "Risiko", wie er im Almo verwendet wird, ist nicht durchgängig 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(ß)".

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

Risiko bei ursächlichen nominalen Variablen

Betrachten wir die beiden obersten Zeilen

```
    unabhaengige Variable       Risiko     relatives    Signifikanz  partielle
                                epx(ß)     Risiko       (1-p)*100    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(ß)
```

Das "relative Risiko" ist

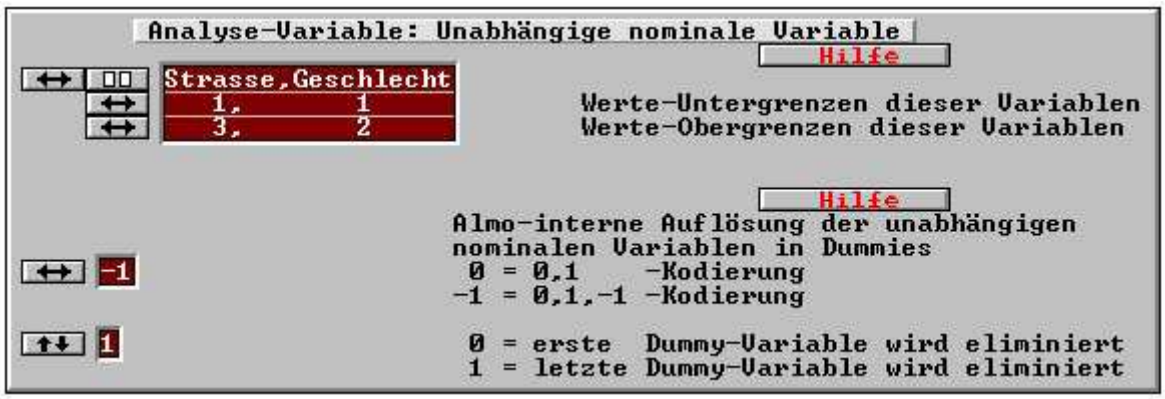
```
(Risiko-1)*100
```

Die Männer haben - im Vergleich zum Durchschnitt aller Personen in der Variablen "Geschlecht" - 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" in der betreffenden unabhängigen Variablen.

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. Auspraegung "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.

P22.5.2.3 Interpretation bei polytomer Logitanalyse

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 in der Variablen "Geschlecht". Unsere Interpretation muß, wenn sie vollständig und korrekt sein soll, folgendermaßen lauten:

Die Männer haben - im Vergleich zum Durchschnitt aller Personen in der Variablen "Geschlecht" - 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 in der Variablen "Geschlecht" 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 in der unabhängigen Variablen (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.

P22.5.3 Referenzkategorie ändern durch Vertauschen der Ausprägungen

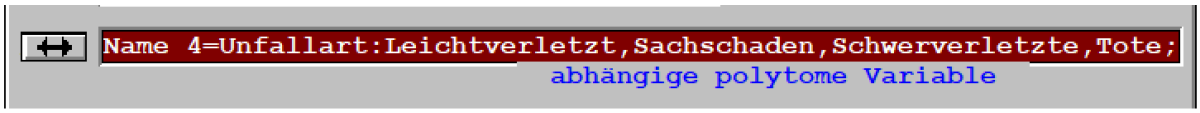
Almo verwendet in unserem Datenbeispiel die erste Ausprägung als Bezugsgruppe der abhängigen Variablen. Der Benutzer kann das nicht ändern. Die Ergebnisse für diese Bezugsgruppe, in unserem Beispiel der Unfall mit Sachschaden werden nicht ausgegeben. Natürlich interessieren auch diese. Wir müssen um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

P22.5.3.1 Vertauschen der 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 Ausprägungs-namen vertauscht werden. Es entsteht



P22.5.3.2 Vertauschen der Ausprägung bei individuellen Daten

Würden die Daten als aufeinander folgende indiduelle 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 epx(ß) 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 epx(ß) 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

```
         R2 = 1 / R1
            = 1 / 0.94539
            = 1.05776
```

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".

Literatur:

Arminger, Küsters: Statistische Verfahren zur Analyse qualitativer Variablen, Bergisch Gladbach, 1986
Dieter Urban: Logit-Analyse, Gustav Fischer, Stuttgart, 1993