Einführung in das Arbeiten mit GLIM zur Analyse mehrdimensionaler Kontingenztafeln mittels loglinearer und Logit-Modelle

Forschungsberichte der Bundesanstalt für Straßenwesen
Bereich Unfallforschung

Einführung in das Arbeiten mit GLIM zur Analyse mehrdimensionaler Kontingenztafeln mittels loglinearer und Logit-Modelle

Gabriele Ernst
Ekkehard Brühning

Bundesanstalt für Straßenwesen
Bereich Unfallforschung
Bergisch Gladbach, Januar 1987

Herausgeber:
Bundesanstalt für Straßenwesen
Bereich Unfallforschung
5060 Bergisch Gladbach 1, Brüderstr. 53
Tel. 02204/430, Telex 8878483 bas d

Druck: Fotodruck J. Mainz, 5100 Aachen

Lfd. Nr. 148

ISSN 0173 - 7066

INHALT

0. VORWORT

Multivariate Analyseverfahren haben in den letzten Jahren in den empirischen Wissenschaften ein immer breiteres Anwendungsfeld gefunden, wobei die Vielfalt der Methoden für den Forschungspraktiker kaum noch zu überblicken ist. Mit diesem Werkbuch wird der Versuch unternommen, verschiedene Methoden zur Analyse von kreuztabellierten Daten mit Hilfe des Programms GLIM (Generalized Linear Interactive Modelling) darzustellen; es soll dem Forschungspraktiker beim Einstieg in die Anwendung der Verfahren helfen.

Bisher sind sowohl der von Nelder und Wedderburn (1972) entwickelte statistische Ansatz der verallgemeinerten linearen Modelle, als auch das auf diesem Ansatz basierende Programm GLIM in der praktischen Anwendung noch relativ wenig verbreitet. Hierfür mag es viele Gründe geben, die wichtigsten sind vermutlich, daß der Grad der statistischen Formalisierung relativ hoch und der Umfang der Programmdokumentation eher niedrig ist. Vor allem fehlt es an konkreten Beispielen, wie man mit einem Datensatz z.B. ein loglineares oder Logit-Modell berechnet.

Vor diesem Hintergrund ist das vorliegende Werkbuch entstanden. Es ist daher keine Beschreibung aller Möglichkeiten, die GLIM bietet, sondern eine Dokumentation von Vorgehensweisen und Erfahrungen. Es beruht auf praktischen Anwendungen, die in der Bundesanstalt für Straßenwesen, Bereich Unfallforschung, im Jahre 1985 bei den Arbeiten der Fachgruppe U4.3 "Statistik" gemacht wurden (vgl. Brühning, Dilling, Ernst u. Schmid, 1986) und berücksichtigt zusätzlich die GLIM-Version Release 3.77 (PC-Version). Die Verfasser danken Herrn Dipl.-Wirtschaftsing. M. Schmid für zahlreiche Hinweise, die zur Verbesserung des Werkbuchs beigetragen haben.

Der Anwender findet in diesem Werkbuch eine Darstellung der erforderlichen Arbeitsschritte. Zunächst wird der Prozeß der Datenbereitstellung und -übergabe nach GLIM beschrieben, wie er sich bei den Arbeiten der BASt-Unfallforschung bewährt hat. Sodann werden diejenigen GLIM-Anweisungen an Beispielen angewendet, die der Benutzer benötigt, um Daten selbständig mittels einfacher

loglinearer bzw. Logit-Modelle zu analysieren. Die vollständige Dokumentation aller GLIM-Anweisungen findet der Anwender im GLIM-Manual[1]; eine vergleichsweise ausführliche Kommandoübersicht mit deutschsprachigen Erläuterungen (ohne Beispiele) ist im Anhang 7 zusammengestellt.

Zur statistischen Theorie der Logit-Modelle und ihrer Anwendung wird auf Arminger (1986) verwiesen. In Anhang 6 ist zur Verdeutlichung einiger grundlegender statistischer Zusammenhänge ein von den Verfassern erarbeitets Referat auszugsweise wiedergegeben. Zum Einsatz der angesprochenen Modelle wird ferner auf den Tagungsbericht des Seminars "Multivariate Analyse mittels loglinearer Modelle - Ein Analyseinstrument für die Verkehrs- und Unfallforschung" verwiesen (DVWG, 1987).

[1] Das Programm **GLIM** wird von der Numerical Algorithmus Group, NAG Central Office, Mayfield House, 256 Banburg Road, Oxford OX2 7DE; U.K. vertrieben; neben Großrechner-Versionen steht auch eine PC-Version zur Verfügung. Um GLIM auf einem PC einsetzen zu können, müssen die folgenden Hardware- bzw. Softwarevoraussetzungen erfüllt sein: 640 KB, Numeric Co-Prozessor, zwei Diskettenlaufwerke oder ein Diskettenlaufwerk und Festplatte ; als Betriebssystem muß MS-DOS bzw. PC-DOS ab Version 2.0 im Einsatz sein.

1. VERWENDETE DATEN

In diesem Werkbuch soll das Arbeiten mit GLIM an zwei Beispielen der multivariaten Kontingenztafelanalyse verdeutlicht werden. In beiden Beispielen wird die gleiche Datenbasis für jeweils unterschiedliche Modelle verwendet:

1. Beispiel: Zur Erklärung der Zahl der Fahrunfälle durch die qualitativen unabhängigen Variablen: Alter des Beteiligten (A), Geschlecht (G), Alkoholeinfluß (AL) und Straßenklasse (SK) wird ein loglineares Modell ermittelt.

2. Beispiel: Zur Erklärung des Anteils der Fahrunfälle durch die qualitativen unabhängigen Variablen: Alter des Beteiligten (A), Geschlecht (G), Alkoholeinfluß (AL), und Straßenklasse (SK) wird ein Logit-Modell bestimmt.

Die in die multivariate Analyse eingeschlossenen Variablen und ihre Ausprägungen sind:

A = Alter des Beteiligten
- A1 25 bis unter 60 Jahre
- A2 unter 25 Jahre
- A3 60 Jahre und mehr

G = Geschlecht des Beteiligten
- G1 männlich
- G2 weiblich

AL = Alkoholeinfluß beim Beteiligten
- AL1 ohne Alkohol
- AL2 mit Alkohol

SK = Straßenklasse
- SK1 Landesstraßen
- SK2 Bundesstraßen
- SK3 Kreis-, Gemeindestraßen

2. ERZEUGUNG DER GLIM-EINGABE-KONTINGENZTAFEL

Die mehrdimensionale Kontingenztafel, welche mit GLIM analysiert werden soll, kann mit Hilfe des Statistik-Programm-Systems SPSS erstellt werden.
SPSS bietet hierzu drei Prozeduren an:

WRITE CASES
AGGREGATE
CROSSTABS

In diesem Werkbuch soll die Verwendung der Prozedur CROSSTABS beschrieben werden, da sie sich in guter Weise für die Verarbeitung von Massendaten eignet, wie sie z.B. im Arbeitsbereich der Verfasser anfallen. Die Prozedur CROSSTABS bietet im Integermode in Verbindung mit der Anweisung "RAW OUTPUT UNIT 15" die Möglichkeit, die Zellenhäufigkeiten der Kontingenztafel in eine externe Datei zu schreiben.
Die entsprechenden Anweisungen [2] für das in diesem Werkbuch verwendete Beispiel lauten:

```
RAW OUTPUT UNIT 15
CROSSTABS VARIABLES = ALTBET (1,3)          [A]
                      GESCHL (1,2)          [G]
                      UNFALL (1,2)          [AL]
                      STRKLASS (1,3)        [SK]
                      UNFTYP (1,2)/         [Fahrunfall ja/nein]
            TABLES = STRKLASS BY
                     UNFALL BY GESCHL BY
                     ALTBET BY UNFTYP
OPTIONS              11
```

In diesem Zusammenhang sei noch vermerkt: die Prozedur CROSSTABS hat im hier verwendeten Integermode die Beschränkung, daß in der TABLES-Anweisung nur 8 Variablen mit BY miteinander verknüpft werden können. Soll die zu analysierende Kontingenztafel mehr als 8 Variablen (Merkmale) enthalten, gibt es die Möglichkeit, mit der Datenselektionsanweisung: SELECT IF über eine oder mehrere Variablen (Merkmale) zu selektieren. Man enthält dann 2...n externe

[2] (vgl. Anhang 1 (SPSS-JOB) und für SIEMENS-Benutzer in BS2000 Anhang 2 (SPSS-ENTER-DATEI))

Ausgabefiles [3], die mit einem Editor (z.B. EDOR) bearbeitet und zu einer Kontingenztafel zusammengefügt werden können.

Die aus den SPSS-Auswertungen entstandene Ergebnisdatei (eine ISAM-Datei), die die Kontingenztafel beschreibt, ist wie folgt aufgebaut (vgl. Anhang 3):
die ersten 8 Spalten enthalten den ISAM-Schlüssel, gefolgt von zwei Spalten, die lediglich "1" enthalten; sie werden SPSS-intern erzeugt. Die darauf folgende Spalte enthält die Zellenhäufigkeiten. Dann folgen die Ausprägungen der in der TABLES-Anweisung durch BY verknüpften Variablen; STRKLASS - 3 Ausprägungen; UNFALL - 2 Ausprägungen; GESCHL - 2 Ausprägungen; ALTBET - 3 Ausprägungen und UNFTYP - 2 Ausprägungen.

Diese Datei kann vor dem Einlesen nach GLIM noch mit Hilfe von Dienstprogrammen, die in größeren Rechenzentren meist verfügbar sind, oder durch ein selbst zu erstellendes kleines Programm bearbeitet werden.

In Abschnitt 3.1 wird die Handhabung eines entsprechendes FORTRAN-Programms ("TRANS.SPSS") dargestellt, das die SPSS-Ergebnisdatei vor dem Einlesen nach GLIM optimiert. Es ist jedoch auch möglich, die SPSS-Ergebnisdatei direkt mit GLIM weiterzuverarbeiten, in diesem Fall muß die Definition der Eingabedaten entsprechend der SPSS-Ausgabe definiert werden; d.h. die zwei Spalten der SPSS-Ergebnisdatei, die lediglich "1" enthalten müssen mit einem Variablennamen versehen werden; sie werden in GLIM als Variablen behandelt. In die Analyse werden diese Variablen jedoch nicht einbezogen.

Die Analyse von binomialen Logit-Modellen zur Modellierung von Anteilswerten ist mit der von SPSS ausgegebenen Ergebnisdatei ohne weitere Behandlung mit einem Dienstprogramm oder einem selbst erstellten kleinen Programm, welches Gesamthäufigkeiten errechnet, nicht möglich. Die Verfasser setzten ein kleines FORTRAN-Programm ("TRANS.SPSS.2") ein, welches die Ausgabedatei von "TRANS.SPSS"

[3] Anmerkung: Den 2...n Ausgabefiles müssen auf Systemebene im Parameter OUTPUTA = "Dateiname" unterschiedliche Dateinamen zugewiesen werden; vgl. Anhang 2, Aufruf der Enterdatei

als Eingabedatei verwendet um jeweils zwei korrespondierende Zeilen zusammenzufassen und Gesamthäufigkeiten zu berechnen.

Grundsätzlich ist zwar die manuelle Dateneingabe in GLIM vorgesehen, dies ist jedoch nur bei kleinen Kontingenztafeln vorteilhaft.

3. BEARBEITUNG DER MEHRDIMENSIONALEN KONTINGENZTAFEL MIT SPEZIELLEN FORTRAN-PROGRAMMEN

Die vorstehend beschriebene Ergebnisdatei wird mit Hilfe von zwei FORTRAN-Programmen bearbeitet. Beide FORTRAN-Programme "TRANS.SPSS", welches SPSS-Ergebnisdateien zur Analyse von loglinearen Modellen optimiert und "TRANS.SPSS.2", welches Eingabedateien für binomiale Logit-Modelle erstellt, sind interaktiv und selbstdokumentierend, d.h. dem Benutzer wird während des Programmablaufs vom Programm mitgeteilt, was das Programm an Eingaben erwartet (die Programme sind in Anhang 4 und 5 abgedruckt).

3.1 PROGRAMM "TRANS.SPSS"

Nach dem anlagespezifischen Programmaufruf, (Benutzer von Siemens-Anlagen die unter BS2000 arbeiten und den Source-Code des Programms unter ihrer Benutzer-ID (Kennung) abgelegt haben, können mit: EXEC TRANS.SPSS das Programm aufrufen) wird der Benutzer aufgefordert, die Eingabedatei (in Hochkommata eingeschlossen), dann die Ausgabedatei anzugeben. Anschließend fragt das Programm nach der Anzahl der Variablen; in diesem Beispiel sind es fünf Variablen (vgl. Anhang 4).
Das Programm "TRANS.SPSS" erzeugt eine neue (ISAM-)Datei, die in den ersten 8 Spalten wiederum den (ISAM-)Schlüssel enthält, im Format I10 die Variablenausprägungen und die Zellenhäufigkeiten. Das Programm tauscht alle Spalten der "SPSS-Ausgabedatei" und entfernt die SPSS-intern erzeugten Spalten, die lediglich "1" enthalten.
Die Ausgabedatei enthält hier die Beispiels-Variablen im Format I10 in folgender Reihenfolge: UNFTYP - 2 Ausprägungen; ALTBET - 3 Ausprägungen; GESCHL - 2 Ausprägungen; UNFALL - 2 Ausprägungen; STRKLASS - 3 Ausprägungen und die Zellenhäufigkeiten (vgl. Anhang 4).

3.2 PROGRAMM "TRANS.SPSS.2"

Das Programm "TRANS.SPSS.2" erzeugt Dateien, wie sie bei binomialen Logit-Modellen zur Modellierung von Anteilswerten für GLIM benötigt werden.

"TRANS.SPSS.2" formt die von "TRANS.SPSS" bereitgestellte Ausgabedatei um, indem jeweils aus zwei zusammengehörenden Zeilen Summen (Gesamthäufigkeiten) berechnet werden. Die Zellenhäufigkeiten werden nach den Codes der ersten Variablen - mit den Ausprägungen 1 und 2 -, die immer dichotom sein muß (in unserem Beispiel UNFTYP) unter Beibehaltung der Reihenfolge der Ausprägungen der übrigen Variablen addiert und in einer weiteren Spalte mit dem Format I10 der Datei angefügt; die Zeilen zur Ausprägung 2 werden gelöscht, die erste Variable entfällt. Die Ausgabedatei enthält danach nur noch halb so viele Zeilen wie die Eingabedatei (vgl. Anhang 5).
Der Aufruf von "TRANS.SPSS." ist auch hier anlagespezifisch (für Benutzer von Siemens-Anlagen gilt das oben ausgeführte; der Aufruf erfolgt mit: EXEC TRANS.SPSS.2").
Wiederum wird der Benutzer aufgefordert, die Eingabedatei in Hochkommata eingeschlossen anzugeben (dies ist die Ausgabedatei von TRANS.SPSS). Danach ist die Ausgabedatei anzugeben.
Anschließend muß die Anzahl der Spalten der Eingabedatei angegeben werden; in unserem Beispiel sind es sechs, fünf für die Variablen und eine für die Zellenhäufigkeiten.

4. ÜBERGABE DER MEHRDIMENSIONALEN KONTINGENZTAFEL AN GLIM

GLIM bietet mit dem Befehl DINPUT [Kanalnummer] die Möglichkeit Eingaben von einer externen Datei nach GLIM einzulesen. Aus diesem Grunde muß in einem "FILE-Kommando", das vor dem Aufruf von GLIM erfolgen muß, der externen Datei eine Kanalnummer zugewiesen werden. Die Befehle für die Zuweisung einer Kanalnummer sind Anlage- bzw. Betriebssystem spezifisch; (für Benutzer von Siemens-Anlagen, die unter BS2000 arbeiten, lautet das entsprechende "File-Kommando": FILE Dateiname, LINK = DSET70 (70 ist die Kanalnummer)).
Dateiname kann z.B. der Name einer Ausgabedatei von "TRANS.SPSS." sein.

5. SCHÄTZUNG EINES LOGLINEAREN MODELLS

5.1 VORBEREITUNG

Das 1. Beispiel zeigt, wie die Anzahl der Fahrunfälle durch die unabhängigen Variablen Alter des Beteiligten (A), Geschlecht (G), Alkoholeinfluß (AL) und Straßenzustand (SK) erklärt werden kann.

Die Analysen können sowohl im Batch-Betrieb als auch interaktiv durchgeführt werden. Beim interaktiven Arbeiten auf Großanlagen muß der Benutzer darauf achten, daß der Dialog mitprotokolliert wird. Erfolgt keine automatische Protokollierung kann der Benutzer durch entsprechende Befehle, die anlageabhängig sind, eine Protokollierung veranlassen.

Benutzer von Siemens-Anlagen müssen, wenn sie mit GLIM unter BS2000 interaktiv arbeiten, noch bevor GLIM aufgerufen wird, dem System mitteilen, daß der Dialog in einer Datei protokolliert werden soll:

```
OPTION MSG = FHL                \ spezielle Siemens-
SYSFILE SYSLST = BEISPIEL.1     / Steuerbefehle
```

Sofern die Eingabedaten (mehrdimensionale Kontingenztafel) in einer mit "TRANS.SPSS" oder durch ähnliche Dienstprogramme erzeugten Datei stehen, muß der Eingabedatei noch eine Kanalnummer zugewiesen werden (vgl. Abschnitt 4).

Bei der PC-Version wird automatisch eine Protokolldatei mit dem Namen "GLIM.LOG" angelegt. Diese Datei wird jedoch bei einem weiteren GLIM-Aufruf wieder überschrieben; sollen die Ergebnisse in Dateiform über längere Zeit aufbewahrt werden, muß die Datei "GLIM.LOG" umbenannt werden.

Da im Beispiel 1 nur die Häufigkeit der Fahrunfälle betrachtet werden soll, werden bei der in Abschnitt 3.1 spezifizierten Ausgabedatei mittels Editor die erste Spalte und die Zeilen 37 bis 72 gelöscht.

5.2 DATEN-DEFINITION

Nach dem Dialogaufruf von GLIM erfolgt die Definition der Eingabedaten (des Beispiels 1):

```
$UNIT    36   $FAC  A 3  G 2  AL 2  SK 3
$DATA   A G AL SK R
$DINPUT  70   $LOOK 1 5 R $
```

Mit $UNIT wird dem System die Zahl der Eingabezeilen mitgeteilt; mit $FAC werden die unabhängigen Variablen und ihre Ausprägungen definiert; mit $DATA wird angegeben, in welcher Reihenfolge die unabhängigen und die abhängige Variable R (hier: Zahl der Fahrunfälle) in der Datei stehen; mit $DINPUT 70 wird die Kanalnummer definiert unter der die Daten abgelegt sind und mit $LOOK werden hier die ersten 5 Werte von R aufgelistet. $LOOK braucht nicht aufgerufen zu werden, die Anweisung dient dem Benutzer lediglich zu Kontrollzwecken.

5.3 DEFINITION DES MODELLTYPS

Die Komponenten der Modellschätzung sind in den hier behandelten Beispielen:

a) Eine beobachtete abhängige Variable Y, hier die gegebene Zellenhäufigkeit.
b) Ein lineares Modell, gebildet aus den unabhängigen (erklärenden) Variablen, mit denen der Vektor η geschätzt wird.
c) Eine Wahrscheinlichkeitsverteilung für jedes Element der Y-Variablen: jedem Wert der Y-Variablen ist ein Erwartungswert μ zugewiesen, der ebenfalls in einem Vektor μ zusammengefaßt wird.
d) Eine Link-Funktion,
   $$\eta = g(\mu)$$
   die den linearen Prädiktor η mit dem Erwartungswert μ verknüpft.

Zur Modelldefinition ist zunächst folgende GLIM-Anweisung zu formulieren:

```
$YVAR R $ERR P $LINK L
```

$YVAR definiert die abhängige Variable;
$ERR die Verteilung der Fehler $(y - \mu)$ (hier: Poisson-Verteilung);
$LINK definiert das Verhältnis zwischen η und μ (s. oben d).
Der Benutzer kann zwischen folgenden Dichte- und Link-Funktionen wählen:

$ERR	$LINK
NORMAL	IDENTITY
POISSON	LOG
BINOMIAL	LOGIT
GAMMA	RECIPROCAL
	COMPLEMENTARY LOG-LOG
	NUMBER EXPONENT
	PROBIT
	SQUARE ROOT

Voreinstellung ist der natürliche Parameter der Dichtefunktion (z.B. N I, P L,...; Abkürzungen s. Anhang 7).

5.4 MODELLENTWICKLUNG

Es empfiehlt sich, in einem ersten Schritt die unabhängigen Variablen, auch Haupteffekte genannt, einzeln und gemeinsam zu testen, um den Einfluß der einzelnen Haupteffekte beurteilen zu können.

```
$FIT   $DIS ME $
$FIT  A   $DIS ME $
$FIT  -A+G  $DIS ME $
$FIT  -G+AL  $DIS ME $
$FIT  -AL+SK  $DIS ME $
$FIT  +A+G+AL   $DIS ME $
```

Mittels $FIT wird ein Modell geschätzt, welches neben der Regressionskonstanten die angegebenen Effekte (z.B. "A") enthält. Durch $FIT können, wie o.g. angegeben, Modellformeln gesetzt oder verändert werden; durch die zweite $FIT-Anweisung wird hier der Haupteffekt A zusätzlich modelliert. In der folgenden $FIT-Anweisung ($FIT -A+G) wird A der Modellformel entzogen und der Haupteffekt G in die Modellformel aufgenommen.

```
(IN)      $FIT $DIS ME $
(OUT)              SCALED
(OUT)       CYCLE  DEVIANCE       DF
(OUT)           5  0.3128E+05     35
(OUT)
(OUT)     Y-VARIATE  R
(OUT)     ERROR POISSON  LINK LOG
(OUT)
(OUT)     LINEAR PREDICTOR
(OUT)     %GM
(OUT)
(OUT)              ESTIMATE       S.E.       PARAMETER
(OUT)           1     6.415    0.6743E-02   %GM
(OUT)         SCALE PARAMETER TAKEN AS       1.000
(OUT)
(IN)      $FIT A $DIS ME $
(OUT)              SCALED
(OUT)       CYCLE  DEVIANCE       DF
(OUT)           5  0.1830E+05     33
(OUT)
(OUT)     Y-VARIATE  R
(OUT)     ERROR POISSON  LINK LOG
(OUT)
(OUT)     LINEAR PREDICTOR
(OUT)     %GM A
(OUT)
(OUT)              ESTIMATE       S.E.       PARAMETER
(OUT)           1     6.689    0.1018E-01   %GM
(OUT)           2    0.1916    0.1376E-01   A(2)
(OUT)           3    -2.671    0.4003E-01   A(3)
(OUT)         SCALE PARAMETER TAKEN AS       1.000
(OUT)
(IN)      $FIT -A+G $DIS ME $
(OUT)              SCALED
(OUT)       CYCLE  DEVIANCE       DF
(OUT)           5  0.2382E+05     34
(OUT)
(OUT)     Y-VARIATE  R
(OUT)     ERROR POISSON  LINK LOG
(OUT)
(OUT)     LINEAR PREDICTOR
(OUT)     %GM G
(OUT)
(OUT)              ESTIMATE       S.E.       PARAMETER
(OUT)           1     6.863    0.7622E-02   %GM
(OUT)           2    -1.281    0.1635E-01   G(2)
(OUT)         SCALE PARAMETER TAKEN AS       1.000
```

```
(IN)      $FIT -G+AL $DIS ME $
(OUT)             SCALED
(OUT)      CYCLE  DEVIANCE      DF
(OUT)         5   0.2290E+05    34
(OUT)
(OUT)     Y-VARIATE  R
(OUT)     ERROR POISSON  LINK LOG
(OUT)
(OUT)     LINEAR PREDICTOR
(OUT)     %GM AL
(OUT)
(OUT)             ESTIMATE       S.E.       PARAMETER
(OUT)          1    6.883      0.7547E-02  %GM
(OUT)          2   -1.376      0.1681E-01  AL(2)
(OUT)        SCALE PARAMETER TAKEN AS      1.000
(OUT)
(IN)      $FIT -AL+SK $DIS ME $
(OUT)             SCALED
(OUT)      CYCLE  DEVIANCE      DF
(OUT)         5   0.3011E+05    33
(OUT)
(OUT)     Y-VARIATE  R
(OUT)     ERROR POISSON  LINK LOG
(OUT)
(OUT)     LINEAR PREDICTOR
(OUT)     %GM SK
(OUT)
(OUT)             ESTIMATE       S.E.       PARAMETER
(OUT)          1    6.685      0.1020E-01  %GM
(OUT)          2 -0.5582       0.1691E-01  SK(2)
(OUT)          3 -0.3325       0.1579E-01  SK(3)
(OUT)        SCALE PARAMETER TAKEN AS      1.000
(OUT)
(IN)      $FIT +A+G+AL $DIS ME $
(OUT)             SCALED
(OUT)      CYCLE  DEVIANCE      DF
(OUT)         4    1284.        29
(OUT)
(OUT)     Y-VARIATE  R
(OUT)     ERROR POISSON  LINK LOG
(OUT)
(OUT)     LINEAR PREDICTOR
(OUT)     %GM SK A G AL
(OUT)
(OUT)             ESTIMATE       S.E.       PARAMETER
(OUT)          1    7.876      0.1365E-01  %GM
(OUT)          2 -0.5582       0.1691E-01  SK(2)
(OUT)          3 -0.3325       0.1579E-01  SK(3)
(OUT)          4  0.1916       0.1376E-01  A(2)
(OUT)          5  -2.671       0.4004E-01  A(3)
(OUT)          6  -1.281       0.1635E-01  G(2)
(OUT)          7  -1.376       0.1681E-01  AL(2)
(OUT)        SCALE PARAMETER TAKEN AS      1.000
```

$DIS zeigt die Ergebnisse des zuletzt vorgegebenen Fits: mit M werden alle Modell-Spezifikationen und mit E die Parameterschätzwerte und ihre Standardfehler abgerufen. Die "SCALED DEVIANCE" des Haupteffekt-Modells (1284 bei "DF" = Freiheitsgraden von 29; vgl. S. 21) zeigt, daß die Modellanpassung noch nicht hinreichend gut ist. Die "SCALED DEVIANCE" ist ein Maß für die Abweichungen zwischen den Schätzungen des gerechneten Modells und den empirischen Daten (= Schätzwerten des saturierten Modells); sie ist eine Funktion der Likelihoodfunktionen des gerechneten und des saturierten Modells. "CYCLE" gibt an, nach wieviel Iterationsschritten die Berechnung abgebrochen wurde.

Als nächstes wird ein Modell geschätzt, in welches noch alle Interaktionseffekte erster Ordnung aufgenommen werden. Die Modellformel lautet:

```
(IN)      $FIT +SK.A+SK.G+SK.AL+A.G+A.AL+G.AL $DIS ME $
(OUT)              SCALED
(OUT)       CYCLE  DEVIANCE       DF
(OUT)           3    24.35         16
(OUT)
(OUT)     Y-VARIATE  R
(OUT)     ERROR POISSON  LINK LOG
(OUT)
(OUT)     LINEAR PREDICTOR
(OUT)     %GM SK A G AL SK.A SK.G A.G SK.AL A.AL G.AL
(OUT)
(OUT)              ESTIMATE       S.E.       PARAMETER
(OUT)          1    7.727       0.1935E-01  %GM
(OUT)          2 -0.4519        0.2929E-01  SK(2)
(OUT)          3 -0.4117        0.2846E-01  SK(3)
(OUT)          4  0.3650        0.2379E-01  A(2)
(OUT)          5  -2.210        0.6257E-01  A(3)
(OUT)          6 -0.8553        0.3124E-01  G(2)
(OUT)          7 -0.9926        0.3257E-01  AL(2)
(OUT)          8 -0.2155        0.3465E-01  SK(2).A(2)
(OUT)          9  0.1541E-02    0.9428E-01  SK(2).A(3)
(OUT)         10  0.1115        0.3248E-01  SK(3).A(2)
(OUT)         11 -0.3048        0.1018      SK(3).A(3)
(OUT)         12  0.2486E-01    0.4132E-01  SK(2).G(2)
(OUT)         13 -0.6055E-01    0.3947E-01  SK(3).G(2)
(OUT)         14 -0.3535        0.3383E-01  A(2).G(2)
(OUT)         15 -0.9414        0.1169      A(3).G(2)
(OUT)         16 -0.1285E-01    0.4391E-01  SK(2).AL(2)
(OUT)         17  0.1880        0.3962E-01  SK(3).AL(2)
(OUT)         18 -0.3666        0.3471E-01  A(2).AL(2)
(OUT)         19  -1.425        0.1483      A(3).AL(2)
(OUT)         20  -1.491        0.6018E-01  G(2).AL(2)
(OUT)         SCALE PARAMETER TAKEN AS      1.000
```

Die Modellanpassung ist zwar mit einer "SCALED DEVIANCE" von 24.35 bei 16 Freiheitsgraden hinreichend gut, es werden jedoch trotzdem noch einige Interaktionseffekte zweiter Ordnung (Dreier-Interaktionen) geprüft. Diese Interaktionseffekte zweiter Ordnung werden nacheinander in die Modellformel aufgenommen, um die jeweils auftretenden Verbesserungen des Modells beurteilen zu können. Die entsprechenden Anweisungen lauten:

```
$FIT +SK.A.AL $DIS ME$
$FIT +A.G.AL  $DIS ME$
```

```
(IN)      $FIT +SK.A.AL $DIS ME $
(OUT)              SCALED
(OUT)      CYCLE  DEVIANCE      DF
(OUT)          3    10.46        12
(OUT)
(OUT)     Y-VARIATE  R
(OUT)     ERROR POISSON  LINK LOG
(OUT)
(OUT)     LINEAR PREDICTOR
(OUT)     %GM SK A G AL SK.A SK.G A.G SK.AL A.AL G.AL SK.A.AL
(OUT)
(OUT)              ESTIMATE       S.E.       PARAMETER
(OUT)          1    7.711       0.2004E-01  %GM
(OUT)          2 -0.4318        0.3082E-01  SK(2)
(OUT)          3 -0.3770        0.3020E-01  SK(3)
(OUT)          4  0.3925        0.2512E-01  A(2)
(OUT)          5  -2.205        0.6424E-01  A(3)
(OUT)          6 -0.8539        0.3132E-01  G(2)
(OUT)          7 -0.9255        0.3781E-01  AL(2)
(OUT)          8 -0.2526        0.3866E-01  SK(2).A(2)
(OUT)          9  0.5439E-02    0.9818E-01  SK(2).A(3)
(OUT)         10  0.5215E-01    0.3671E-01  SK(3).A(2)
(OUT)         11 -0.3250        0.1067      SK(3).A(3)
(OUT)         12  0.2334E-01    0.4134E-01  SK(2).G(2)
(OUT)         13 -0.6334E-01    0.3948E-01  SK(3).G(2)
(OUT)         14 -0.3538        0.3383E-01  A(2).G(2)
(OUT)         15 -0.9418        0.1169      A(3).G(2)
(OUT)         16 -0.1001        0.6045E-01  SK(2).AL(2)
(OUT)         17  0.4609E-01    0.5763E-01  SK(3).AL(2)
(OUT)         18 -0.5052        0.5300E-01  A(2).AL(2)
(OUT)         19  -1.359        0.2085      A(3).AL(2)
(OUT)         20  -1.491        0.6018E-01  G(2).AL(2)
(OUT)         21  0.1891        0.8719E-01  SK(2).A(2).AL(2)
(OUT)         22 -0.2475        0.3661      SK(2).A(3).AL(2)
(OUT)         23  0.2755        0.7871E-01  SK(3).A(2).AL(2)
(OUT)         24 -0.3507E-01    0.3593      SK(3).A(3).AL(2)
(OUT)         SCALE PARAMETER TAKEN AS      1.000
```

```
(IN)      $FIT +A.G.AL $DIS ME $
(OUT)              SCALED
(OUT)       CYCLE  DEVIANCE      DF
(OUT)           3     8.268      10
(OUT)
(OUT)     Y-VARIATE  R
(OUT)     ERROR POISSON  LINK LOG
(OUT)
(OUT)     LINEAR PREDICTOR
(OUT)     %GM SK A G AL SK.A SK.G A.G SK.AL A.AL G.AL SK.A.AL A.G.AL
(OUT)
(OUT)              ESTIMATE       S.E.       PARAMETER
(OUT)          1    7.711       0.2011E-01  %GM
(OUT)          2 -0.4319        0.3082E-01  SK(2)
(OUT)          3 -0.3771        0.3020E-01  SK(3)
(OUT)          4  0.3919        0.2527E-01  A(2)
(OUT)          5  -2.201        0.6426E-01  A(3)
(OUT)          6 -0.8547        0.3182E-01  G(2)
(OUT)          7 -0.9261        0.3825E-01  AL(2)
(OUT)          8 -0.2526        0.3866E-01  SK(2).A(2)
(OUT)          9  0.5573E-02    0.9819E-01  SK(2).A(3)
(OUT)         10  0.5220E-01    0.3670E-01  SK(3).A(2)
(OUT)         11 -0.3251        0.1068      SK(3).A(3)
(OUT)         12  0.2367E-01    0.4134E-01  SK(2).G(2)
(OUT)         13 -0.6305E-01    0.3948E-01  SK(3).G(2)
(OUT)         14 -0.3514        0.3533E-01  A(2).G(2)
(OUT)         15 -0.9706        0.1196      A(3).G(2)
(OUT)         16 -0.1001        0.6045E-01  SK(2).AL(2)
(OUT)         17  0.4618E-01    0.5762E-01  SK(3).AL(2)
(OUT)         18 -0.5026        0.5411E-01  A(2).AL(2)
(OUT)         19  -1.408        0.2130      A(3).AL(2)
(OUT)         20  -1.485        0.7933E-01  G(2).AL(2)
(OUT)         21  0.1892        0.8719E-01  SK(2).A(2).AL(2)
(OUT)         22 -0.2486        0.3663      SK(2).A(3).AL(2)
(OUT)         23  0.2753        0.7871E-01  SK(3).A(2).AL(2)
(OUT)         24 -0.3227E-01    0.3595      SK(3).A(3).AL(2)
(OUT)         25 -0.3222E-01    0.1228      A(2).G(2).AL(2)
(OUT)         26  0.8567        0.5394      A(3).G(2).AL(2)
(OUT)         SCALE PARAMETER TAKEN AS      1.000
```

Bei Modellen mit wenigen Variablen (Haupteffekten) ist es möglich, alle Dreier-Interaktionen zu testen. Ob dies aber eine sinnvolle Vorgehensweise ist, muß der Benutzer selber entscheiden; es sei nur angemerkt, daß der Aufwand bei Modellen mit 5 und mehr Variablen erheblich ist. Aus diesem Grunde sollten Interaktionseffekte

höherer Ordnung entweder aufgrund fachlicher Überlegungen oder (probehalber) in Abhängigkeit von den Ausprägungen niedrigerer Interaktionseffekte in die Modellierung einbezogen werden.

Die Interaktion A.G.AL ist in beiden Teilinteraktionen nicht signifikant von 0 verschieden, d.h. sie liefert keinen zusätzlichen Erklärungswert (vgl. S. 21, 2. Absatz). Sie kann somit wieder aus dem Modell ausgeschlossen werden.

```
(IN)      $FIT -A.G.AL $DIS ME $
(OUT)              SCALED
(OUT)      CYCLE  DEVIANCE      DF
(OUT)          3     10.46      12
(OUT)
(OUT)     Y-VARIATE  R
(OUT)     ERROR POISSON  LINK LOG
(OUT)
(OUT)     LINEAR PREDICTOR
(OUT)     %GM SK A G AL SK.A SK.G A.G SK.AL A.AL G.AL SK.A.AL
(OUT)
(OUT)              ESTIMATE       S.E.       PARAMETER
(OUT)          1    7.711       0.2004E-01  %GM
(OUT)          2 -0.4318        0.3082E-01  SK(2)
(OUT)          3 -0.3770        0.3020E-01  SK(3)
(OUT)          4  0.3925        0.2512E-01  A(2)
(OUT)          5  -2.205        0.6424E-01  A(3)
(OUT)          6 -0.8539        0.3132E-01  G(2)
(OUT)          7 -0.9255        0.3781E-01  AL(2)
(OUT)          8 -0.2526        0.3866E-01  SK(2).A(2)
(OUT)          9  0.5439E-02    0.9818E-01  SK(2).A(3)
(OUT)         10  0.5215E-01    0.3671E-01  SK(3).A(2)
(OUT)         11 -0.3250        0.1067      SK(3).A(3)
(OUT)         12  0.2334E-01    0.4134E-01  SK(2).G(2)
(OUT)         13 -0.6334E-01    0.3948E-01  SK(3).G(2)
(OUT)         14 -0.3538        0.3383E-01  A(2).G(2)
(OUT)         15 -0.9418        0.1169      A(3).G(2)
(OUT)         16 -0.1001        0.6045E-01  SK(2).AL(2)
(OUT)         17  0.4609E-01    0.5763E-01  SK(3).AL(2)
(OUT)         18 -0.5052        0.5300E-01  A(2).AL(2)
(OUT)         19  -1.359        0.2085      A(3).AL(2)
(OUT)         20  -1.491        0.6018E-01  G(2).AL(2)
(OUT)         21  0.1891        0.8719E-01  SK(2).A(2).AL(2)
(OUT)         22 -0.2475        0.3661      SK(2).A(3).AL(2)
(OUT)         23  0.2755        0.7871E-01  SK(3).A(2).AL(2)
(OUT)         24 -0.3507E-01    0.3593      SK(3).A(3).AL(2)
(OUT)         SCALE PARAMETER TAKEN AS      1.000
```

5.5 KONSTRUKTION DES OPTIMALEN MODELLS

Die Zweier-Interaktion SK.G ist in beiden Teilinteraktionen nicht signifikant von 0 verschieden, sie sollte deshalb im optimalen Modell nicht mehr enthalten sein.

```
(IN)      $FIT -SK.G $DIS ME $
(OUT)              SCALED
(OUT)      CYCLE  DEVIANCE      DF
(OUT)          3    14.68       14
(OUT)
(OUT)     Y-VARIATE  R
(OUT)     ERROR POISSON  LINK LOG
(OUT)
(OUT)     LINEAR PREDICTOR
(OUT)     %GM SK A G AL SK.A A.G SK.AL A.AL G.AL SK.A.AL
(OUT)
(OUT)             ESTIMATE       S.E.       PARAMETER
(OUT)          1   7.715       0.1919E-01  %GM
(OUT)          2 -0.4248       0.2818E-01  SK(2)
(OUT)          3 -0.3955       0.2794E-01  SK(3)
(OUT)          4  0.3925       0.2505E-01  A(2)
(OUT)          5  -2.208       0.6414E-01  A(3)
(OUT)          6 -0.8653       0.2488E-01  G(2)
(OUT)          7 -0.9276       0.3759E-01  AL(2)
(OUT)          8 -0.2542       0.3855E-01  SK(2).A(2)
(OUT)          9  0.1769E-02   0.9796E-01  SK(2).A(3)
(OUT)         10  0.5643E-01   0.3661E-01  SK(3).A(2)
(OUT)         11 -0.3153       0.1066      SK(3).A(3)
(OUT)         12 -0.3572       0.3377E-01  A(2).G(2)
(OUT)         13 -0.9374       0.1168      A(3).G(2)
(OUT)         14 -0.1051       0.5980E-01  SK(2).AL(2)
(OUT)         15  0.5920E-01   0.5705E-01  SK(3).AL(2)
(OUT)         16 -0.5054       0.5299E-01  A(2).AL(2)
(OUT)         17  -1.358       0.2085      A(3).AL(2)
(OUT)         18  -1.494       0.6017E-01  G(2).AL(2)
(OUT)         19  0.1902       0.8717E-01  SK(2).A(2).AL(2)
(OUT)         20 -0.2451       0.3661      SK(2).A(3).AL(2)
(OUT)         21  0.2727       0.7869E-01  SK(3).A(2).AL(2)
(OUT)         22 -0.4162E-01   0.3593      SK(3).A(3).AL(2)
(OUT)         SCALE PARAMETER TAKEN AS     1.000
```

Mit dem Modell

SK+A+G+AL+SK.A+A.G+SK.AL+A.AL+G.AL+SK.A.AL

kann die Suche nach weiteren signifikanten Einflußgrößen (Interaktionen) beendet werden.

Die "SCALED DEVIANCE", die die Abweichungen der auf der Basis der Parameter der Modellformel erwarteten von den beobachteten Zellenhäufigkeiten kennzeichnet, beträgt 14.68 bei "DF" 14. Da die Devianzen mit den jeweils angegebenen Freiheitsgraden asymptotisch X^2 verteilt sind, stellen wir fest, daß dieses Modell eine gute Anpassung hat.

Im nächsten Schritt muß nun geprüft werden, ob alle im Modell enthaltenen Parameter (Effekte) signifikant von 0 verschieden sind, denn ein optimales Modell sollte nur signifikante Effekte enthalten. Als Faustregel kann die Forderung gelten: "ESTIMATE"/"S.E.">2 ($\alpha \leq 0{,}05$). Der Benutzer hat 2 Möglichkeiten der Prüfung:

1. durch Augenschein bzw. Benutzung eines Taschenrechners
2. durch die Formulierung eines cal-Statements im GLIM-Programm z.B. $CAL 0.3925/0.02505 $

Es läßt sich leicht feststellen, daß alle Haupteffekte signifikant von 0 verschieden sind. Bei den Interaktionen erster Ordnung (Zweier-Interaktionen) sind die Interaktionen A.G, A.AL, G.AL in allen Teilinteraktionen signifikant von 0 verschieden ; die Interaktion SK.A ist in den zwei Teilinteraktionen SK(2).A(3), SK(3).A(3) und die Interaktion SK.AL nur in der Teilinteraktion SK(2).AL(2) signifikant von 0 verschieden.
Die Interaktion zweiter Ordnung (Dreier-Interaktion) SK.A.AL weist nur in den Teilinteraktionen SK(2).A(2).AL(2) und SK(3).A(2).AL(2) signifikant von 0 verschiedene Effekte auf.

Da nicht signifikante Teilinteraktionen aus dem Modell verschwinden sollen, werden im nächsten Schritt signifikante Teilinteraktionen durch entsprechende cal-Statements einzeln definiert. Der Benutzer muß zur Identifikation der Parameter sinnvolle Abkürzungen finden, die nur aus max. 4 Zeichen bestehen dürfen.

```
(IN)      $CAL SK2=%IF(%EQ(SK,2),1,0)
(IN)      $CAL SK3=%IF(%EQ(SK,3),1,0)
(IN)      $CAL A2=%IF(%EQ(A,2),1,0)
(IN)      $CAL A3=%IF(%EQ(A,3),1,0)
(IN)      $CAL AL2=%IF(%EQ(AL,2),1,0)
(IN)      $CAL S2A2=SK2*A2
(IN)      $CAL S3A3=SK3*A3
(IN)      $CAL S2AA=SK2*A2*AL2
(IN)      $CAL S3AA=SK3*A2*AL2
(IN)      $CAL S2AL=SK2*AL2
```

Nach der Kalkulation dieser neuen Parameter (Effekte), müssen diese noch in die Modellformel eingebracht werden.

$FIT -SK.A-SK.AL-SK.A.AL+S2A2+S3A3+S2AA+S3AA $DIS ME $

Sofern nicht durch den Ausschluß nichtsignifikanter Teilinteraktionen Umlagerungen erfolgen, die die Signifikanz anderer Teilinteraktionen beeinträchtigen, ist nach diesem Arbeitsschritt bereits das optimale Modell gefunden.

optimales Modell:

```
(IN)      $FIT -SK.AL-SK.A-SK.A.AL+S2AL+S2A2+S3A3+S2AA+S3AA
(IN)      $DIS ME $
(OUT)              SCALED
(OUT)       CYCLE  DEVIANCE       DF
(OUT)           3    17.85        19
(OUT)
(OUT)     Y-VARIATE  R
(OUT)     ERROR POISSON  LINK LOG

(OUT)     LINEAR PREDICTOR
(OUT)     %GM SK A G AL S2A2 S3A3 S2AA S3AA S2AL A.G A.AL G.AL
(OUT)
(OUT)              ESTIMATE       S.E.       PARAMETER
(OUT)          1    7.700       0.1687E-01  %GM
(OUT)          2 -0.4111        0.2573E-01  SK(2)
(OUT)          3 -0.3600        0.1692E-01  SK(3)
(OUT)          4  0.4154        0.2001E-01  A(2)
(OUT)          5  -2.200        0.5046E-01  A(3)
(OUT)          6 -0.8653        0.2488E-01  G(2)
(OUT)          7 -0.9022        0.2913E-01  AL(2)
(OUT)          8 -0.2766        0.3487E-01  S2A2
(OUT)          9 -0.3387        0.9345E-01  S3A3
(OUT)         10  0.2277        0.8306E-01  S2AA
(OUT)         11  0.3529        0.5162E-01  S3AA
(OUT)         12 -0.1339        0.5404E-01  S2AL
(OUT)         13 -0.3572        0.3377E-01  A(2).G(2)
(OUT)         14 -0.9374        0.1168      A(3).G(2)
(OUT)         15 -0.5394        0.4686E-01  A(2).AL(2)
(OUT)         16  -1.434        0.1483      A(3).AL(2)
(OUT)         17  -1.494        0.6017E-01  G(2).AL(2)
(OUT)         SCALE PARAMETER TAKEN AS      1.000
```

Die "SCALED DEVIANCE" des optimalen Modells beträgt 17.85 bei "DF" 19. Damit ist das Modell gut angepaßt.
Das optimale Modell enthält 17 Parameter. Die Effekte aller Parameter sind signifikant von 0 verschieden.

Für das optimale Modell sollen die Residuen ausgegeben werden. Dies erfolgt mit der $DIS-Anweisung:

```
$DIS R $
```

R = Residuen.

```
(IN)      $DIS R $
(OUT)       UNIT  OBSERVED     FITTED     RESIDUAL
(OUT)          1      2245       2209.       0.7629
(OUT)          2      1458       1465.      -0.1702
(OUT)          3      1513       1541.      -0.7203
(OUT)          4       892       896.2      -0.1395
(OUT)          5       520       519.6       0.1566E-01
(OUT)          6       628       625.2       0.1101
(OUT)          7       939       929.9       0.2994
(OUT)          8       624       616.4       0.3044
(OUT)          9       631       648.8      -0.6970
(OUT)         10        78       84.71      -0.7292
(OUT)         11        51       49.12       0.2683
(OUT)         12        65       59.10       0.7672
(OUT)         13      3290       3347.      -0.9812
(OUT)         14      1655       1683.      -0.6706
(OUT)         15      2416       2335.        1.677
(OUT)         16       793       791.6       0.4819E-01
(OUT)         17       434       437.1      -0.1479
(OUT)         18       791       786.0       0.1778
(OUT)         19      1005       985.6       0.6175
(OUT)         20       523       495.5        1.236
(OUT)         21       644       687.6       -1.664
(OUT)         22        51       52.36      -0.1873
(OUT)         23        32       28.91       0.5752
(OUT)         24        47       51.98      -0.6912
(OUT)         25       243       244.8      -0.1154
(OUT)         26       168       162.3       0.4482
(OUT)         27       120       121.7      -0.1563
(OUT)         28        24       23.68       0.6633E-01
(OUT)         29        11       13.73      -0.7365
(OUT)         30        12       11.77       0.6617E-01
(OUT)         31        44       40.36       0.5733
(OUT)         32        20       26.75       -1.306
(OUT)         33        21       20.07       0.2083
(OUT)         34         2      0.8766        1.200
(OUT)         35         1      0.5083       0.6897
(OUT)         36         1      0.4359       0.8545
```

Die standardisierten Residuen sind grundsätzlich näherungsweise normalverteilt. Abweichungen, die größer als (+/-. 2) sind, sollten daher bei einem gut angepaßten Modell mit einer Wahrscheinlichkeit von höchstens 5% auftreten.
In unserem Beispiel liegen alle standardisierten Residuen im Intervall (+2, -2). Dies zeigt ebenfalls, wie gut das Modell angepaßt ist.

Neben den Residuen soll für das optimale Modell auch die (Co-)VarianzMatrix und die Matrix der Standard-Fehler ausgegeben werden; dies erfolgt mit der $DIS-Anweisung

```
$DIS VS $
V = (Co-)Varianz-Matrix
S = Standardfehler der Differenzen (S.E. of Differences)
```

```
(IN)      $DIS VS $
(CO)VARIANCE MATRIX
 1  2.8466E-04
 2 -2.1830E-04  6.6226E-04
 3 -1.1532E-04  1.0935E-04  2.8630E-04
 4 -2.3553E-04  1.7336E-04 -2.3337E-06  4.0037E-04
 5 -1.9854E-04 -4.1246E-05  7.1755E-05  1.6750E-04  2.5464E-03
 6 -1.8337E-04  5.4212E-20 -8.5873E-20  1.7751E-04  1.8216E-04  6.1900E-04
 7 -2.3104E-04  1.7314E-04 -1.7119E-06  2.2558E-04  1.6588E-04  1.6193E-04
 8  1.7091E-04 -6.1732E-04  8.3099E-06 -3.1028E-04  7.0735E-05 -2.7397E-19
 9  3.3775E-05  1.3527E-04 -2.4281E-04  6.6009E-05 -2.0296E-03 -6.0684E-21
10 -1.2400E-04  5.7153E-04 -1.2352E-04  3.1072E-04 -8.4142E-05  5.4835E-19
11  1.1532E-04 -1.0935E-04 -2.8630E-04  2.3337E-06 -7.1755E-05  6.3809E-19
12  1.7139E-04 -6.1647E-04  5.8679E-06 -1.7380E-04  5.4653E-05  7.9152E-20
13  1.7574E-04 -2.4799E-19  7.9633E-20 -3.0022E-04 -1.7535E-04 -5.9326E-04
14  1.8085E-04 -1.1454E-20  1.3223E-20 -1.7630E-04 -2.0275E-03 -6.1049E-04
15  1.7557E-04 -1.2821E-04  1.1937E-04 -3.7953E-04 -1.2918E-04 -1.3467E-04
16  1.7370E-04  1.5060E-05  2.7052E-06 -1.7083E-04 -1.9794E-03 -1.4778E-04
17  1.0983E-04  3.3815E-19  5.0637E-20 -5.2657E-05 -9.8067E-05 -3.7076E-04
        1           2           3           4           5           6

 7  8.4828E-04
 8 -1.7385E-04  1.2158E-03
 9  4.8424E-05 -2.3505E-04  8.7334E-03
10  7.6047E-04 -1.2174E-03  2.6576E-04  6.8995E-03
11  1.7119E-06 -8.3099E-06  2.4281E-04  1.3084E-03  2.6644E-03
12 -7.5977E-04  6.1888E-04 -1.6598E-04 -2.9229E-03 -5.8679E-06  2.9205E-03
13 -1.3482E-04  5.4093E-19 -1.2953E-19 -8.5670E-19 -5.8175E-19  3.1262E-19
14 -1.5297E-04  1.3456E-19  2.3744E-18 -5.7498E-19 -5.3369E-19  3.3747E-21
15 -8.2027E-04  3.1323E-04 -1.4821E-04 -2.1320E-03 -1.3042E-03  7.6218E-04
16 -6.1691E-04 -1.3949E-05 -7.6518E-05  7.0027E-05 -2.7052E-06 -7.1139E-05
17 -3.9044E-04 -2.3602E-19  1.2993E-19  8.8928E-19 -5.6788E-19 -1.5056E-18
        7           8           9          10          11          12

13  1.1404E-03
14  5.9052E-04  1.3652E-02
15  2.1794E-04  1.3181E-04  2.1962E-03
16  1.3026E-04  1.5235E-03  5.9947E-04  2.1984E-02
17  1.1945E-04  2.8769E-04  1.2430E-04  2.5233E-04  3.6199E-03
       13          14          15          16          17
   SCALE PARAMETER TAKEN AS    1.000
```

S.E. OF DIFFERENCES

1	0.0000					
2	3.7196E-02	0.0000				
3	2.8313E-02	2.7016E-02	0.0000			
4	3.4001E-02	2.6756E-02	2.6293E-02	0.0000		
5	5.6817E-02	5.7369E-02	5.1857E-02	5.1105E-02	0.0000	
6	3.5643E-02	3.5795E-02	3.0088E-02	2.5775E-02	5.2925E-02	0.0000
7	3.9938E-02	3.4121E-02	3.3734E-02	2.8240E-02	5.5344E-02	3.3815E-02
8	3.4039E-02	5.5792E-02	3.8542E-02	4.7295E-02	6.0173E-02	4.2835E-02
9	9.4607E-02	9.5525E-02	9.7495E-02	9.4877E-02	0.1238	9.6708E-02
10	8.6210E-02	8.0117E-02	8.6214E-02	8.1722E-02	9.8052E-02	8.6709E-02
11	5.2139E-02	5.9543E-02	5.9358E-02	5.5319E-02	7.3173E-02	5.7301E-02
12	5.3501E-02	6.9395E-02	5.6525E-02	6.0568E-02	7.3195E-02	5.9493E-02
13	3.2766E-02	4.2458E-02	3.7772E-02	4.6273E-02	6.3541E-02	5.4276E-02
14	0.1165	0.1196	0.1180	0.1200	0.1423	0.1244
15	4.6149E-02	5.5811E-02	4.7368E-02	5.7928E-02	7.0717E-02	5.5538E-02
16	0.1481	0.1504	0.1492	0.1508	0.1688	0.1513
17	6.0703E-02	6.5438E-02	6.2499E-02	6.4230E-02	7.9765E-02	7.0572E-02
	1	2	3	4	5	6
7	0.0000					
8	4.9110E-02	0.0000				
9	9.7390E-02	0.1020	0.0000			
10	7.8910E-02	0.1027	0.1228	0.0000		
11	5.9239E-02	6.2425E-02	0.1044	8.3350E-02	0.0000	
12	7.2721E-02	5.3838E-02	0.1094	0.1252	7.4811E-02	0.0000
13	4.7522E-02	4.8541E-02	9.9367E-02	8.9665E-02	6.1683E-02	6.3725E-02
14	0.1216	0.1219	0.1496	0.1434	0.1277	0.1287
15	6.8447E-02	5.2779E-02	0.1059	0.1155	8.6424E-02	5.9936E-02
16	0.1551	0.1524	0.1757	0.1695	0.1570	0.1583
17	7.2450E-02	6.9539E-02	0.1111	0.1025	7.9274E-02	8.0872E-02
	7	8	9	10	11	12
13	0.0000					
14	0.1166	0.0000				
15	5.3858E-02	0.1248	0.0000			
16	0.1512	0.1805	0.1516	0.0000		
17	6.7241E-02	0.1292	7.4615E-0	0.1584	0.0000	
	13	14	15	16	17	

SCALE PARAMETER TAKEN AS 1.000

Mit Hilfe der Standardfehler der Differenzen kann man testen, ob sich zwei Effekte voneinander unterscheiden.
Im Beispiel der Effekte A(2) und A(3) geschieht dies wie folgt:

Nullhypothese: H_0 : A(2) - A(3) = 0

Alternativhypothese: H_1 : A(2) - A(3) $\neq$ 0

Testgröße: T = (A(2) - A(3)) / S.E. (A(2),A(3))

Schwellenwert: $\alpha/2$ und $(1-\alpha/2)$ - Quantile der Standardnormalverteilung (für α=0,05; $u_{1-\alpha/2}$=1,96)

Da A(2) - A(3) = 0.4154 - (-2.200) = 2.6154 und der Standardfehler der Differenzen zwischen den Parametern 4 und 5 = 0.051105 beträgt, kann H_0 mit $\alpha < 0.01$ abgelehnt werden.
Die Ermittlung von A(2) - A(3) kann auch als GLIM-Anweisung formuliert werden:

```
$CAL 0.4154-(-2.200)/0.051105$
```

5.6 GRAPHISCHE DARSTELLUNG DER RESIDUEN DES OPTIMALEN MODELLS

Mit GLIM ist es möglich, die standardisierten Residuen graphisch darzustellen. Die graphische Darstellung der Residuen ermöglicht eine sehr gute Beurteilung der Güte der Anpassung, denn der Benutzer sieht auf einen Blick die Verteilung der standardisierten Residuen. Die standardisierten Residuen sind in etwa normalverteilt. Abweichungen, die größer (±2) sind, sind daher bei einem gut angepaßten Modell mit einer Wahrscheinlichkeit von höchstens 5% zu erwarten.
Da die mit $DIS R ausgegebenen standardisierten Residuen nicht weiterverarbeitet werden können, muß zur weiteren Verwendung der standardisierten Residuen ein Makro (d.h. ein kleines Unterprogramm) geschrieben werden:

```
$MAC RES
$CAL RE = (%YU - %FV)/%SQRT (%FV)
ENDMAC $
```

%YU = beobachtete Werte

%FV = geschätzte Werte

%SQRT = √

Die Anweisung $PLO RE A$ erzeugt ein Scattergramm. Die standardisierten Residuen RE werden auf der Y-Achse und A (3 Lebensaltersklassen) auf der X-Achse abgebildet. Wie das Scattergramm verdeutlicht ist die Streuung der standardisierten Residuen in der Altersklasse A(2) (Fahrer unter 25 Jahre) am größten, in der Altersklasse A(1) (Fahrer zwischen 25 und 60 Jahre) am geringsten. Alle standardisierten Residuen liegen im Intervall (+2,-2).

```
(IN)      $MAC RES $CAL RE=(%YV-%FV)/%SQRT(%FV)  $ENDMAC $
(IN)      $USE RES $PLO RE A $
(OUT)    2.00   *
(OUT)    1.75   *                          R
(OUT)    1.50   *
(OUT)    1.25   *                          R                          R
(OUT)    1.00   *
(OUT)   0.750   *   2                                                 2
(OUT)   0.500   *                          2                          2
(OUT)   0.250   *   3                      R                          R
(OUT)   0.000   *   2                      R                          3
(OUT) -0.250    *   2                      2                          R
(OUT) -0.500    *
(OUT) -0.750    *   3                      2                          R
(OUT)   -1.00   *                          R
(OUT)   -1.25   *                                                     R
(OUT)   -1.50   *
(OUT)   -1.75   *                          R
(OUT)   -2.00   *
.................*........*........*........*........*........*........
(OUT)          0.800     1.20     1.60     2.00     2.40     2.80
```

Auf der X-Achse können auch die geschätzten (%FV) oder die beobachteten (R) Werte abgetragen werden. Im ersten Diagramm sind die standardisierten Residuen gegen die (geschätzte) erwartete Zahl der Fahrunfälle, im zweiten gegen die Zahl beobachtete der Fahrunfälle geplottet. In beiden Diagrammen sind die standardisierten Residuen gleichmäßig um die Nullachse verteilt.

```
 (IN)     $PLO RE %FV $
 (OUT)      2.00     *
 (OUT)      1.75     *                              R
 (OUT)      1.50     *
 (OUT)      1.25     R      R
 (OUT)      1.00     *
 (OUT)     0.750     2R                           R
 (OUT)     0.500     RRR           R
 (OUT)     0.250     RR       R R R
 (OUT)     0.000     2  R   R R R
 (OUT)    -0.250     *RR  R      R       R
 (OUT)    -0.500     *
 (OUT)    -0.750     R2       R            R R
 (OUT)     -1.00     *                                                R
 (OUT)     -1.25     R
 (OUT)     -1.50     *
 (OUT)     -1.75     *         R
 (OUT)     -2.00     *
...........*.........*.........*.........*.........*.........*.......
 (OUT)             0.000       800.      0.160E+04 0.240E+04 0.320E+04
```

```
 (IN)     $PLO RE R $
 (OUT)      2.00     *
 (OUT)      1.75     *                               R
 (OUT)      1.50     *
 (OUT)      1.25     R      R
 (OUT)      1.00     *
 (OUT)     0.750     2R                           R
 (OUT)     0.500     RRR           R
 (OUT)     0.250     RR       R R R
 (OUT)     0.000     2  R   R R R
 (OUT)    -0.250     *2   R      R       R
 (OUT)    -0.500     *
 (OUT)    -0.750     R2       R            R R
 (OUT)     -1.00     *                                               R
 (OUT)     -1.25     R
 (OUT)     -1.50     *
 (OUT)     -1.75     *         R
 (OUT)     -2.00     *
...........*.........*.........*.........*.........*.........*.......
 (OUT)             0.000       800.      0.160E+04 0.240E+04 0.320E+04
```

Ebenso ist es möglich, die standardisierten Residuen zu sortieren und gegen die gleiche Anzahl sortierter Zufallszahlen der Standardnormalverteilung abzubilden. Folgende GLIM-Statements sind zu formulieren:

```
 (IN)     $USE RES
 (IN)     $CAL GRE=RE $SORT GRE $CAL NOR=%ND(%SR(0))
 (IN)     $SORT NOR $PLO GRE NOR $
 (OUT)      2.00   *
 (OUT)      1.75   *                                              G
 (OUT)      1.50   *
 (OUT)      1.25   *                                        G G
 (OUT)      1.00   *
 (OUT)     0.750   *                                   G GGG
 (OUT)     0.500   *                                  G3
 (OUT)     0.250   *                              G2G  G
 (OUT)     0.000   *                         G2 2G
 (OUT)    -0.250   *                    GG GGG
 (OUT)    -0.500   *
 (OUT)    -0.750   *            22  2
 (OUT)     -1.00   *           G
 (OUT)     -1.25   *   G
 (OUT)     -1.50   *
 (OUT)     -1.75   *G
 (OUT)     -2.00   *
.........*.........*.........*.........*.........*.........*.........*
 (OUT)             -2.00      -1.00      0.000      1.00       2.00
```

Der letzte Plot zeigt, daß die standardisierten Residuen auf einer 45°-Geraden liegen; das Modell ist damit den Daten gut angepaßt. Wenn keine weiteren Analysen durchgeführt werden sollen, wird GLIM mit $STOP verlassen.

5.7 INTERPRETATION DES OPTIMALEN MODELLS

Zur Analyse wurden in den Abschnitten 5.3 bis 5.5 Modelle unter Einschluß eines bzw. aller Haupteffekte sowie der Interaktionen 1. Ordnung (Zweier-Interaktionen) und unter sachlichen Kriterien ausgewählten Interaktionen 2. Ordnung (Dreier-Interaktionen) untersucht. Nach schrittweisem Ausschluß aller nicht signifikanten Parameter ergab sich ein optimales Modell mit 4 Haupteffekten, 8 Interaktionseffekten erster Ordnung und 2 Interaktionseffekten zweiter Ordnung.

Da die Haupteffekte Alter des Beteiligten und Straßenklasse mehr als zwei Kategorien (einschließlich der Basiskategorie) haben, enthält das optimale Modell 17 Parameter. Durch dieses Modell werden 99,9% der Devianz des Basismodells (vgl. Anhang 6.6) erklärt. Wir wollen dieses Modell als das optimale Modell bezeichnen, da es das Ziel einer jeden multivariaten Analyse erfüllt, einerseits den empirischen Befund so einfach wie möglich, d.h. mit möglichst wenig Parametern, andererseits aber auch so genau wie möglich darzustellen. Diese beiden Kriterien stehen prinzipiell in Konkurrenz miteinander und der Anwender muß einen Kompromiß zwischen beiden Forderungen suchen. Somit kann das Endresultat einer loglinearen Analyse auch kein absolut gesehen bestes Modell sein, sondern nur ein optimales, das diesen Namen hinsichtlich ganz bestimmter Kriterien besitzt. Hier war das Optimierungskriterium: Einschluß aller signifikanter Parameter bei einem Signifikanzniveau von $\alpha \leq 0{,}01$.

Positive Werte der "ESTIMATE" weisen auf eine im Vergleich zur Basiskategorie größere Besetzungszahl hin, während negative Werte eine geringere Anzahl an Fahrunfällen der Merkmalskategorie bezeichnen.

Als Basiskategorie der Variablen (bis auf A) wurde jeweils die Ausprägung mit der größten Besetzungshäufigkeit an Fahrunfällen gewählt. Daraus ergibt sich für die Regressionskonstante GM (= Geltung aller Basiskategorien) ein vergleichsweise hoher Wert. Entsprechend treten für die übrigen Haupteffekt-Parameter SK(2)... AL(2), ausgenommen A(2), negative Vorzeichen auf.

Die "ESTIMATE" der Parameter sind einzeln und in Kombination miteinander interpretierbar (vgl. Anhang 6.1). So wird aus den relativ hohen negativen "ESTIMATE" der Haupteffekte A(3), G(2), AL(2) deutlich, daß die Absolutzahl der Fahrunfälle von alten Fahrern, Frauen und bei Alkohol vergleichsweise niedrig ist. Wie die gleichfalls großen negativen "ESTIMATE" der Interaktionen G(2).AL(2), A(3).AL(2) und A(3).G(2) zeigen, treten bei Kombination der zuvor genannten Haupteffekte nochmals geringere Absolutzahlen auf.

Grundsätzlich ist es auch möglich, mittels loglinearer Modelle zu Aussagen über den Anteil der Fahrunfälle an allen Unfällen, d.h. über das bedingte Risiko für einen Fahrunfall, wenn sich ein Unfall ereignet, zu gelangen. Gegenüber der vorangegangenen Auswertung hätte man hierfür nicht nur die Anzahl der Fahrunfälle in den verschiedenen Merkmalskombinationen untersuchen müssen, sondern die Anzahl der Fahrunfälle im Verhältnis zu allen Unfällen (in Abhängigkeit von den Variablenkonstellationen) betrachten müssen. Dazu wäre eine weitere Variable "Unfallart" mit den Kategorien "Fahrunfall","kein Fahrunfall" in die Analyse zusätzlich einzubeziehen.

Für eine Fragestellung dieser Art ist es aber wesentlich günstiger, Modelle zu entwickeln, die nicht Unfallanzahlen sondern unmittelbar den Anteil der Fahrunfälle an allen Unfällen schätzen. Diese Analyse mittels Logit-Modellen soll im folgenden Abschnitt 6 durchgeführt werden.

6. SCHÄTZUNG EINES BINOMIALEN LOGIT-MODELLS

6.1 VORBEREITUNG

Bevor GLIM aufgerufen wird, sollten auch hier (vgl. Abschnitt 5.1) entsprechende Vorkehrungen zur Protokollierung des Dialogprozesses getroffen werden.
Sofern die Eingabe-Daten (mehrdimensionale Kontingenztafel), in einer mit "TRANS.SPSS.2" oder durch ähnliche Dienstprogramme erzeugten Datei (vgl. Abschnitt 3.2) abgelegt sind, muß dieser noch eine Kanalnummer zugewiesen werden (vgl. Abschnitt 4).

6.2 DATEN-DEFINITION

Nach dem Dialogaufruf von GLIM erfolgt die Definition der Eingabedaten (des Beispiels 2):

```
$UNIT 36 $FAC  A 3  G 2  AL 2  SK 3
$DATA A G AL SK R N
$DINPUT 70
$LOOK 1 5 A G AL SK R N $
```

$UNIT enthält die Zahl der Eingabezeilen. Mit $FAC erfolgt die Benennung der Variablen mit der Zahl ihrer Ausprägungen. Mit $DATA wird angegeben, in welcher Reihenfolge die unabhängigen Variablen, sowie die Variablen R (Zahl der Fahrunfälle) und N (Zahl der Unfälle insgesamt) in der Datei abgelegt sind. Mit $DINPUT 70 wird die Kanalnummer definiert, unter der die Daten abgelegt sind. Mit $LOOK werden hier zur Kontrolle die ersten 5 Werte von A, G, AL, SK, R, N, aufgelistet.

6.3 DEFINITION DES MODELLTYPS

Die Komponenten der Modellschätzung sind im Abschnitt 5.2 - a) bis d) - angegeben worden. Die Modelldefinition erfolgt für das Beispiel 2 mit folgender Anweisung:

```
$YVAR R $ERR B N $LINK G
```

$YVAR definiert die unabhängige Variable;
Mit $ERR B N wird die Verteilung der Fehler (y und μ) als binomialverteilt angegeben. N bezeichnet den Nenner, hier das Verhältnis zwischen n und μ. Mit $LINK G wird die Logit-Funktion bestimmt.

6.4 MODELLENTWICKLUNG

Wie in Beispiel 1 empfiehlt es sich auch hier, die unabhängigen Variablen, auch Haupteffekte genannt, einzeln und gemeinsam zu testen, um den Einfluß der einzelnen Haupteffekte beurteilen zu können.

```
$FIT       $ DIS ME $
$FIT  A    $ DIS ME $
$FIT -A+G  $ DIS ME $
$FIT -G+AL $ DIS ME $
$FIT -AL+SK $ DIS ME $
$FIT +A+G+AL $ DIS ME $
```

Mittels $FIT wird ein Modell geschätzt, welches neben der Regressionskonstanten die angegebenen Effekte (z.B. "A"; "G"...) enthält. Durch "+, -" in der $FIT-Anweisung werden die Haupteffekte einzeln in die Modellformel aufgenommen bzw. dieser wieder entzogen.
$DIS zeigt die Ergebnisse des jeweiligen Fits; mit M werden alle Modell-Spezifikationen und mit E die Parameterschätzwerte und ihre Standardfehler abgerufen.

```
(IN)      $FIT $DIS ME $
(OUT)              SCALED
(OUT)      CYCLE  DEVIANCE      DF
(OUT)          3    5497.       35
(OUT)
(OUT)     Y-VARIATE  R
(OUT)     ERROR BINOMIAL LINK LOGIT
(OUT)     BINOMIAL DENOMINATOR N
(OUT)
(OUT)     LINEAR PREDICTOR
(OUT)     %GM
(OUT)
(OUT)             ESTIMATE      S.E.       PARAMETER
(OUT)          1 -0.6375      0.8337E-02  %GM
(OUT)        SCALE PARAMETER TAKEN AS      1.000
(OUT)
(IN)      $FIT A $DIS ME $
(OUT)              SCALED
(OUT)      CYCLE  DEVIANCE      DF
(OUT)          3    3238.       33
(OUT)
(OUT)     Y-VARIATE  R
(OUT)     ERROR BINOMIAL LINK LOGIT
(OUT)     BINOMIAL DENOMINATOR N
(OUT)
(OUT)     LINEAR PREDICTOR
(OUT)     %GM A
(OUT)
(OUT)             ESTIMATE      S.E.       PARAMETER
(OUT)          1 -0.8785      0.1211E-01  %GM
(OUT)          2  0.6466      0.1733E-01  A(2)
(OUT)          3 -0.8236      0.4381E-01  A(3)
(OUT)        SCALE PARAMETER TAKEN AS      1.000
(OUT)
(IN)      $FIT -A+G $DIS ME $
(OUT)              SCALED
(OUT)      CYCLE  DEVIANCE      DF
(OUT)          3    5420.       34
(OUT)
(OUT)     Y-VARIATE  R
(OUT)     ERROR BINOMIAL LINK LOGIT
(OUT)     BINOMIAL DENOMINATOR N
(OUT)
(OUT)     LINEAR PREDICTOR
(OUT)     %GM G
(OUT)
(OUT)             ESTIMATE      S.E.       PARAMETER
(OUT)          1 -0.5971      0.9490E-02  %GM
(OUT)          2 -0.1739      0.1990E-01  G(2)
(OUT)        SCALE PARAMETER TAKEN AS      1.000
```

```
(IN)      $FIT -G+AL $DIS ME $
(OUT)              SCALED
(OUT)      CYCLE  DEVIANCE       DF
(OUT)         3     3025.        34
(OUT)
(OUT)     Y-VARIATE  R
(OUT)     ERROR BINOMIAL LINK LOGIT
(OUT)     BINOMIAL DENOMINATOR N
(OUT)
(OUT)     LINEAR PREDICTOR
(OUT)     %GM AL
(OUT)
(OUT)            ESTIMATE       S.E.       PARAMETER
(OUT)         1 -0.7938       0.9093E-02  %GM
(OUT)         2   1.265       0.2588E-01  AL(2)
(OUT)        SCALE PARAMETER TAKEN AS     1.000
(OUT)
(IN)      $FIT -AL+SK $DIS ME $
(OUT)              SCALED
(OUT)      CYCLE  DEVIANCE       DF
(OUT)         3     4627.        33
(OUT)
(OUT)     Y-VARIATE  R
(OUT)     ERROR BINOMIAL LINK LOGIT
(OUT)     BINOMIAL DENOMINATOR N
(OUT)
(OUT)     LINEAR PREDICTOR
(OUT)     %GM SK
(OUT)
(OUT)            ESTIMATE       S.E.       PARAMETER
(OUT)         1 -0.5383       0.1284E-01  %GM
(OUT)         2 -0.4638       0.2034E-01  SK(2)
(OUT)         3  0.1483       0.2021E-01  SK(3)
(OUT)        SCALE PARAMETER TAKEN AS     1.000
(OUT)
(IN)      $FIT +A+G+AL $DIS ME $
(OUT)              SCALED
(OUT)      CYCLE  DEVIANCE       DF
(OUT)         3     125.7        29
(OUT)
(OUT)     Y-VARIATE  R
(OUT)     ERROR BINOMIAL LINK LOGIT
(OUT)     BINOMIAL DENOMINATOR N
(OUT)
(OUT)     LINEAR PREDICTOR
(OUT)     %GM SK A G AL
(OUT)
(OUT)            ESTIMATE       S.E.       PARAMETER
(OUT)         1 -0.9390       0.1752E-01  %GM
(OUT)         2 -0.4402       0.2106E-01  SK(2)
(OUT)         3  0.1009       0.2102E-01  SK(3)
(OUT)         4  0.6788       0.1788E-01  A(2)
(OUT)         5 -0.6853       0.4447E-01  A(3)
(OUT)         6 -0.5246E-01   0.2079E-01  G(2)
(OUT)         7   1.261       0.2686E-01  AL(2)
(OUT)        SCALE PARAMETER TAKEN AS     1.000
```

Die "SCALED DEVIANCE" des Haupteffekt-Modells (125.7 bei "DF"=29) zeigt, daß die Modellanpassung noch nicht hinreichend gut ist. Die "SCALED DEVIANCE" ist ein Maß für die Abweichungen zwischen den Schätzungen des gerechneten Modells und den empirischen Daten (= Schätzwerten des saturierten Modells); sie ist eine Funktion der Likelihoodfunktionen des gerechneten und des saturierten Modells. Die Devianzen sind asymptotisch X^2-verteilt mit den jeweils angegebenen Freiheitsgraden.

Wie im Beispiel 1 wird auch hier als nächstes ein Modell gefittet, welches alle Interaktionseffekte erster Ordnung enthält.

```
(IN)      $FIT +A.G+A.AL+A.SK+G.AL+G.SK+SK.AL $DIS ME $
(OUT)              SCALED
(OUT)       CYCLE  DEVIANCE       DF
(OUT)          3    19.65         16
(OUT)
(OUT)     Y-VARIATE  R
(OUT)     ERROR BINOMIAL LINK LOGIT
(OUT)     BINOMIAL DENOMINATOR N
(OUT)
(OUT)     LINEAR PREDICTOR
(OUT)     %GM SK A G AL SK.A SK.G A.G SK.AL A.AL G.AL
(OUT)
(OUT)              ESTIMATE       S.E.       PARAMETER
(OUT)           1 -0.9888       0.2326E-01  %GM
(OUT)           2 -0.3782       0.3441E-01  SK(2)
(OUT)           3  0.7713E-01   0.3516E-01  SK(3)
(OUT)           4  0.8035       0.3043E-01  A(2)
(OUT)           5 -0.6275       0.6921E-01  A(3)
(OUT)           6  0.9100E-01   0.3809E-01  G(2)
(OUT)           7   1.186       0.4779E-01  AL(2)
(OUT)           8 -0.1809       0.4335E-01  SK(2).A(2)
(OUT)           9 -0.3207E-01   0.1031      SK(2).A(3)
(OUT)          10  0.6599E-02   0.4330E-01  SK(3).A(2)
(OUT)          11 -0.6023E-01   0.1140      SK(3).A(3)
(OUT)          12  0.7107E-01   0.4976E-01  SK(2).G(2)
(OUT)          13  0.8225E-02   0.5033E-01  SK(3).G(2)
(OUT)          14 -0.3495       0.4218E-01  A(2).G(2)
(OUT)          15 -0.1424       0.1276      A(3).G(2)
(OUT)          16  0.4687E-01   0.6498E-01  SK(2).AL(2)
(OUT)          17  0.1340       0.6563E-01  SK(3).AL(2)
(OUT)          18  0.7434E-01   0.5646E-01  A(2).AL(2)
(OUT)          19 -0.3213E-02   0.1873      A(3).AL(2)
(OUT)          20  0.3002E-01   0.9621E-01  G(2).AL(2)
(OUT)          SCALE PARAMETER TAKEN AS     1.000
```

Die Modellanpassung ist zwar mit einer "SCALED DEVIANCE" von "DF"= 19.65 bei 16 Freiheitsgraden hinreichend gut, es werden jedoch trotzdem wiederum noch einige Interaktionseffekte zweiter Ordnung auf Signifikanz geprüft. Sie werden einzeln nacheinander in die Modellformel aufgenommen um die jeweils auftretenden Verbesserungen des Modells beurteilen zu können. Die entsprechenden Anweisungen lauten:

```
$FIT  +A.G.AL   $DIS ME $
$FIT  +A.AL.SK  $DIS ME $
```

```
(IN)     $FIT +A.G.AL $DIS ME $
(OUT)             SCALED
(OUT)     CYCLE  DEVIANCE      DF
(OUT)        4    15.39        14
(OUT)
(OUT)    Y-VARIATE  R
(OUT)    ERROR BINOMIAL LINK LOGIT
(OUT)    BINOMIAL DENOMINATOR N
(OUT)
(OUT)    LINEAR PREDICTOR
(OUT)    %GM SK A G AL SK.A SK.G A.G SK.AL A.AL G.AL A.G.AL
(OUT)
(OUT)           ESTIMATE      S.E.       PARAMETER
(OUT)        1 -0.9895      0.2330E-01  %GM
(OUT)        2 -0.3780      0.3442E-01  SK(2)
(OUT)        3  0.7708E-01  0.3517E-01  SK(3)
(OUT)        4  0.8044      0.3052E-01  A(2)
(OUT)        5 -0.6219      0.6925E-01  A(3)
(OUT)        6  0.9361E-01  0.3836E-01  G(2)
(OUT)        7   1.190      0.4822E-01  AL(2)
(OUT)        8 -0.1810      0.4335E-01  SK(2).A(2)
(OUT)        9 -0.3339E-01  0.1032      SK(2).A(3)
(OUT)       10  0.6623E-02  0.4330E-01  SK(3).A(2)
(OUT)       11 -0.5872E-01  0.1141      SK(3).A(3)
(OUT)       12  0.7056E-01  0.4978E-01  SK(2).G(2)
(OUT)       13  0.8081E-02  0.5034E-01  SK(3).G(2)
(OUT)       14 -0.3525      0.4314E-01  A(2).G(2)
(OUT)       15 -0.1788      0.1299      A(3).G(2)
(OUT)       16  0.4637E-01  0.6500E-01  SK(2).AL(2)
(OUT)       17  0.1347      0.6566E-01  SK(3).AL(2)
(OUT)       18  0.6946E-01  0.5893E-01  A(2).AL(2)
(OUT)       19 -0.8327E-01  0.1928      A(3).AL(2)
(OUT)       20 -0.8054E-02  0.1182      G(2).AL(2)
(OUT)       21  0.5398E-01  0.2053      A(2).G(2).AL(2)
(OUT)       22   2.080       1.147      A(3).G(2).AL(2)
(OUT)       SCALE PARAMETER TAKEN AS    1.000
```

```
(IN)      $FIT +A.AL.SK $DIS ME $
(OUT)              SCALED
(OUT)       CYCLE  DEVIANCE       DF
(OUT)          4     8.936        10
(OUT)
(OUT)     Y-VARIATE  R
(OUT)     ERROR BINOMIAL LINK LOGIT
(OUT)     BINOMIAL DENOMINATOR N
(OUT)
(OUT)     LINEAR PREDICTOR
(OUT)     %GM SK A G AL SK.A SK.G A.G SK.AL A.AL G.AL SK.A.AL A.G.AL
(OUT)
(OUT)              ESTIMATE       S.E.       PARAMETER
(OUT)           1 -0.9908       0.2364E-01  %GM
(OUT)           2 -0.3725       0.3532E-01  SK(2)
(OUT)           3  0.7566E-01   0.3619E-01  SK(3)
(OUT)           4  0.8095       0.3144E-01  A(2)
(OUT)           5 -0.6452       0.7076E-01  A(3)
(OUT)           6  0.9333E-01   0.3837E-01  G(2)
(OUT)           7   1.198       0.5412E-01  AL(2)
(OUT)           8 -0.1969       0.4613E-01  SK(2).A(2)
(OUT)           9  0.2577E-02   0.1062      SK(2).A(3)
(OUT)          10  0.4560E-02   0.4609E-01  SK(3).A(2)
(OUT)          11 -0.6066E-02   0.1178      SK(3).A(3)
(OUT)          12  0.7096E-01   0.4978E-01  SK(2).G(2)
(OUT)          13  0.8832E-02   0.5035E-01  SK(3).G(2)
(OUT)          14 -0.3524       0.4315E-01  A(2).G(2)
(OUT)          15 -0.1780       0.1298      A(3).G(2)
(OUT)          16  0.1066E-01   0.8203E-01  SK(2).AL(2)
(OUT)          17  0.1430       0.8475E-01  SK(3).AL(2)
(OUT)          18  0.1620E-01   0.8942E-01  A(2).AL(2)
(OUT)          19  0.4161       0.2967      A(3).AL(2)
(OUT)          20 -0.8232E-02   0.1183      G(2).AL(2)
(OUT)          21  0.1464       0.1352      SK(2).A(2).AL(2)
(OUT)          22 -0.8008       0.4650      SK(2).A(3).AL(2)
(OUT)          23  0.3174E-01   0.1343      SK(3).A(2).AL(2)
(OUT)          24 -0.8995       0.4690      SK(3).A(3).AL(2)
(OUT)          25  0.5480E-01   0.2051      A(2).G(2).AL(2)
(OUT)          26   2.144        1.167      A(3).G(2).AL(2)
(OUT)          SCALE PARAMETER TAKEN AS     1.000
```

Da die Interaktionen zweiter Ordnung nicht (auch nicht in Teilinteraktionen) signifikant von 0 verschieden sind, d.h. keinen zusätzlichen Erklärungswert besitzen, werden sie dem Modell wieder entnommen.

```
$FIT   -A.G.AL   $DIS ME $
$FIT   -A.AL.SK  $DIS ME $
```

```
(IN)    $FIT -A.G.AL $DIS ME $
(OUT)             SCALED
(OUT)     CYCLE  DEVIANCE      DF
(OUT)        3    13.26        12
(OUT)
(OUT)    Y-VARIATE  R
(OUT)    ERROR BINOMIAL LINK LOGIT
(OUT)    BINOMIAL DENOMINATOR N
(OUT)
(OUT)    LINEAR PREDICTOR
(OUT)    %GM SK A G AL SK.A SK.G A.G SK.AL A.AL G.AL SK.A.AL
(OUT)
(OUT)             ESTIMATE      S.E.       PARAMETER
(OUT)          1 -0.9900      0.2360E-01  %GM
(OUT)          2 -0.3727      0.3532E-01  SK(2)
(OUT)          3  0.7558E-01  0.3619E-01  SK(3)
(OUT)          4  0.8086      0.3134E-01  A(2)
(OUT)          5 -0.6508      0.7075E-01  A(3)
(OUT)          6  0.9066E-01  0.3811E-01  G(2)
(OUT)          7   1.194      0.5373E-01  AL(2)
(OUT)          8 -0.1969      0.4613E-01  SK(2).A(2)
(OUT)          9  0.2751E-02  0.1062      SK(2).A(3)
(OUT)         10  0.4582E-02  0.4609E-01  SK(3).A(2)
(OUT)         11 -0.5503E-02  0.1178      SK(3).A(3)
(OUT)         12  0.7155E-01  0.4977E-01  SK(2).G(2)
(OUT)         13  0.9125E-02  0.5034E-01  SK(3).G(2)
(OUT)         14 -0.3494      0.4218E-01  A(2).G(2)
(OUT)         15 -0.1420      0.1276      A(3).G(2)
(OUT)         16  0.1076E-01  0.8204E-01  SK(2).AL(2)
(OUT)         17  0.1429      0.8476E-01  SK(3).AL(2)
(OUT)         18  0.2113E-01  0.8772E-01  A(2).AL(2)
(OUT)         19  0.4864      0.2926      A(3).AL(2)
(OUT)         20  0.2987E-01  0.9622E-01  G(2).AL(2)
(OUT)         21  0.1464      0.1352      SK(2).A(2).AL(2)
(OUT)         22 -0.7535      0.4552      SK(2).A(3).AL(2)
(OUT)         23  0.3169E-01  0.1343      SK(3).A(2).AL(2)
(OUT)         24 -0.9067      0.4638      SK(3).A(3).AL(2)
(OUT)         SCALE PARAMETER TAKEN AS    1.000
```

```
(IN)      $FIT -SK.A.AL $DIS ME $
(OUT)              SCALED
(OUT)      CYCLE   DEVIANCE       DF
(OUT)         3     19.65         16
(OUT)
(OUT)     Y-VARIATE  R
(OUT)     ERROR BINOMIAL LINK LOGIT
(OUT)     BINOMIAL DENOMINATOR N
(OUT)
(OUT)     LINEAR PREDICTOR
(OUT)     %GM SK A G AL SK.A SK.G A.G SK.AL A.AL G.AL
(OUT)
(OUT)              ESTIMATE       S.E.       PARAMETER
(OUT)          1 -0.9888       0.2326E-01  %GM
(OUT)          2 -0.3782       0.3441E-01  SK(2)
(OUT)          3  0.7713E-01   0.3516E-01  SK(3)
(OUT)          4  0.8035       0.3043E-01  A(2)
(OUT)          5 -0.6275       0.6921E-01  A(3)
(OUT)          6  0.9100E-01   0.3809E-01  G(2)
(OUT)          7   1.186       0.4779E-01  AL(2)
(OUT)          8 -0.1809       0.4335E-01  SK(2).A(2)
(OUT)          9 -0.3207E-01   0.1031      SK(2).A(3)
(OUT)         10  0.6599E-02   0.4330E-01  SK(3).A(2)
(OUT)         11 -0.6023E-01   0.1140      SK(3).A(3)
(OUT)         12  0.7107E-01   0.4976E-01  SK(2).G(2)
(OUT)         13  0.8225E-02   0.5033E-01  SK(3).G(2)
(OUT)         14 -0.3495       0.4218E-01  A(2).G(2)
(OUT)         15 -0.1424       0.1276      A(3).G(2)
(OUT)         16  0.4687E-01   0.6498E-01  SK(2).AL(2)
(OUT)         17  0.1340       0.6563E-01  SK(3).AL(2)
(OUT)         18  0.7434E-01   0.5646E-01  A(2).AL(2)
(OUT)         19 -0.3213E-02   0.1873      A(3).AL(2)
(OUT)         20  0.3002E-01   0.9621E-01  G(2).AL(2)
(OUT)         SCALE PARAMETER TAKEN AS     1.000
```

Die o.g. Fit-Ergebnisse zeigen, daß auch die Interaktionen erster Ordnung SK.G, A.AL und G.AL in keiner der Teilinteraktionen signifikant von 0 verschiedene Ergebnisse zeigen. Sie werden ebenfalls nacheinander wieder aus dem Modell ausgeschlossen. Dabei ist nach jedem Schritt zu prüfen, ob sich Auswirkungen auf die Signifikanz anderer Parameter ergeben haben. In diesem Fall ergibt sich wieder das bereits auf S. 37 dargestellte Modell.

```
(IN)      $FIT -SK.G $DIS ME $
(OUT)              SCALED
(OUT)       CYCLE  DEVIANCE      DF
(OUT)          3    21.88        18
(OUT)
(OUT)     Y-VARIATE  R
(OUT)     ERROR BINOMIAL LINK LOGIT
(OUT)     BINOMIAL DENOMINATOR N
(OUT)
(OUT)     LINEAR PREDICTOR
(OUT)     %GM SK A G AL SK.A A.G SK.AL A.AL G.AL
(OUT)
(OUT)              ESTIMATE      S.E.        PARAMETER
(OUT)          1 -0.9955       0.2222E-01  %GM
(OUT)          2 -0.3581       0.3135E-01  SK(2)
(OUT)          3  0.7950E-01   0.3200E-01  SK(3)
(OUT)          4  0.8047       0.3043E-01  A(2)
(OUT)          5 -0.6244       0.6919E-01  A(3)
(OUT)          6  0.1142       0.2953E-01  G(2)
(OUT)          7   1.191       0.4756E-01  AL(2)
(OUT)          8 -0.1829       0.4333E-01  SK(2).A(2)
(OUT)          9 -0.4163E-01   0.1028      SK(2).A(3)
(OUT)         10  0.6642E-02   0.4322E-01  SK(3).A(2)
(OUT)         11 -0.6114E-01   0.1138      SK(3).A(3)
(OUT)         12 -0.3507       0.4220E-01  A(2).G(2)
(OUT)         13 -0.1416       0.1276      A(3).G(2)
(OUT)         14  0.3310E-01   0.6425E-01  SK(2).AL(2)
(OUT)         15  0.1322       0.6492E-01  SK(3).AL(2)
(OUT)         16  0.7385E-01   0.5645E-01  A(2).AL(2)
(OUT)         17 -0.2967E-02   0.1873      A(3).AL(2)
(OUT)         18  0.3088E-01   0.9632E-01  G(2).AL(2)
(OUT)         SCALE PARAMETER TAKEN AS     1.000
```

```
(IN)      $FIT -A.AL $DIS ME $
(OUT)              SCALED
(OUT)      CYCLE  DEVIANCE      DF
(OUT)         3     23.63       20
(OUT)
(OUT)     Y-VARIATE  R
(OUT)     ERROR BINOMIAL LINK LOGIT
(OUT)     BINOMIAL DENOMINATOR N
(OUT)     LINEAR PREDICTOR
(OUT)     %GM SK A G AL SK.A A.G SK.AL G.AL
(OUT)
(OUT)             ESTIMATE       S.E.       PARAMETER
(OUT)          1   -1.001      0.2184E-01   %GM
(OUT)          2 -0.3581      0.3137E-01   SK(2)
(OUT)          3  0.7901E-01  0.3202E-01   SK(3)
(OUT)          4   0.8148      0.2941E-01   A(2)
(OUT)          5 -0.6208      0.6826E-01   A(3)
(OUT)          6   0.1184      0.2935E-01   G(2)
(OUT)          7    1.217      0.4256E-01   AL(2)
(OUT)          8 -0.1822      0.4331E-01   SK(2).A(2)
(OUT)          9 -0.4193E-01  0.1029       SK(2).A(3)
(OUT)         10  0.7386E-02  0.4321E-01   SK(3).A(2)
(OUT)         11 -0.6175E-01  0.1138       SK(3).A(3)
(OUT)         12 -0.3586      0.4175E-01   A(2).G(2)
(OUT)         13 -0.1443      0.1273       A(3).G(2)
(OUT)         14  0.3375E-01  0.6410E-01   SK(2).AL(2)
(OUT)         15  0.1364      0.6467E-01   SK(3).AL(2)
(OUT)         16  0.2723E-01  0.9607E-01   G(2).AL(2)
(OUT)         SCALE PARAMETER TAKEN AS     1.000
(OUT)
(IN)      $FIT -G.AL $DIS ME $
(OUT)              SCALED
(OUT)      CYCLE  DEVIANCE      DF
(OUT)         3     23.71       21
(OUT)
(OUT)     Y-VARIATE  R
(OUT)     ERROR BINOMIAL LINK LOGIT
(OUT)     BINOMIAL DENOMINATOR N
(OUT)     LINEAR PREDICTOR
(OUT)     %GM SK A G AL SK.A A.G SK.AL
(OUT)
(OUT)             ESTIMATE       S.E.       PARAMETER
(OUT)          1   -1.001      0.2178E-01   %GM
(OUT)          2 -0.3581      0.3137E-01   SK(2)
(OUT)          3  0.7903E-01  0.3202E-01   SK(3)
(OUT)          4   0.8150      0.2940E-01   A(2)
(OUT)          5 -0.6204      0.6825E-01   A(3)
(OUT)          6   0.1201      0.2873E-01   G(2)
(OUT)          7    1.219      0.4172E-01   AL(2)
(OUT)          8 -0.1822      0.4331E-01   SK(2).A(2)
(OUT)          9 -0.4196E-01  0.1029       SK(2).A(3)
(OUT)         10  0.7357E-02  0.4321E-01   SK(3).A(2)
(OUT)         11 -0.6182E-01  0.1138       SK(3).A(3)
(OUT)         12 -0.3593      0.4167E-01   A(2).G(2)
(OUT)         13 -0.1454      0.1272       A(3).G(2)
(OUT)         14  0.3373E-01  0.6410E-01   SK(2).AL(2)
(OUT)         15  0.1363      0.6467E-01   SK(3).AL(2)
(OUT)         SCALE PARAMETER TAKEN AS     1.000
```

6.5 KONSTRUKTION DES OPTIMALEN MODELLS

Mit dem Modell

SK+A+G+AL+SK.A+A.G+SK.AL

kann die Suche nach weiteren signifikanten Einflußgrößen (Interaktionen) beendet werden. Vergleichen wir dieses Modell mit dem entsprechenden des Beispiels 1, stellen wir fest, daß es wesentlich sparsamer ist, d.h. weniger Interaktionen enthält, nur 3 Interaktionen erster Ordnung, keine Interaktion zweiter Ordnung.

Die "SCALED DEVIANCE" (vgl. 6.4) beträgt 23.71 bei "DF" = 21. Da die Devianzen mit den jeweils angegebenen Freiheitsgraden asymptotisch X^2 verteilt sind, stellen wir fest, daß dieses Modell eine gute Anpassung hat.

Im nächsten Schritt ist zu prüfen, ob alle im Modell befindlichen Parameter signifikant von 0 verschieden sind, denn ein optimales Modell sollte nur signifikante Effekte enthalten.
Hier gilt wie im Beispiel 1 der Test "ESTIMATE"/"S.E."> 2. Es läßt sich leicht feststellen, daß alle Parameter der Haupteffekte signifikant von 0 verschieden sind. Von den drei Interaktionen erster Ordnung, die noch im Modell enthalten sind, ist jeweils nur eine Teilinteraktion signifikant von 0 verschieden: SK(2).A(2), A(2).G(2) und SK(3).AL(2).
Um die nichtsignifikanten Teilinteraktionen aus dem Modell auszuschließen, werden im nächsten Schritt die signifikanten Teilinteraktionen durch entsprechende cal-Statements einzeln definiert. Hier ist wieder darauf zu achten, sinnvolle Abkürzungen zu finden, die nur aus maximal 4 Zeichen bestehen dürfen.

```
(IN)      $CAL SK2=%IF(%EQ(SK,2),1,0)
(IN)      $CAL SK3=%IF(%EQ(SK,3),1,0)
(IN)      $CAL AL2=%IF(%EQ(AL,2),1,0)
(IN)      $CAL A2=%IF(%EQ(A,2),1,0)
(IN)      $CAL G2=%IF(%EQ(G,2),1,0)
(IN)      $CAL SKA2=SK2*A2
(IN)      $CAL A2G2=A2*G2
(IN)      $CAL SKAL=SK3*AL2
```

Nach der Kalkulation dieser neuen Parameter (Effekte), müssen diese noch in die Modellformel eingebracht werden.

optimales Modell:

```
(IN)      $FIT -SK.A-A.G-SK.AL+SKA2+A2G2+SKAL   $DIS ME $
(OUT)            SCALED
(OUT)     CYCLE  DEVIANCE      DF
(OUT)         3    25.74        26
(OUT)
(OUT)     Y-VARIATE  R
(OUT)     ERROR BINOMIAL LINK LOGIT
(OUT)     BINOMIAL DENOMINATOR N

(OUT)     LINEAR PREDICTOR
(OUT)     %GM SK A G AL SKA2 A2G2 SKAL
(OUT)
(OUT)           ESTIMATE      S.E.       PARAMETER
(OUT)        1  -1.001      0.1917E-01  %GM
(OUT)        2 -0.3546      0.2729E-01  SK(2)
(OUT)        3  0.8196E-01  0.2231E-01  SK(3)
(OUT)        4  0.8161      0.2334E-01  A(2)
(OUT)        5 -0.6690      0.4448E-01  A(3)
(OUT)        6  0.1126      0.2797E-01  G(2)
(OUT)        7   1.233      0.3187E-01  AL(2)
(OUT)        8 -0.1849      0.3871E-01  SKA2
(OUT)        9 -0.3520      0.4115E-01  A2G2
(OUT)       10  0.1224      0.5861E-01  SKAL
(OUT)       SCALE PARAMETER TAKEN AS    1.000
```

Da bei Ausschluß der nichtsignifikanten Teilinteraktionen keine Umlagerungen erfolgten, die die Signifikanz anderer Parameter beeinträchtigten, ist das optimale Modell gefunden.

Die "SCALED DEVIANCE" des optimalen Modells beträgt 25.74 bei "DF" = 26. Damit ist das Modell gut angepaßt. Das optimale Modell enthält nur 10 Parameter. Vergleicht man das o.g. optimale Modell mit dem aus Beispiel 1, so stellt man fest, daß dieses Modell, in dem die Fahrunfallanteile geschätzt werden, bei gleich guter Anpassung 7 Parameter weniger enthält. Dies ist auch im Hinblick auf die Ergebnisinterpretation vorteilhaft.

Auch für dieses Modell sollen die Residuen ausgegeben werden. Dies erfolgt mit der $DIS-Anweisung:

```
$DIS R $
R = Residuen
```

```
(IN)     $DIS R $
(OUT)         UNIT   OBSERVED OUT OF   FITTED      RESIDUAL
(OUT)            1       2245    8271   2223.        0.5539
(OUT)            2       1458    7241   1484.       -0.7586
(OUT)            3       1513    5251   1497.        0.4810
(OUT)            4        892    1610   897.9       -0.2946
(OUT)            5        520    1129   529.9       -0.5896
(OUT)            6        628    1043   633.4       -0.3451
(OUT)            7        939    3262   950.6       -0.4481
(OUT)            8        624    2620   586.6         1.751
(OUT)            9        631    2091   645.4       -0.6796
(OUT)           10         78     143   83.69       -0.9659
(OUT)           11         51     100   49.75        0.2509
(OUT)           12         65      99   62.75        0.4695
(OUT)           13       3290    7254   3293.      -0.5923E-01
(OUT)           14       1655    5121   1672.       -0.4945
(OUT)           15       2416    5097   2417.      -0.3811E-01
(OUT)           16        793    1071   792.9       0.4765E-02
(OUT)           17        434     670   418.4         1.247
(OUT)           18        791    1011   786.2        0.3604
(OUT)           19       1005    2543   1006.      -0.2713E-01
(OUT)           20        523    1906   526.3       -0.1671
(OUT)           21        644    1550   643.6       0.2120E-01
(OUT)           22         51      78   53.96       -0.7255
(OUT)           23         32      49   27.78         1.218
(OUT)           24         47      61   44.75        0.6524
(OUT)           25        243    1511   239.4        0.2559
(OUT)           26        168    1362   158.9        0.7708
(OUT)           27        120     713   121.0      -0.9601E-01
(OUT)           28         24      49   19.23         1.396
(OUT)           29         11      45   14.03       -0.9753
(OUT)           30         12      37   16.36        -1.442
(OUT)           31         44     268   46.64       -0.4246
(OUT)           32         20     226   29.10        -1.807
(OUT)           33         21     110   20.47        0.1293
(OUT)           34          2       2  0.8391         1.663
(OUT)           35          1       2  0.6729        0.4895
(OUT)           36          1       1  0.4700         1.062
```

Die standardisierten Residuen sind grundsätzlich näherungsweise normalverteilt. Abweichungen, die größer als (+/- 2) sind, sollten daher bei einem gut angepaßten Modell mit einer Wahrscheinlichkeit von höchstens 5% auftreten.
Wie sich leicht feststellen läßt, liegen alle standardisierten Residuen im Intervall (+2, -2). Dies zeigt, daß das Modell gut angepaßt ist.
Neben den Residuen soll für das optimale Modell auch die (Co-) Varianz-Matrix und die Matrix der Standard-Fehler ausgegeben werden; dies erfolgt wieder mit der $DIS-Anweisung:

```
$DIS VS $
V = (Co-)Varianz-Matrix
S = Standardfehler der Differenzen (S.E. of Differences)
```

```
(IN)     $DIS VS $
(CO)VARIANCE MATRIX
 1  3.6740E-04
 2 -2.7282E-04  7.4501E-04
 3 -1.8951E-04  1.6997E-04  4.9761E-04
 4 -2.7502E-04  2.0163E-04 -1.0654E-05  5.4453E-04
 5 -1.9754E-04 -3.3973E-06  2.2506E-05  1.7949E-04  1.9784E-03
 6 -2.1906E-04  8.8345E-06 -1.0855E-07  2.0806E-04  9.5268E-05  7.8208E-04
 7 -1.7279E-04 -1.1258E-05  1.1851E-04  5.8589E-05  9.2981E-05  1.0287E-04
 8  2.0167E-04 -6.7370E-04  1.5659E-05 -4.2906E-04  1.0303E-05 -1.0214E-05
 9  2.0288E-04 -4.4700E-06  1.1121E-05 -4.1276E-04 -8.7758E-05 -7.7357E-04
10  1.1992E-04  3.3157E-05 -4.2308E-04  1.0306E-06 -3.4156E-05 -7.7050E-06
       1           2           3           4           5           6
 7  1.0157E-03
 8  8.5720E-06  1.4986E-03
 9 -4.1549E-05  2.9077E-06  1.6932E-03
10 -9.9333E-04 -5.4163E-05  9.0251E-06  3.4347E-03
       7           8           9          10
  SCALE PARAMETER TAKEN AS    1.000


  S.E. OF DIFFERENCES
 1  0.0000
 2  4.0719E-02  0.0000
 3  3.5271E-02  3.0045E-02  0.0000
 4  3.8236E-02  2.9771E-02  3.2611E-02  0.0000
 5  5.2354E-02  5.2252E-02  4.9306E-02  4.6519E-02  0.0000
 6  3.9845E-02  3.8851E-02  3.5776E-02  3.0174E-02  5.0695E-02  0.0000
 7  4.1577E-02  4.2228E-02  3.5725E-02  3.7987E-02  5.2992E-02  3.9900E-02
 8  3.8244E-02  5.9925E-02  4.4327E-02  5.3863E-02  5.8791E-02  4.7970E-02
 9  4.0680E-02  4.9469E-02  4.6568E-02  5.5347E-02  6.2026E-02  6.3423E-02
10  5.9685E-02  6.4136E-02  6.9127E-02  6.3065E-02  7.4037E-02  6.5056E-02
       1           2           3           4           5           6
 7  0.0000
 8  4.9971E-02  0.0000
 9  5.2840E-02  5.6445E-02  0.0000
10  8.0231E-02  7.1004E-02  7.1484E-02  0.0000
       7           8           9          10
  SCALE PARAMETER TAKEN AS    1.000
```

Mit Hilfe der Standardfehler der Differenzen wollen wir testen, ob sich die Effekte signifikant voneinander unterscheiden (vgl. Abschnitt 5.5). Z. B. Vergleich von SK(2) mit SK(3):
(0,3546-0,08196)/0,030045=2.727908 > 2,0.α < 0,01.

Es ergibt sich, daß sich alle zu einer Variablen gehörenden Effekte signifikant voneinander unterscheiden. Damit sind alle Bedingungen, die an ein optimales Modell gestellt werden erfüllt:

- das Modell bildet den empirischen Befund hinreichend gut ab
- alle im Modell befindlichen Parameter sind signifikant von 0 verschieden
- alle im Modell befindlichen Parameter sind signifikant voneinander verschieden.

Für binomiale Logit-Modelle (Schätzung von Anteilswerten) lassen sich in GLIM unter Verwendung des optimalen Modells für alle Variablen-Konstellationen der untersuchten Kontingenztafel Erwartungswerte für den Anteil der Fahrunfälle schätzen. Dabei fließen je nach Variablenkonstellation, neben den Haupteffekten auch die Interaktionseffekte des optimalen Modells mit ein.

Dazu dient ein Makro (Unterprogramm) folgenden Aufbaus:

```
$MAC PRO
$CAL PR = %FV/N
$ENDMAC$
```

%FV = geschätzte Fahrunfallanteilswerte
N = Gesamtzahl der Unfälle

Mit $USE PRO $ wird das Makro wieder aufgerufen.
Mit $LOOK A G AL SK PR $ wird für jede Zelle der untersuchten Kontingenztafel die sie kennzeichnende Variablenkonstellation und der geschätzte Erwartungswert für den Anteil der Fahrunfälle ausgegeben.

```
(IN)     $MAC PRO
(IN)     $CAL PR=%FV/N
(IN)     $ENDMAC $
(IN)     $USE PRO $
(IN)     $LOOK A G AL SK PR $
(OUT)      1    1.000       1.000       1.000       1.000       0.2687
(OUT)      2    1.000       1.000       1.000       2.000       0.2050
(OUT)      3    1.000       1.000       1.000       3.000       0.2851
(OUT)      4    1.000       1.000       2.000       1.000       0.5577
(OUT)      5    1.000       1.000       2.000       2.000       0.4693
(OUT)      6    1.000       1.000       2.000       3.000       0.6073
(OUT)      7    1.000       2.000       1.000       1.000       0.2914
(OUT)      8    1.000       2.000       1.000       2.000       0.2239
(OUT)      9    1.000       2.000       1.000       3.000       0.3086
(OUT)     10    1.000       2.000       2.000       1.000       0.5853
(OUT)     11    1.000       2.000       2.000       2.000       0.4975
(OUT)     12    1.000       2.000       2.000       3.000       0.6338
(OUT)     13    2.000       1.000       1.000       1.000       0.4539
(OUT)     14    2.000       1.000       1.000       2.000       0.3264
(OUT)     15    2.000       1.000       1.000       3.000       0.4743
(OUT)     16    2.000       1.000       2.000       1.000       0.7404
(OUT)     17    2.000       1.000       2.000       2.000       0.6244
(OUT)     18    2.000       1.000       2.000       3.000       0.7777
(OUT)     19    2.000       2.000       1.000       1.000       0.3955
(OUT)     20    2.000       2.000       1.000       2.000       0.2761
(OUT)     21    2.000       2.000       1.000       3.000       0.4152
(OUT)     22    2.000       2.000       2.000       1.000       0.6918
(OUT)     23    2.000       2.000       2.000       2.000       0.5668
(OUT)     24    2.000       2.000       2.000       3.000       0.7336
(OUT)     25    3.000       1.000       1.000       1.000       0.1584
(OUT)     26    3.000       1.000       1.000       2.000       0.1166
(OUT)     27    3.000       1.000       1.000       3.000       0.1697
(OUT)     28    3.000       1.000       2.000       1.000       0.3924
(OUT)     29    3.000       1.000       2.000       2.000       0.3118
(OUT)     30    3.000       1.000       2.000       3.000       0.4420
(OUT)     31    3.000       2.000       1.000       1.000       0.1740
(OUT)     32    3.000       2.000       1.000       2.000       0.1288
(OUT)     33    3.000       2.000       1.000       3.000       0.1861
(OUT)     34    3.000       2.000       2.000       1.000       0.4196
(OUT)     35    3.000       2.000       2.000       2.000       0.3364
(OUT)     36    3.000       2.000       2.000       3.000       0.4700
```

6.6 GRAPHISCHE DARSTELLUNG DER RESIDUEN

Wie für das optimale Modell des Beispiels 1 (vgl. Abschnitt 5.6) sollen auch hier die Residuen graphisch dargestellt werden.
Dazu wird in GLIM ein Makro geschrieben, das die standardisierten Residuen berechnet:

```
$MAC RES
$CAL P = %FV/N
$CAL RE = (R/N-P)
$CAL H = P * (1-P)/N
$CAL RE = RE/(%SQRT(H))
$ENDMAC$
```

%FV = geschätzte Werte
N = Zahl der Unfälle insgesamt
R = Zahl der Fahrunfälle
%SQRT(H) = $\sqrt{H}$

Mit $USE RES $ wird das Makro wieder aufgerufen.
Die Anweisung $PLO RE A $ erzeugt ein Scattergramm. Die standardisierten Residuen werden auf der y-Achse und A (3 Lebensaltersklassen) auf der x-Achse abgebildet.
Das Diagramm zeigt, daß alle standardisierten Residuen im Intervall (+2, -2) liegen. Die standardisierten Residuen der Altersgruppe A(3) (über 60 Jahre) weisen die stärkste Streuung auf.

```
(IN)  $MAC RES
(IN)  $CAL P=%FV/N
(IN)  $CAL RE=(R/N)-P
(IN)  $CAL H=P*(1-P)/N
(IN)  $CAL RE=RE/(%SQRT(H))
(IN)  $ENDMAC $
(IN)  $USE RES $
(IN)  $PLO RE A $
(OUT)  2.00     *
(OUT)  1.75     *    R                                                  R
(OUT)  1.50     *                                                       R
(OUT)  1.25     *                              2
(OUT)  1.00     *                                                       R
(OUT)  0.750    *                              R                       R
(OUT)  0.500    *    3                                                  R
(OUT)  0.250    *    R                         R                       2
(OUT)  0.000    *                              5                       R
(OUT) -0.250    *    2                         R
(OUT) -0.500    *    2                         R                       R
(OUT) -0.750    *    2                         R
(OUT) -1.00     *    R                                                  R
(OUT) -1.25     *
(OUT) -1.50     *                                                       R
(OUT) -1.75     *                                                       R
(OUT) -2.00     *
(OUT)  .........*.........*.........*.........*.........*.........*.......
(OUT)          0.800     1.20      1.60      2.00      2.40      2.80
```

Wie in Beispiel 1, ist es auch hier möglich, die standardisierten Residuen zu sortieren und gegen die gleiche Anzahl sortierter Zufallszahlen der Standardnormalverteilung abzubilden. Folgende GLIM-Statements sind zu formulieren:

```
(IN)     $USE RES
(IN)     $CAL GRE=RE $SORT GRE
(IN)     $CAL NOR=%ND(%SR(0))  $SORT NOR
(IN)     $PLO GRE NOR $
(OUT)      2.00     *                                                  G    G
(OUT)      1.75     *                                                G
(OUT)      1.50     *                                             GG
(OUT)      1.25     *                                           G
(OUT)      1.00     *                                         2
(OUT)     0.750     *                                       22
(OUT)     0.500     *                                 G2G
(OUT)     0.250     *                             G2 2G
(OUT)     0.000     *                          GGG
(OUT)    -0.250     *                       3G
(OUT)    -0.500     *                  G2
(OUT)    -0.750     *                GG
(OUT)     -1.00     *
(OUT)     -1.25     *
(OUT)     -1.50     *   G
(OUT)     -1.75     *G
(OUT)     -2.00     *
(OUT)               .........*.........*.........*.........*.........*.........*.........*
(OUT)                     -2.00      -1.00      0.000      1.00       2.00       3.00
```

Der letzte Plot zeigt, daß die standardisierten Residuen auf einer 45°-Geraden liegen; das Modell ist damit den Daten gut angepaßt; Ausreißer liegen nicht vor.
Wenn keine weiteren Analysen durchgeführt werden sollen, wird GLIM mit $STOP verlassen.

6.7 INTERPRETATION DES OPTIMALEN MODELLS

Tabelle 1: Einbeziehung der Haupteffekte in das Logit-Modell (vgl. Abschnitt 6.4)

Modell	Einbezogener Effekt	Devianz	Freiheitsgrade (DF)	reduz.Devianz zum Basismodell (Mod.1)	erklärte Devianz[1]
1	GM	5497,0	35	-	-
2	GM + A	3238,0	33	2259,0	41,1%
3	GM + G	5420,0	34	77,0	0,1%
4	GM + AL	3025,0	34	2472,0	45,0%
5	GM + SK	4627,0	33	870,0	15,8%
6	GM+A+G+AL+SK	125,7	29	5371,3	97,5%

[1] Wird die mit Modell 2 bis 6 errechnete Devianz auf die durch Modell 1 gegebene Devianz (Devianz des Basismodells) bezogen, erhält man den Anteil der durch die Parameter des jeweiligen Modells erklärten Devianz (vgl. Anhang 6, Beurteilung der Modellanpassung).

Modell 1, welches zur Schätzung der Zellenhäufigkeiten nur den GM (= Regressionskonstante) verwendet, gibt die Hypothese wieder, daß die Variablen A,AL,G,SK keinen Einfluß auf den Anteil der Fahrunfälle haben. Wie die erklärten Devianzen der Modelle 2 bis 6 zeigen, sind die Erklärungsbeiträge der einzelnen Variablen höchst unterschiedlich. Eine Devianzreduzierung von 45,0% übt die Variable Alkoholeinfluß (AL), gefolgt von der Variablen Alter des Beteiligten (A) mit 41,1% aus. Deutlich geringer ist der Einfluß der Variablen Straßenklasse (SK) mit 15,8% und insbesondere der Variablen Geschlecht (G) mit 0.1%. Alle im Modell enthaltenen Haupteffekte zusammen reduzieren die Devianz des Modells um 97,5%, d.h. es bleibt noch ein unerklärter Rest von 2,5%.

Es wurden erweiterte Modelle unter Einschluß aller Haupteffekte sowie der Interaktionen 1. Ordnung (Zweier-Interaktionen) und unter sachlichen Kriterien ausgewählte Interaktionen 2. Ordnung (Dreier-Interaktionen) untersucht. Nach schrittweisem Ausschluß aller nichtsignifikanten Parameter ergab sich ein optimales Modell mit 4 Haupteffekten und 3 Interaktionseffekten 1. Ordnung.

Da die Haupteffekte Alter des Beteiligten und Straßenklasse mehr als zwei Kategorien (einschließlich der Basiskategorie) haben, enthält das optimale Modell 10 Parameter. Durch dieses Modell werden 99,5% der Devianz des Basismodells erklärt. Damit ist das optimale Modell gefunden.

Wie schon in Abschnitt 5.7 ausgeführt, ist bei der Entwicklung eines optimalen Modells ein Kompromiß zwischen den konkurrierenden Zielen zu suchen, einerseits mit möglichst wenig Parametern auszukommen und andererseits den empirischen Befund so genau wie möglich abzubilden. Hier wurde als Optimierungskriterium der Einfluß aller signifikanten ($\alpha \leq 0{,}05$) Parameter gewählt. (Als weitere Kriterien kämen z.B. infrage: Suche eines Modells, welches mit möglichst wenig signifikanten Parametern (1) 95% der Devianz des Basismodells erklärt oder (2) die Bedingung erfüllt Devianz $\leq X^2_{0,05;DF}$.)

Wie gut die angegebenen 10 Parameter des optimalen Modells den empirischen Befund abbilden, zeigt auch die graphische Darstellung der standardisierten Residuen (vgl. Abschnitt 6.6).

Die im optimalen Modell enthaltenen Modellparameter sind einer ersten Interpretation leicht zugänglich:

Negative Werte der "ESTIMATE" bedeuten, daß der Fahrunfallanteil bei Geltung dieser Parameter abnimmt, positive Werte, daß er zunimmt. Der größte positive Haupteffekt ist mit dem Parameter AL2 (mit Alkohol) gegeben; dementsprechend hoch ist der Einfluß des Alkohols auf den Anstieg des Fahrunfallanteils.

Aus dem Vorzeichen des Parameters A2 (junge Fahrer) und A3 (alte Fahrer) läßt sich direkt ablesen, daß junge Fahrer häufiger und alte Fahrer weniger häufig Fahrunfälle verursachen, als Fahrer der mittleren Altersgruppe (A1 ist Basiskategorie, d.h. der Parameter dieser Ausprägung wird gleich Null gesetzt und liegt damit zwischen den Werte von A2 und A3). Der Abstand zwischen den Werten ("ESTIMATE") von A2 und A3 ist sogar noch größer als der Wert von AL2; dies bedeutet, daß mit dem Alter ein vergleichsweise sehr starker Einfluß einhergeht.

Der Einfluß des Parameters G2 (weibliche Fahrer) ist, betrachtet man nur den Haupteffekt, relativ schwach (schwach positiv), d.h. der Einfluß des Geschlechts auf den Anteil der Fahrunfälle ist entsprechend gering.
Im Falle der Interaktion von G2 mit A2 wird der schwach positive Haupteffekt durch eine Interaktion mit negativem Vorzeichen überlagert. Das bedeutet bei der Interaktion A2G2: junge Frauen unter 25 Jahre haben etwas niedrigere Fahrunfallanteile als die jungen Männer, während in den darüber liegenden Altersgruppen Frauen einen etwas höheren Fahrunfallanteil haben.

Die Werte der Haupteffekte SK2 (Bundesstraßen) und SK3 (Kreis- und Gemeindestraßen) zeigen in Verbindung mit der Basiskategorie SK1 (Landesstraßen), daß der Fahrunfallanteil von Bundesstraßen zu Kreis- und Gemeindestraßen zunimmt.

Unter Verwendung des optimalen Modells lassen sich mit GLIM für beliebige Kovariatenkonstellationen Erwartungswerte für den Anteil der Fahrunfälle schätzen. Die Schätzwerte des optimalen Modells wurden bereits in Abschnitt 6.5, letzter Ausdruck, dargestellt. Dazu sei angemerkt, daß je nach Variablenkonstellation neben den Haupteffekten die Interaktionseffekte des optimalen Modells mit einfließen. Die folgende inhaltliche Interpretation beruht auf einer Auswahl der o.g. Erwartungswerte des Abschnitts 6.5:

Unter der Bedingung, daß für alle im Modell vorhandenen Parameter die Basiskategorie gilt, beträgt der Erwartungswert des Fahrunfallanteils 26,9% (s. S. 49, lfd.Nr.1).
Der Erwartungswert beträgt bei einem alten Fahrer (A3) 15,8%, bei einem Fahrer der Altersgruppe A2 (unter 25 Jahre) 45,4% (unter der Bedingung, daß für alle übrigen Variablen die Basiskategorie gilt), lfd.Nr.25 bzw. 13.
Liegt die Angabe "mit Alkohol" (AL2) beim Fahrer vor,(bei den übrigen im Modell vorhandenen Variablen gilt wiederum die Basiskategorie), so steigt der Erwartungswert des Fahrunfallanteils auf 55,8%. Unter Beibehaltung von "mit Alkohol" (AL2) schnellt der Erwartungswert des Fahrunfallanteils in der Altersgruppe der unter 25 Jahre alten männlichen Fahrer auf 74,0%. Betrachtet man jedoch nicht Fahrunfälle auf Landesstraßen (Basiskategorie) sondern auf Gemeinde- und Kreisstraßen, so steigt der Fahrunfallanteil bei

jungen alkoholisierten Fahrern sogar auf 77,8%; d.h. mehr als drei Viertel aller unter Alkoholeinfluß stehenden PKW-Fahrer im Alter unter 25 Jahren, die auf Gemeinde- und Kreisstraßen einen Unfall verursachen, sind in einen Fahrunfall verwickelt.
Bei den weiblichen Fahrern im Alter von 25 bis unter 60 Jahren liegt der Erwartungswert mit 29,1% um 2,3% höher als bei den männlichen Fahrern gleichen Alters. Betrachtet man im Vergleich die Gruppe der jungen Fahrer (A2), läßt sich bei Frauen mit 39,6% ein um 5,7% niedrigerer Erwartungswert als bei den jungen Männern beobachten; deren Wert liegt bei 45,4%.

Die Beispiele zeigen, daß man für jede in die multidimensionale Analyse mittels Logit-Modellen einbezogene Merkmalskombination einen Erwartungswert errechnen kann. Dabei gilt aber, daß die Ermittlung eines Erwartungswertes für Merkmalskombinationen, die strukturelle Nullen enthalten, unzulässig ist.

Die Erwartungswerte weichen von den Prozentsätzen einer mehrdimensionalen Kontingenztafel deshalb ab, weil alle im optimalen Modell enthaltenen, sowie bei der Modellentwicklung ausgeschlossenen Effekte bei der Berechnung eines Erwartungswertes implizit berücksichtigt sind, wobei aber nur die statistisch signifikanten Parameter in die Modellrechnung eingehen.

Die in das Modell einbezogenen Parameter erklären 99,5% der Devianz des Basismodells (PEDAD = 99,5%; vgl. Anhang 6.6). Das multiple Bestimmungsmaß PED beträgt für dieses Modell jedoch nur 6.2%. Es bleibt also noch eine Restvariation von 93,8%, die auf andere als in das Modell explizit einbezogene Variablen (Einflüsse) zurückzuführen ist.
Obwohl PEDAD (s. Abschnitt 2.6) bei fast 100% liegt, wird die Gesamtvariation der Individualdaten nur zu 6,2% erklärt. Es ist also angeraten, weitere Variablen in die Analyse einzubeziehen, wenn man die Realität hinreichend abbilden will; dies ist an anderer Stelle beschrieben (s. Brühning u. Ernst in DVWG, 1987).

LITERATUR

Arminger,G.; Küsters,U.; 1986
Statistische Verfahren zur Analyse qualitativer Variablen.
Bericht zum FP 8302/3 der Bundesanstalt für Straßenwesen
Bergisch Gladbach, 1986

Brühning,E.; Ernst,G; 1985
Loglinear Models in Effectivness Studies -
An Application to "Simultaneous Before-After-Comparisons with Control Groups.
In: Evaluation 85 International Meeting on the Evaluation of Local Traffic Safety Measures,Tome 2, S. 291-311, hrsg. von der ONSER, Paris, 1985

Brühning,E.; Dilling,J.; Ernst,G.; Schmid,M.; 1986
Auswertung zu "Fahrunfällen auf Landstraßen" im Rahmen der Unfallforschung.
In: Zusammenführung und Auswertung von Fahrzeug- und Unfalldaten, BASt/KBA 1986 (Veröffentlichung in Vorbereitung)

DVWG; 1987
Multivariate Analyse mittels loglinearer Modelle -
Ein Analyseinstrument für die Verkehrs- und Unfallforschung.
Schriftenreihe der Deutschen Verkehrswissenschaftlichen Gesellschaft DVWG, Reihe B, 1987

Brühning,E.; Ernst G.; 1987
Kontingenztafelanalyse mittels loglinearer und Logit-Modelle im Bereich der Verkehrsforschung.
In: Verkehrsstatistik heute: Theorie und Praxis vor neuen Aufgaben, Schriftenreihe der Deutschen Verkehrswissenschaftlichen Gesellschaft DVWG, Reihe B; 1987

The GLIM-System Manual, Release 3.77.
Numerical Algorithmus Group, NAG Central Office, Mayfield House, 256 Banbury Road, Oxford OX2 7DE; U. K.

ANHANG 1: SPSS-JOB

```
GET FILE        BET
SELECT IF       (FZAGRB80 EQ 6)
RECODE          UNF1UR, UNF2UR, UNF3UR (0,2 THRU 69=1) (1=2)
VALUE LABELS    UNF1UR, UNF2UR, UNF3UR (1) KEIN ALKOHOL
                                       (2) ALKOHOL
IF              (UNF1UR EQ 1 OR UNF2UR EQ 1 OR  UNF3UR EQ 1)
                UNFALL=1
IF              (UNF1UR EQ 2 OR UNF2UR EQ 2 OR  UNF3UR EQ 2)
                UNFALL=2
VALUE LABELS    UNFALL (1) KEIN ALKOHOLUNFALL
                       (2) ALKOHOLUNFALL
RECODE          UNFTYP (0,2 THRU 7=2) (1=1)
VALUE LABELS    UNFTYP (1) FAHRUNFALL
                       (2) KEIN FAHRUNFALL
RECODE          ALTBET (BLANK=101)
MISSING VALUE   ALTBET (101)
RECODE          ALTBET (0   THRU 24=2)
                       (25  THRU 59=1)
                       (60  THRU 99=3)
                       (100=4)
VALUE LABELS    ALTBET   (2) 0  BIS 24 JAHRE
                         (1) 25 BIS 59 JAHRE
                         (3) 60 BIS 99 JAHRE
                         (4) OHNE ANGABE
                         (3) 8 JAHRE U.MEHR
RECODE          STRKLASS (3=1)(4,5=3)
RAW OUTPUT UNIT 15
CROSSTABS       VARIABLES=ALTBET    (1,3)
                          GESCHL(1,2)
                          UNFALL(1,2)
                          STRKLASS(1,3)
                          UNFTYP  (1,2)/
                TABLES=STRKLASS BY UNFALL BY GESCHL
                BY ALTBET BY UNFTYP
OPTIONS         11
FINISH
```

ANHANG 2: SPSS-ENTER-DATEI

```
/.ERNST LOGON U401BRUE,0016
/OPTION MSG=FHL
/SYSFILE SYSLST=AUS.ERNST.500
/DO PROC.SPSS9,(CONTROL=WERKBUCH,GTFILE=$U415UFZ1.KBA.BET3,
/OUTPUTA=WERKBUCH.100)
/SYSFILE SYSLST=( )
/ER AUS.ERNST.500
/LOGOFF
```

ANHANG 3: NACH SPSS-AUSWERTUNG ENTSTANDENE ERGEBNISDATEI

00000001	1	1	2245	1	1	1	1	1
00000002	1	1	1458	2	1	1	1	1
00000003	1	1	1513	3	1	1	1	1
00000004	1	1	892	1	2	1	1	1
00000005	1	1	520	2	2	1	1	1
00000006	1	1	628	3	2	1	1	1
00000007	1	1	939	1	1	2	1	1
00000008	1	1	624	2	1	2	1	1
00000009	1	1	631	3	1	2	1	1
00000010	1	1	78	1	2	2	1	1
00000011	1	1	51	2	2	2	1	1
00000012	1	1	65	3	2	2	1	1
00000013	1	1	3290	1	1	1	2	1
00000014	1	1	1655	2	1	1	2	1
00000015	1	1	2416	3	1	1	2	1
00000016	1	1	793	1	2	1	2	1
00000017	1	1	434	2	2	1	2	1
00000018	1	1	791	3	2	1	2	1
00000019	1	1	1005	1	1	2	2	1
00000020	1	1	523	2	1	2	2	1
00000021	1	1	644	3	1	2	2	1
00000022	1	1	51	1	2	2	2	1
00000023	1	1	32	2	2	2	2	1
00000024	1	1	47	3	2	2	2	1
00000025	1	1	243	1	1	1	3	1
00000026	1	1	168	2	1	1	3	1
00000027	1	1	120	3	1	1	3	1
00000028	1	1	24	1	2	1	3	1
00000029	1	1	11	2	2	1	3	1
00000030	1	1	12	3	2	1	3	1
00000031	1	1	44	1	1	2	3	1
00000032	1	1	20	2	1	2	3	1
00000033	1	1	21	3	1	2	3	1
00000034	1	1	2	1	2	2	3	1
00000035	1	1	1	2	2	2	3	1
00000036	1	1	1	3	2	2	3	1
00000037	1	1	6026	1	1	1	1	2
00000038	1	1	5783	2	1	1	1	2
00000039	1	1	3738	3	1	1	1	2
00000040	1	1	718	1	2	1	1	2
00000041	1	1	609	2	2	1	1	2
00000042	1	1	415	3	2	1	1	2
00000043	1	1	2323	1	1	2	1	2
00000044	1	1	1996	2	1	2	1	2
00000045	1	1	1460	3	1	2	1	2
00000046	1	1	65	1	2	2	1	2
00000047	1	1	49	2	2	2	1	2
00000048	1	1	34	3	2	2	1	2
00000049	1	1	3964	1	1	1	2	2
00000050	1	1	3466	2	1	1	2	2
00000051	1	1	2681	3	1	1	2	2
00000052	1	1	278	1	2	1	2	2
00000053	1	1	236	2	2	1	2	2
00000054	1	1	220	3	2	1	2	2

00000055	1	1	1538	1	1	2	2	2
00000056	1	1	1383	2	1	2	2	2
00000057	1	1	906	3	1	2	2	2
00000058	1	1	27	1	2	2	2	2
00000059	1	1	17	2	2	2	2	2
00000060	1	1	14	3	2	2	2	2
00000061	1	1	1268	1	1	1	3	2
00000062	1	1	1194	2	1	1	3	2
00000063	1	1	593	3	1	1	3	2
00000064	1	1	25	1	2	1	3	2
00000065	1	1	34	2	2	1	3	2
00000066	1	1	25	3	2	1	3	2
00000067	1	1	224	1	1	2	3	2
00000068	1	1	206	2	1	2	3	2
00000069	1	1	89	3	1	2	3	2
00000070	1	1	0	1	2	2	3	2
00000071	1	1	1	2	2	2	3	2
00000072	1	1	0	3	2	2	3	2

ANHANG 4: "TRANS.SPSS"

PROGRAMM TRANS.SPSS

```
      PROGRAM TRANS.SPSS
      CHARACTER CFILE*30
      DIMENSION J(20)
      WRITE(2,11)
11    FORMAT('EINGABEDATEI IN HOCHKOMMATA ?')
      READ(1,*)CFILE
      OPEN(10,FILE=CFILE,STATUS='OLD')
      WRITE(2,22)
22    FORMAT('AUSGABEDATEI IN HOCHKOMMATA ?')
      READ(1,*)CFILE
      OPEN(11,FILE=CFILE)
      WRITE(2,33)
33    FORMAT('ANZAHL DER VARIABLEN ?')
      READ(1,*)I
      N=0
10    N=N+1
      READ(10,*,END=100)(J(I1),I1=1,I+3)
      WRITE(11,'(100I10)')(J(I1),I1=I1+3,4-1),J(3)
      GOTO 10
100   N=N-1
      WRITE(2,55)N
55    FORMAT(5X,I5, '       SAETZE GELESEN')
      STOP
      END
```

AUSGABEDATEI AUS "TRANS.SPSS"

00000001	1	1	1	1	1	2245
00000002	1	1	1	1	2	1458
00000003	1	1	1	1	3	1513
00000004	1	1	1	2	1	892
00000005	1	1	1	2	2	520
00000006	1	1	1	2	3	628
00000007	1	1	2	1	1	939
00000008	1	1	2	1	2	624
00000009	1	1	2	1	3	631
00000010	1	1	2	2	1	78
00000011	1	1	2	2	2	51
00000012	1	1	2	2	3	65
00000013	1	2	1	1	1	3290
00000014	1	2	1	1	2	1655
00000015	1	2	1	1	3	2416
00000016	1	2	1	2	1	793
00000017	1	2	1	2	2	434
00000018	1	2	1	2	3	791
00000019	1	2	2	1	1	1005
00000020	1	2	2	1	2	523
00000021	1	2	2	1	3	644
00000022	1	2	2	2	1	51
00000023	1	2	2	2	2	32
00000024	1	2	2	2	3	47
00000025	1	3	1	1	1	243
00000026	1	3	1	1	2	168
00000027	1	3	1	1	3	120
00000028	1	3	1	2	1	24
00000029	1	3	1	2	2	11
00000030	1	3	1	2	3	12
00000031	1	3	2	1	1	44
00000032	1	3	2	1	2	20
00000033	1	3	2	1	3	21
00000034	1	3	2	2	1	2
00000035	1	3	2	2	2	1
00000036	1	3	2	2	3	1
00000037	2	1	1	1	1	6026
00000038	2	1	1	1	2	5783
00000039	2	1	1	1	3	3738
00000040	2	1	1	2	1	718
00000041	2	1	1	2	2	609
00000042	2	1	1	2	3	415
00000043	2	1	2	1	1	2323
00000044	2	1	2	1	2	1996
00000045	2	1	2	1	3	1460
00000046	2	1	2	2	1	65
00000047	2	1	2	2	2	49
00000048	2	1	2	2	3	34
00000049	2	2	1	1	1	3964
00000050	2	2	1	1	2	3466
00000051	2	2	1	1	3	2681
00000052	2	2	1	2	1	278
00000053	2	2	1	2	2	236
00000054	2	2	1	2	3	220

00000055	2	2	2	1	1	1538
00000056	2	2	2	1	2	1383
00000057	2	2	2	1	3	906
00000058	2	2	2	2	1	27
00000059	2	2	2	2	2	17
00000060	2	2	2	2	3	14
00000061	2	3	1	1	1	1268
00000062	2	3	1	1	2	1194
00000063	2	3	1	1	3	593
00000064	2	3	1	2	1	25
00000065	2	3	1	2	2	34
00000066	2	3	1	2	3	25
00000067	2	3	2	1	1	224
00000068	2	3	2	1	2	206
00000069	2	3	2	1	3	89
00000070	2	3	2	2	1	0
00000071	2	3	2	2	2	1
00000072	2	3	2	2	3	0

ANHANG 5: "TRANS.SPSS.2"

PROGRAMM TRANS.SPSS.2

```
      PROGRAM TRANS.SPSS.2
      CHARACTER IFILE*30,AFILE*30,FMT*30
      INTEGER ANZAHL,SPALTE,ERG,MERK(10000)
      DIMENSION IDUM(50)
      I1=0
      I2=0
      WRITE(2,11)
11    FORMAT('EINGABEDATEI IN HOCHKOMMATA ?')
      READ(1,*)IFILE
      OPEN(10,FILE=IFILE)
      WRITE(2,22)
22    FORMAT('AUSGABEDATEI IN HOCHKOMMATA ?')
      READ(1,*)AFILE
      OPEN(11,FILE=AFILE)
      WRITE(2,33)
33    FORMAT('WIEVIELE SPALTEN INSGESAMT ?')
      READ(1,*)ANZAHL
      FMT(1:1)='('
      FMT(4:7)='I10)'
      WRITE(FMT(2:3),'(I2)')ANZAHL
10    I1=I1+1
      READ(10,FMT,END=100)(IDUM(I),I=1,ANZAHL)
      IF(IDUM(1).EQ.2)THEN
      I2=I2+1
      MERK(I2)=IDUM(ANZAHL)
      ENDIF
      GOTO 10
100   REWIND 10
      I=0
20    I=I+1
      READ(10,FMT)(IDUM(I1),I1=1,ANZAHL)
      IF(IDUM(1).EQ.2)GOTO 200
      ERG=IDUM(ANZAHL)+MERK(I)
      WRITE(11,FMT)(IDUM(I1),I1=2,ANZAHL),ERG
      GOTO 20
200   I=I-1
      WRITE(2,55)I
55    FORMAT(I10,'ZEILEN GESCHRIEBEN')
      STOP
      END
```

AUSGABEDATEI AUS "TRANS.SPSS.2"

00000001	1	1	1	1	2245	8271
00000002	1	1	1	2	1458	7241
00000003	1	1	1	3	1513	5251
00000004	1	1	2	1	892	1610
00000005	1	1	2	2	520	1129
00000006	1	1	2	3	628	1043
00000007	1	2	1	1	939	3262
00000008	1	2	1	2	624	2620
00000009	1	2	1	3	631	2091
00000010	1	2	2	1	78	143
00000011	1	2	2	2	51	100
00000012	1	2	2	3	65	99
00000013	2	1	1	1	3290	7254
00000014	2	1	1	2	1655	5121
00000015	2	1	1	3	2416	5097
00000016	2	1	2	1	793	1071
00000017	2	1	2	2	434	670
00000018	2	1	2	3	791	1011
00000019	2	2	1	1	1005	2543
00000020	2	2	1	2	523	1906
00000021	2	2	1	3	644	1550
00000022	2	2	2	1	51	78
00000023	2	2	2	2	32	49
00000024	2	2	2	3	47	61
00000025	3	1	1	1	243	1511
00000026	3	1	1	2	168	1362
00000027	3	1	1	3	120	713
00000028	3	1	2	1	24	49
00000029	3	1	2	2	11	45
00000030	3	1	2	3	12	37
00000031	3	2	1	1	44	268
00000032	3	2	1	2	20	226
00000033	3	2	1	3	21	110
00000034	3	2	2	1	2	2
00000035	3	2	2	2	1	2
00000036	3	2	2	3	1	1

ANHANG 6: METHODISCHE ANMERKUNGEN ZU LOGLINEAREN UND LOGIT-MODELLEN

(Auszug aus: Brühning, Ernst (1987), ergänzt um Abschnitt 4 (teilweise) und um Abschnitt 5)

1. PRINZIP DES LOGLINEAREN MODELLS

Das Prinzip loglinearer Modelle soll zunächst am einfachen Beispiel einer 2 x 2 Kontingenztafel demonstriert werden.

Bild 1: Schema einer 2 x 2 Kontigenztafel

		B	
		1	2
A	1	y_{11}	y_{12}
	2	y_{21}	y_{22}

Das loglineare Modell für das in Bild 1 gezeigte Beispiel lautet:

$$\eta_{ij} = \ln\mu_{ij} = \beta_0 + \beta^A_i + \beta^B_j + \beta^{AB}_{ij} \qquad i,j = 1,2$$

Der Logarithmus des jeweiligen erwarteten Zellenwertes in der Vierfeldertafel y_{ij} setzt sich additiv aus einem allgemeinen Effekt β_0 (vielfach auch "Great Mean" genannt), den Haupteffekten β^A_i und β^B_j und einem Interaktionseffekt β^{AB}_{ij} zusammen.

Die Feldbesetzungen der Kontingenztafel werden hierdurch auf die zugrunde liegenden Variableneffekte zurückgeführt. Das statistische Modell geht von einer multiplikativen Verknüpfung der Einzelfaktoren aus, es läßt sich jedoch durch logarithmieren in ein additives, sog. loglineares Modell überführen. Die Annahme der multiplikativen Verknüpfung der Variableneffekte ist statistisch begründet. Sie muß aber auch ebensogut sachlich begründet

werden können; so wird sich z.B. in Abhängigkeit vom Lebensalter die Unfallhäufigkeit um einen bestimmten Prozentsatz verändern und nicht um eine konstante Anzahl, was einer Multiplikation mit einem entsprechenden Faktor gleichkommt.

Die Effekte $\beta_0, \ldots, \beta^{AB}_{ij}$ lassen sich sehr anschaulich deuten. Sind die Zellenwerte y_{ij} in der Vierfeldertafel gleich groß (Fall der Gleichverteilung), kommt dies in dem allgemein wirksamen Effekt β_0 zum Ausdruck. Die Effekte β^A_i und β^B_j treten mit der Abweichung von der Gleichverteilung auf. Unterschiede zwischen den Zeilen der Vierfeldertafel werden auf den Einfluß der Variablen A, Unterschiede zwischen den Spalten auf den Einfluß der Variablen B zurückgeführt. Sind die beiden Variablen A und B stochastisch abhängig, so tritt zusätzlich zu den Haupteffekten β^A_i und β^B_j ein Interaktionseffekt β^{AB}_{ij} auf, der das Zusammenwirken der beiden Variablen A und B zum Ausdruck bringt. Sind die beiden Variablen stochastisch unabhängig, so ist $\beta^{AB}_{ij} = 0$.

Bei einem Modell, das genau so viele Effekte (Parameter) enthält, wie die Kontingenztafel Zellen hat, spricht man von einem saturierten Modell.

2. DAS PRINZIP DES LOGIT-MODELLS

Mit Abschnitt 1 wurde das Prinzip des loglinearen Modells verdeutlicht. Dabei werden die absoluten Häufigkeiten in der Kontingenztafel als unabhängige Variable aufgefaßt.

Um Logit-Modelle handelt es sich, wenn nicht absolute Häufigkeiten, sondern der Verhältniswert der Ausprägungen einer z.B. dichotomen Variablen abgebildet werden soll.

Bild 2: Schema einer 3 x 2 x 2 Kontingenztafel

		C_1	C_2
A_1	B_1	y_{111}	y_{112}
	B_2	y_{121}	y_{122}
A_2	B_1	y_{211}	y_{212}
	B_2	y_{221}	y_{222}
A_3	B_1	y_{311}	y_{312}
	B_2	y_{321}	y_{322}

Bei der in Bild 2 dargestellten dreidimensionalen Tabelle soll die Ausprägung von C in Abhängigkeit von den Variablen A und B behandelt werden. Dabei wird die Kategorie C_1 im Vergleich mit der Kategorie C_2 betrachtet.
Dies geschieht, indem für jedes $A_i B_j$ der Wert

$$\eta_{ij} = \ln \frac{\mu_{ij}}{1 - \mu_{ij}} \quad \text{ermittelt wird,}$$

wobei $$\mu_{ij} = \frac{\mu_{ij1}}{\mu_{ij1}+\mu_{ij2}} = \frac{\mu_{ij1}}{N_{ij}}$$

der Erwartungswert von $y_{ij} = \frac{y_{ij1}}{N_{ij}}$ ist.

Der lineare Prädiktor η_{ij} wird als Logit bezeichnet. Das Logit ist einfach interpretierbar als Logarithmus des relativen Verhältnisses der Wahrscheinlichkeit von $C = C_1$ zur Wahrscheinlichkeit $C = C_2$.
Wird η_{ij} als Zellenwert geschätzt, so entsteht ein lineares Modell für die Differenzen der Logarithmen der Wahrscheinlichkeiten

$$\ln \mu_{ij} - \ln (1 - \mu_{ij}) = \beta_0 + \beta^{A}{}_i + \beta^{B}{}_j + \beta^{AB}{}_{ij} \qquad \begin{matrix} i = 1,\ldots,3 \\ j = 1,2 \end{matrix}$$

Ein Modell ist somit definiert durch:

1. Eine beobachtete abhängige Variable (hier: C), beim loglinearen Modell die absoluten Zellenhäufigkeiten.
2. Ein lineares Modell, gebildet aus den unabhängigen (erklärenden) Variablen mit denen der Vektor η geschätzt wird.
3. Die Wahrscheinlichkeitsverteilung der C-Variablen
4. Die Link-Funktion (z.B. Logit), die den linearen Prädiktor η mit dem Erwartungswert μ verknüpft.

Mit der Link-Funktion wird somit der Modelltyp (z.B. loglinear oder Logit) festgelegt.

3.HIERARCHISCHE, NICHTHIERARCHISCHE MODELLE

Hierarchische Modelle unterscheiden sich von nichthierarchischen dadurch, daß Interaktionseffekte höherer Ordnung in einem Modell immer zugleich mit den zugehörigen Interaktionseffekten niedriger Ordnung sowie den Haupteffekten auftreten. Arbeitet man nur mit hierarchischen Modellen, so ergibt sich mit zunehmender Zahl der Variablen und deren Kategorien sehr schnell eine lange Liste von Effekten, von den meist ein Großteil nicht signifikant von 0 verschieden ist und aus dem Modell gestrichen werden könnte. Dies ist aber nicht möglich, wenn ein Effekt höherer Ordnung signifikant von 0 verschieden ist und im Modell enthalten sein soll. Aus diesem Grund sind hierarchische Modelle zur Analyse meist weniger geeignet. Bei der Softwareauswahl ist es daher wichtig, mit Verfahren zu arbeiten, die nichthierarchische Modelle schätzen können.

4. PROBLEME DER STRUKTURELLEN NULLEN UND STICHPROBENNULLEN

Bei der Analyse von Kontingenztafeln kann der Fall eintreten, daß eine bestimmte Ausprägung der abhängigen Variablen innerhalb einer gegebenen Konstellation von unabhängigen Variablen nicht auftreten kann, d.h. es gibt in der Kontingenztafel gewisse Merkmalsausprägungen der unabhängigen Variablen, die logisch unvereinbar sind und deshalb mit der Häufigkeit "0" auftreten müssen. In diesem Fall spricht man von strukturellen Nullen. Beispiele für strukturelle Nullen lassen sich in der Verkehrswissenschaft leicht finden: So gibt es keine Fahrzeuge mit hohem Leistungsgewicht und zugleich niedriger Motorleistung; Kleinkinder treten nicht als Führerscheinbesitzer auf, usw.

Im Gegensatz zu strukturellen Nullen sind Stichprobennullen definiert als zufallsbedingt fehlende Besetzung einer Ausprägung, d.h. es tritt z.B. keinerlei Unfall im empirischen Befund auf, obwohl dieses durchaus hätte erwartet werden können. In einer Untersuchung ergab sich z.B. der Fall, daß alte alkoholisierte Frauen mit hochmotorisierten Fahrzeugen in der Kontingenztafel fehlten. Es ist wichtig, daß strukturelle bzw. Stichprobennullen adäquat behandelt werden. Strukturelle Nullen müssen bei der Berechnung von Freiheitsgraden sowie Interaktionseffekten durch Streichen der Parameter in der Designmatrix berücksichtigt werden können. Dieses ist z.B. beim Programmpaket GLIM sehr einfach möglich (s. folgendes Beispiel). Andere Standardsoftware wie NONMET oder ECTA ersetzen strukturelle Nullen entweder durch kleine Konstante oder durch Gewichte, was zu verzerrten Schätzern führt (näheres s. Arminger, 1986). Stichprobennullen können nur durch Maximum Likelihood (ML) Schätzungen richtig behandelt werden. Dies führt bei bestimmten Softwarelösungen zu Problemen.

Beispiel zur Behandlung struktureller Nullen mittels GLIM:
Im GLIM hat der Benutzer zwei Möglichkeiten der korrekten Behandlung von Kovariatenkostellationen mit strukturellen Nullen:

1. Die Kovariatenkonstellationen können bei der interaktiven Dateneingabe in GLIM entfallen.
2. Erfolgt die Dateneingabe nach GLIM mit $DINPUT aus "TRANS.SPSS" oder "TRANS.SPSS.2" (beide Programme enthalten auch Kovariatenkonstellationen mit strukturellen Nullen) oder ähnlichen

Dienstprogrammen, die Kovariatenkonstellationen mit strukturellen Nullen übergeben, so ist mit Hilfe eines $CAL-Statements und einer anschließenden Gewichtung der Ausschluß von Kovariatenkonstellationen mit strukturellen Nullen möglich.

```
$CAL W=%IF(%EQ(N,0),1,0)
$WEIGHT W
```

N = Beobachtungsvektor der abhängigen zu gewichtenden Variablen
W = Gewichtungsvektor (vgl. Anhang 7)

Beide der oben beschriebenen Vorgehensweisen führen zu einer numerischen Elimination der in GLIM automatisch aufgebauten, aber redundanten Interaktionsparameter zwischen erklärenden Variablen der i,j-ten Kovariatenkonstellation und der fehlenden Merkmalsausprägung der abhängigen Variablen. Die entsprechenden Parameter werden in GLIM als ALIASED (extrinsically) ausgewiesen. Die Anzahl der Freiheitsgrade eines Modells wird durch das weglassen (ausschließen) von Zeilen automatisch nach unten korrigiert.
Als Hinweis sei noch vermerkt, daß strukturelle Nullen nicht in der Basiskategorie enthalten sein sollten.

5. ANMERKUNGEN ZUR BASISKATEGORIE

Die erste Ausprägung jeder unabhängigen Variablen geht als Basiskategorie in die Berechnung ein. Aus diesem Grund ist es zumeist günstig, die erste Kategorie so zu wählen, daß sie auf eine Kategorie mit großer Besetzungszahl fällt.

Dies soll am Beispiel "Alter des Beteiligten" verdeutlicht werden:

Alter des Beteiligten	Häufigkeit		
unter 25 Jahre	25.895	(A2)	
25 bis unter 60 Jahre	32.206	(A1)	Basiskategorie
60 Jahre und älter	4.259	(A3)	

Die übrigen Kategorien der Variablen werden zu der Basiskategorie ins Verhältnis gesetzt; d.h. die Parameterschätzung für die Basiskategorie (erste Ausprägung jeder unabhängigen Variablen) ist für jede Variable gleich 0, da GLIM mit einer 0,1 kodierten Design-

matrix (gecornerten Designmatrix) arbeitet.
Für das Beispiel "Alter des Beteiligten", "Geschlecht" ergibt sich

$$x_{n,6} = \begin{bmatrix} \beta_0 & A2 & A3 & G2 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 \end{bmatrix}$$

Mit dieser Designmatrix ist das Modell $\eta = f(\beta_0, A, G)$ spezifiziert (Haupteffektmodell).

6. BEURTEILUNG DER MODELLANPASSUNG

Bei jeder Modellentwicklung stellt sich die Frage, in welchem Maße die im Modell bzw. in der Kontingenztafel enthaltenen Variablen die Variation des empirischen Befundes erklären. Dies ist bei Logit-Modellen analog zur Varianzanalyse leicht möglich.

Grundsätzlich ist zwischen zwei Maßen zu unterscheiden:
Zum einen kann der Anteil der durch ein Modell erklärten Devianz innerhalb der in der Kontingenztafel enthaltenen Aggregatdaten bestimmt werden. (PEDAD = Proportion of Explained Deviance on Aggregate Data). Dies ist der übliche Weg zur Beurteilung der Modellanpassung an den empirischen Befund.

Häufig wird aber übersehen, daß mit der Aggregation ein Informationsverlust verbunden ist. Deshalb ist es erforderlich, neben dem genannten Maß PEDAD auch zu untersuchen, welcher Anteil der Variation der in der Tabelle aggregierten Individualdaten durch die in der Kontingenztafel enthaltenen unabhängigen Variablen und alle ihre Interaktionen erklärt wird (PED = Proportion of Explained Deviance). Erst dieses Maß ist mit dem aus der Regressionsrechnung bzw. Korrelationsrechnung bekannten R^2 zu vergleichen.

Im Einzelnen gilt:

Devianz des Modells (allgemein)

Der Wert der Devianz nimmt für loglineare bzw. Logit-Modelle den Wert G^2(Bishop et al.,1975) an:

$$D(c,f) = G^2 = \sum_{i=1}^{n} [R_i \ln(R_i/\hat{R}_i) + (N_i - R_i) \ln((N_i - R_i)/(N_i - \hat{R}_i))]$$

G^2 ist - falls die H_0: die geschätzten Werte $\hat{R}_i$ unterscheiden sich nur zufällig von den beobachteten R_i - asymptotisch X^2-verteilt mit DF Freiheitsgraden

DF = n - p
n = Zellenzahl der Kontingenztafel
p = Anzahl der geschätzten Parameter

Devianz des Basismodells der Kontingenztafel

Für ein Basismodell M_{c0}, in dem nur der "Great Mean" β_0 geschätzt wird, ergibt sich gem. obiger Formel die Devianz $D(c_0,f)$,die von GLIM ausgedruckt wird.

Devianz eines erweiterten Modells

Für ein gegenüber M_{c0} um z.B. einige Haupteffekte erweitertes Modell M_{c1} ergibt sich entsprechend $D(c_1,f)$.
$D(c_0,c_1) = D(c_0,f) - D(c_1,f)$ gibt die Devianz an, die durch die Parameter, die in M_{c1},aber nicht in M_{c0} enthalten sind, erklärt wird.

Multiples, partielles Bestimmungsmaß PEDAD

Als multiples partielles Bestimmungsmaß läßt sich der Prozentsatz an erklärter Devianz in aggregierten Daten ermitteln:

$$PEDAD = \frac{D(c_0,f) - D(c_1,f)}{D(c_0,f)} = \frac{D(c_0,c_1)}{D(c_0,f)}$$

Das PEDAD-Maß gibt an, um welchen Prozentsatz die Devianz im Modell $M_{c\,0}$ durch Einfügung der zusätzlichen Parameter (Variablen) in $M_{c\,1}$ verringert wurden.
PEDAD bezieht sich aber nur auf die aggegierten Daten und berücksichtigt nicht den Informationsverlust, der durch die Aggregation entsteht (vgl. das multiple Bestimmungsmaß PED).

Gesamtdevianz

Analysiert werden N Unfälle, die nach ihrer Eigenschaft als Fahrunfälle/übrige Unfälle (mit den Anteilswerten p_1/p_2) binomial verteilt sind. Die Gesamtdevianz D_g ergibt sich zu (vgl. Arminger 1986):

$$D_g = -2N\,(p_1 \ln p_1 + p_2 \ln p_2)$$

$$p_1 = \frac{p_1}{N}$$

$$p_2 = 1 - p_1$$

Multiples Bestimmungsmaß

Das zu R^2 der Regressionsrechnung analoge Bestimmungsmaß PED wird wie folgt ermittelt:

$$PED = \frac{D(c_0,f) - D(c_1,f)}{D_g}$$

Die beiden Bestimmungsmaße PEDAD und PED liegen in der Regel weit auseinander. PEDAD liegt immer deutlich über PED. Auch wenn die Variation innerhalb einer Kontingenztafel zu 95 % und mehr erklärt wird, kann die Variation der Individualdaten nur zu wenigen Prozent erklärt sein, d.h. die Restvariation ist auf andere als die

im Modell explizit einbezogenen Variablen (Einflüsse) zurückzuführen.

ANHANG 7: GLIM - KOMMANDOÜBERSICHT

In diesem Abschnitt wird eine Übersicht über die GLIM Kommandos gegeben. Dabei handelt es sich häufig um eine Kurzfassung der Erläuterungen; für die ausführliche Beschreibung wird auf das Manual verwiesen.

GLIM-Befehle:

Nur die ersten vier Zeichen eines Befehls werden interpretiert. Sie können in vielen Fällen weiter abgekürzt werden (fett gedruckter Teil des jeweiligen Befehls). Befehle beginnen mit einem Befehlszeichen. Befehle und Befehlsspezifikationen müssen durch mindestens ein Leerzeichen oder durch Zeilenschaltung voneinander getrennt sein. $SUBFILE oder $FINISH - falls sie verwendet werden - müssen die ersten Befehle einer Befehlszeile sein. Der Text, der auf die Befehle $FORMAT, $END oder $FINISH folgt, wird ignoriert.

Befehle:

$ACCURACY [Integer]
Anzahl der Stellen bei der Druckausgabe von Zahlen im G-Format
Voreinstellung: 4
zulässig: positive Zahlen

$ALIAS ---
Umschalten auf Ausschluß oder Einschluß von intrinsically aliased Parameters (die erste Ausprägung jeder Variablen ist Basiskategorie und wird auf 0 gesetzt; vgl. Anhang 6, Abschnitt 5); Ja/Nein.
Voreinstellung: Ausschluß

$ARGUMENT macro argument
Setzen von maximal 9 Argumenten für ein definiertes Makro. Die Argumente brauchen erst beim ersten Aufruf des Makros definiert zu sein.
Es dürfen mehr Argumente definiert als benutzt werden.

$CALCULATE Ausdruck
Berechnung von arithmetischen oder logischen Ausdrücken. Die Ergebnisse werden Zieloperanden zugewiesen, die als Vektor implizit definiert werden können.

$COMMENT Kommentar

$CYCLE [Integer1 [Integer2]]
Integer1 = Anzahl maximaler Iterationsschritte (Voreinstellung: 10)
Integer2 = jeder zweite Iterationsschritt und der letzte Iterationsschritt werden ausgedruckt (Voreinstellung: nur der letzte Iterationsschritt wird ausgedruckt)

$DATA [Integer] Name
Variablenliste für $READ oder $DINPUT
falls die Längenangabe "Integer" fehlt, erhalten noch undefinierte Vektoren (Variablen) die Länge des ersten Vektors (Variable) und ggf. der noch nicht definierte erste Vektor die Standardlänge.

$DELETE [Name]
Löschung von definierten Datenstrukturen mit "Namen".

$DINPUT Integer1 [Integer2]
Lesen von Zahlen (Daten) von der Kanalnummer "Inter1" mit der maximalen Satzlänge "Integer2" (32 bis 299) gemäß der Definition durch $DATA.

$DISPLAY Buchstaben
zeigt die Ergebnisse des letzten FITs
Buchstaben:
A: wie E, incl. intrinsically aliased Parameter (vgl. $ALIAS)
C: Korrelation der Parameterschätzwerte
D: Anpassungsmaß (Devianz) und Freiheitsgrade

E: Parameterschätzwerte, ihre Standardfehler und Definition, incl. extrinsically aliased Parameter (vgl. Anhang 6, Abschnitt 4)
L: Formel für den linearen Prädiktor
M: alle Modell - Spezifikationen
R: die Y-Variable, ihre Schätzwerte,generalisierte Residuen
S: Standardfehler der Differenzen geschätzter Parameter
T: generalisierte Inverse der SSP-Matrix (SUM-of-SQUARE-and-Product-Matrix)
U: wie E; excl. extrinsically alaised Parameter
V: Covarianz - Matrix der Parameterschätzwerte
W: wie R, aber abhängig von der Maske %RE (u.s.)

$DUMP [Integer]
speichert aktuellen Programmstatus auf dem Kanal mit der Kanalnummer "Integer".

$ECHO ---
Umschalten auf Protokollierung bzw. Nichtprotokollierung der Eingabe.
Voreinstellung: interaktiv = Nichtprotokollierung
batch = Protokollierung

$END ---
Ende des GLIM - Jobs

$ENDMAC ---
Ende des Makros

$ENVIRONMENT Buchstaben
Informationen über den aktuellen Zustand des Programms
Buchstaben:
C: Channel
Kanäle für Eingabe, Ausgabe und Dump
D: Directory
benutzerdefinierte Identifikatoren und benutzte Systemvektoren sowie deren Speicherplätze

I: Implementation
implementationsabhängige Eigenschaften
P: Program Control Stack
Programmebene
R: Pseudo - Random - Number
Startwerte für den Standardzufallszahlengenerator
S: System
vom GLIM - System benutzter Speicherplatz
U: Usage
Speicherplatz der Daten, Identifikatoren, Vektoren, Modellterme, Programmebene

$ERROR Buchstabe
systemdefinierte Dichtefunktion
Buchstabe : B = Binomial
N = Normal
P = Poisson
G = Gamma

$EXIT [Integer]
Gehe um die in "Integer" angegebene Anzahl an Programmebenen zurück.

$EXTRACT Systemvektor
weist Werte aus der SSP - Matrix den korrespondierenden Systemvektoren (s.u.) %VL, %PL, %VC zu.

$FACTOR [Integer1] [Name Integer]
definiert einen Vektor (Variable) der Länge "Integer1" als Faktor (nomiale Größe) mit der maximalen Ausprägung "Integer". Eine fehlende Vektorlänge wird durch die Standardlänge ersetzt. Bei mehreren Faktoren muß "Name Integer" entsprechend wiederholt werden.

$FINISH ---
END - of - FILE - Marke
beim Aufrufen von Subfiles in einer Sekundärdatei ($INPUT). Wenn das gesuchte Subfile nicht gefunden wurde, bewirkt das erste Erreichen von $FINISH ein

Rewind der Datei und die Fortsetzung des Suchvorgangs bis zum nächsten $FINISH. $FINISH darf nicht auf dem Primäreingabekanal verwendet werden.

$FIT [Modell - Formel]
berechnet die Anpassung an das durch $ERROR und $LINK (bzw. $OWN, $YVARIATE, $WEIGHT, $OFFSET) definierte Modell. Die Berechnung kann durch $CYCLE, $RECYCLE und $ALIAS beeinflußt werden.
Die "Modell - Formel" kann bestehen aus:
Operatoren (+,-,*,/,.)
Operanden (Vektoren (Variablen))
Klammern ( )
Mit $FIT können "Modell - Formeln" entweder gesetzt oder verändert werden.

$FORMAT ---
Eingabe eines Datenformats
definiert auf der folgenden Zeile das Format zu lesender Daten entweder als leer oder FREE (freies Format, d.h. Trennung durch Blank(s) oder Zeilenwechsel) oder als gültiges FORTRAN-Format
Voreinstellung: FREE
Hinweis: stimmen die einzulesenden Daten nicht mit dem Format überein, so wird mit einer FORTRAN-Fehlermeldung GLIM beendet. Nach "(" des FORTRAN-Formates sollten noch 4 Blanks eingegeben werden.

$HELP ---
Art der Erkärung der Fehlermeldungen: kurz oder ausführlich (Ja/Nein).
Voreinstellung: ausführlich

$INPUT Inter1 [Integer2] [Subfiles]
liest Befehle oder Subfiles von der Kanal-Nummer "Integer1" (1 bis 99) mit der maximalen Satzlänge "Integer2" (30 bis 299).

$LINK Buchstabe
definiert die Link - Funktion für eine durch $ERROR definierte Dichtefunktion.
Buchstabe: C = complementary log - log
E = number exponent
G = logit
I = identity
L = log
P = probit
R = reciprocal
S = square root
Voreinstellung: der natürliche Parameter der Dichtefunktion.
Fehler: nicht jede Zuordnung von $ERROR und $LINK ist erlaubt (bzw. sinnvoll).

$LOOK [Integer1 [Integer2]] Vektoren oder Skalare
Spaltenweise Ausgabe von Vektoren (Variablen) oder Skalaren mit den Indexgrenzen "Integer1" bis "Integer2". Bei der Ausgabe von Teilvektoren wird eine Spalte mit Zeilenindizes vorangestellt; bei Überschreitung der definierten Zeilenlänge wird rechts abgeschnitten.

$LSEED [Integer1 [Integer2 [Integer3]]]
setzt Startwerte für den lokalen Zufallszahlengenerator (wie $SSEED).

$MACRO Name Inhalt
definiert ein Makro mit "Namen" und "Inhalt" oder überschreibt ein schon bestehendes Makro mit neuem Inhalt.

$OFFSET [Name]
"Name" desjenigen Vektors (Variablen), dessen Inhalt bei $FIT zum linearen Prädiktor hinzuaddiert werden soll.
Voreinstellung: undefiniert.

$OWN macro1 ... macro4
definiert ein benutzereigenes GL - Modell durch 4 Makros. Die Verbindungsfunktion wird durch macro1 und macro2 definiert, die Dichtefunktion durch macro3 und macro4 (Systemvektoren % ..., s.u.):

```
macro1 = berechnet %FV aus %LP
macro2 = berechnet %DR
macro3 = berechnet %VA
macro4 = berechnet %DI
```

$PAUSE ---
unterbricht GLIM für Betriebssystem - Kommandos (nicht überall implementiert).

$PLOT Y-Vektoren X-Vektor
zeichnet ein Scattergramm mit maximal 9 Vektoren (Variablen) auf der y-Achse gegen einen Vektor (Variable) auf der x-Achse.

$PRINT [Text]
Druckbefehl: druckt Zahlen und Text.

$READ Zahlen
Einlesen von Zahlen für Vektoren (Variablen). Die gelesenen Zahlen werden zeilenweise in die durch $DATA aus Spaltenvektoren definierte Matrix abgelegt. Die gelesenen Zeilen dürfen kürzer oder länger als die Matrixzeilen sein. Das zeilenweise Lesen endet, wenn die Matrix gefüllt ist.

$RECYCLE [Integer1 [Integer2]]
wie $CYCLE, jedoch werden beim FIT eines Standardmodells keine neuen Anfangswerte für die %FV gesetzt.

$RESTORE [Integer]
erneuter Programmbeginn von einem vorher gespeicherten Dump von Kanal "Integer".
Hinweis: wird versucht über das Datenende hinaus zu lesen, endet GLIM mit einer FORTRAN-Fehlermeldung.

$REWIND [Integer]
setzt den Kanal "Integer" auf den Anfang zurück.

$SCALE [Zahl]
setzt "Zahl" als Anfangswert für Skalierungsparameter; 0 oder leer bedeutet iteratives Ersetzen durch den jeweiligen Quotienten aus Devianz und Freiheitsgraden .
Voreinstellung: durch $ERROR wird der Anfangswert bestimmt:
0 (Normal-, Gammaverteilung)
1 (Binomial-, Poissonverteilung)

$SKIP [Integer]
wie $EXIT
mit dem Unterschied, daß die zu verlassende Programmebene in Programmschleifen nur bedingt velassen wird.

$SORT [Vektor1 [Vektor2 oder Integer2 [Vektor3 oder Integer3]]]
sortieren von Vektoren (Variablen).

$SSEED [Integer1 [Integer2 [Integer3]]]
Setzen der Startwerte des Standardzufallszahlengenerators.

$STOP ---
beendet GLIM

$SUBFILE Name
definiert einen externen Subfile mit "Namen". Ein Subfile endet mit $FINISH.

$SUBPEND ---
ruft den Standardeingabekanal auf.

$SWITCH scalar macros
Verzweigung zu einem von mehreren Makros
"scalar" bezeichnet die Reihenfolgenummer des aufzurufenden Makros, falls "scalar" kleiner 1 oder größer der Anzahl der Makros, erfolgt kein Aufruf.

$UNIT Integer
"Integer" definiert die Standardlänge der Vektoren (Variablen).

$USE macro
Aufruf eines Makros mit dem Namen "macro".

$VARIATE [Integer] Name
Definiert Vektoren (Variablen) der Länge "Integer". Eine fehlende Vektorlänge "Integer" für einen noch nicht definierten Vektor wird durch die Standardlänge ersetzt. Eine angegebene Vektorlänge "Integer" für einen schon definierten Vektor muß gleich der bekannten Vektorlänge sein (Verbot der Redefinition von Vektorlängen).

$WARNING ---
Ausgabe von Warnungen: Ja/Nein
Voreinstellung: Ja

$WEIGHT [Name]
Deklaration eines Gewichtungsfaktors
der Vektor (Variable) "Name" enthält nicht-negative a-priori-Gewichte für FITs. Es erfolgt keine automatische Normierung der Gewichte.

$YVARIATE Name
der Vektor "Name" wird als abhängige Variable definiert.

Sonderzeichen

$	Befehlszeichen
:	Wiederholungszeichen
%	Funktionszeichen
#	Ersetzungszeichen
!	Zeilenendzeichen
"	Anführungsstrich

Systemspezifische Skalare

%A, %B,....., %Z	einfache Skalare
%JN	Jobnummer
%NU	Anzahl der Beobachtungen
%DV	skalierte Devianz
%DF	Freiheitsgrade
%X2	verallgemeinertes (Pearson) Chi-Quadrat
%SC	Skalenparameter
%CL	aktuelle Programmebene
%ML	Anzahl der Elemente der Varianz-Kovarianz-Matrix
%PL	Anzahl der geschätzten Parameter im Modell
%PI	3.14159

Systemspezifische Vektoren (Länge in Klammern)

%FV	durch $FIT geschätzte Werte (%NU)
%LP	linearer Prädiktor (%NU)
%WT	iterative Gewichte (%NU)
%YV	abhängige Variable (%NU)
%BD	binomialer Nenner (%NU)
%PW	a priori Gewichte (%NU)
%OS	vereinbartes Offset (%NU)
%DR	vom Benutzer vereinbarte -macro2- Verbindungsfunktion %LP/%FV (%NU)
%VA	Varianzfunktion (%NU)
%DI	Beitrag jeder Beobachtung zur Gesamtdevianz (%NU)

%GM Great Mean (%NU)
%VC Varianz-Kovarianz-Matrix -nur über $EXTRACT- (%ML)
%VL Varianz des linearen Prädiktors -nur über $EXTRACT- (%NU)
%RE Maske für $PLOT und $DISPLAY (%NU)

Funktionen

X und Y sind entweder Vektoren (Variablen) oder Skalare

%ANG(X) Angulartransformation [arcsin (sqrt X)]
%EXP(X) Exponentialfunktion [e**X]
%LOG(X) natürlicher Logarithmus [ln X]
%SIN(X) Sinusfunktion [sin X]
%SQRT(X) Quadratwurzel [√X]
%NP(X) Integral der Standard-Normalverteilung von - ∞ bis X
%ND(X) standardnormalverteilte Zufallsvariable Z mit der Wahrscheinlichkeit X (Umkehrfunktion von %NP) [0 < X < 1]
%TR(X) ganzzahliger Teil von X (Rundung zur nächtskleineren ganzen Zahl)
%GL(k,n) füllt nacheinander einen Vektor (Variable) mit Blöcken aus je n natürlichen Zahlen, beginnend bei 1 und spätestens endend bei k. Falls eine vordefinierte Länge des Vektors groesser ist als k*n, wird die Zuweisung zyklisch fortgesetzt. Die Argumente k bzw. n dürfen auch Vektoren sei.
%CU(X) Kumulierte Summen
%SR(n) Standard-Zufallszahlengenerator im Intervall [0,1]
%LR(n) Lokaler Zufallszahlengenerator im Intervall [0,1]
%LT(X,Y) 1, wenn X kleiner als Y, sonst 0
%LE(X,Y) 1, wenn X kleiner gleich Y, sonst 0
%EQ(X,Y) 1, wenn X gleich Y, sonst 0
%NE(X,Y) 1, wenn X ungleich Y, sonst 0
%GE(X,Y) 1, wenn X größer gleich Y,sonst 0
%GT(X,Y) 1, wenn X größer als Y, sonst 0
%IF(bed,X,Y) X, wenn Bedingung wahr, sonst Y

Arithmetische Operatoren
(in der Reihenfolge ihrer Bearbeitung)

**	Exponentiation
/	Division
*	Multiplikation
-	Subtraktion
+	Addition
=	Zuweisung