Benutzerdefinierte Design-Matrizen in log-linearen Analysen: Realisierungsmöglichkeiten in den SPSS-Prozeduren GENLOG und LOGLINEAR

Kühnel, Steffen M.

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

Benutzerdefinierte Design-Matrizen in log-linearen Analysen: Realisierungsmöglichkeiten in den SPSS-Prozeduren GENLOG und LOGLINEAR

von Steffen M. Kühnel [1]

Zusammenfassung

Der Anwendung log-linearer Modelle in der Sozialforschung steht oft die Vorstellung entgegen, daß diese Modelle recht kompliziert und daher kaum zu interpretieren seien. Das Verständnis für log-lineare Analysen wird erleichtert, wenn die Verwandtschaft zur multiplen Regression mit nominalskalierten Prädiktoren gesehen wird. Gleichzeitig kann so auch die Bedeutung der sogenannten Design-Matrix nahegebracht werden. Die volle Flexibilität log-linearer Modelle wird nämlich erst durch die Formulierung benutzerdefinierter Design-Matrizen erreicht. Anhand von Beispieldaten aus dem ALLBUS 1996 wird gezeigt, wie sich bei Anwendung der SPSS-Prozeduren GENLOG oder LOGLINEAR log-lineare Analysen mit benutzerdefinierten Design-Matrizen realisieren lassen.

Abstract

Applications of log-linear modelling are sometimes prevented by the impression that this technique is not user-friendly. Nevertheless, log-linear modelling is nothing more than multiple regression of the logarithms of cell counts on categorical predictors. Within this view the importance of the design matrix is easy to understand. The specification of user-defined design matrices within log-linear models allows for very flexible analyses of categorical data. It is shown how such analyses can be done using the SPSS procedures GENLOG or LOGLINEAR. An empirical example is given based on data from the ALLBUS 1996.

[1] Anschrift des Autors: Prof. Dr. ***Steffen M. Kühnel***, Justus-Liebig Universität Gießen, Fachbereich Gesellschaftswissenschaften, Institut für Politikwissenschaft, Karl-Glöcker-Str. 21E, 35394 Gießen

Log-lineare Modelle ermöglichen die multivariate Zusammenhangsanalyse bei kategorialen Daten. Die vielfältigen Möglichkeiten dieser Modellklasse lassen sich erst dann voll ausnutzen, wenn der Anwender die Möglichkeit hat, benutzerspezifische Design-Matrizen zu definieren. Es ist leider kaum bekannt, daß log-lineare Modelle mit benutzerdefinierten Design-Matrizen auch mit den SPSS-Prozeduren GENLOG oder LOGLINEAR geschätzt werden können. Im vorliegenden Beitrag möchte ich anhand eines einfachen Beispiels aus dem ALLBUS 1996 zeigen, wie hierbei vorzugehen ist. Für Leser, die mit der Logik log-linearer Modelle nicht so vertraut sind, will ich zunächst die Grundidee der log-linearen Analyse vorstellen.[2]

1. Die Logik log-linearer Tabellenanalysen

In der bekannten linearen Regression ergeben sich die Werte einer abhängigen Variable Y als lineare Funktion der Werte von erklärenden Variablen X_1, X_2, ..., und der Residual variable E:

$$Y = b_0 + b_1 \cong X_1 + b_2 \cong X_2 + ... + E \ .$$

Ganz analog kann auch die log-lineare Analyse einer mehrdimensionalen Tabelle als ein spezielles Regressionsmodell aufgefaßt werden. Dabei bilden die logarithmierten Häufigkeiten der Tabellenzellen die Werte der abhängigen Variable. Vorhergesagt werden diese Werte durch eine lineare Funktion von erklärenden (Design-) Variablen. Diese Sichtweise verdeutlicht auch die Bezeichnung "log-linear".

Über diese Analogie zur linearen Regression läßt sich die Logik der log-linearen Analyse und die Interpretation der Modellparameter relativ leicht nachvollziehen. Als Beispiel soll im folgenden der bivariate Zusammenhang zwischen der Wahlbeteiligung und der wahrgenommenen Bürgernähe von Politikern mit log-linearen Modellen untersucht werden. Im ALLBUS 1996 (ZA-Studiennummer 2800) finden sich hierzu zwei Fragen. Die perzipierte Bürgernähe von Politikern (V20) wird über ein Item aus der Anomia-Skala erfaßt, bei dem nach der Zustimmung zu der Meinung *"Die meisten Politiker interessieren sich in Wirklichkeit gar nicht für die Probleme der einfachen Leute"* gefragt wird. Als Antwortmöglichkeiten werden die Kategorien *"Bin derselben Meinung"* (Kode: 1), *"Bin anderer Meinung"* (Kode: 2) und *"Weiß nicht"* (Kode: 3) berücksichtigt. Die Wahlbeteiligung (V326) wird über die berichtete Wahlbeteiligung bei der letzten Bundestagswahl operationalisiert. Tabelle 1 gibt die in den ALLBUS-Daten vorzufindenden bivariaten Antworthäufigkeiten auf die beiden Fragen wieder. Neben den absoluten Häufigkeiten sind auch die logarith-

[2] Aus Platzgründen kann hier keine umfassende Einführung in die log-lineare Datenanalyse gegeben werden. Hierzu sei auf entsprechende Monographien verwiesen, etwa das kürzlich erschienene Lehrbuch von ***Andreß***, ***Hagenaars*** und ***Kühnel*** (1997) zur kategorialen Datenanalyse.

mierten Häufigkeiten aufgeführt. Der Wert 7.6917 der ersten Tabellenzelle ist beispielsweise der natürliche Logarithmus der Häufigkeit 2190 (ln(2190)≈7.6917).

Tabelle 1: Wahlbeteiligung und Bürgernähe von Politikern

Wahlbeteiligung bei letzter Bundestagswahl (V326)	Politiker nicht an Problemen interessiert (V20)		
	ja (1)	nein (2)	weiß nicht (3)
ja (1)	2190 *7.6917*	451 *6.1115*	180 *5.1930*
nein (2)	547 *6.3044*	75 *4.3175*	68 *4.2195*

Erster Wert: absolute Häufigkeit, zweiter Wert (kursiv): logarithmierte Häufigkeit
(Quelle: ALLBUS 1996)

In der log-linearen Analyse der Tabelle bilden die sechs logarithmierten Zellenhäufigkeiten die Werte der abhängigen Variable. Erklärt werden diese Häufigkeiten durch die Variablen, die die Tabelle definieren, hier also durch die Einschätzung der Desinteressiertheit von Politikern (V20) und durch die Wahlbeteiligung (V326). Beide Variablen werden zunächst als nominalskaliert aufgefaßt. In der linearen Regression kann eine nominalskalierte unabhängige Variable nicht direkt in die Modellgleichung aufgenommen werden. Die Berücksichtigung solcher Erklärungsgrößen erfolgt statt dessen indirekt über sogenannte Design-Variablen.[3] Gleiches gilt auch für log-lineare Modelle.

Die Umsetzung nominalskalierter Variablen in Design-Variablen kann nach unterschiedli-chen Regeln erfolgen. Gemeinsam ist allen Regeln, daß aus einer Variable mit insgesamt K verschiedenen Ausprägungen maximal K-1 Design-Variablen gebildet werden können. Im Beispiel werden also die drei Kategorien der Bürgernähe von Politikern in zwei DesignVariablen und die zwei Kategorien der berichteten Wahlbeteiligung in eine DesignVariable umgewandelt.[4] In der linearen Regression wird bei einer solchen Transformation häufig die *Dummykodierung* verwendet. Dabei werden den einzelnen Kategorien der Ausgangsvariable dichotome 0/1-kodierte Variablen zugeordnet, die immer dann den Wert '1' aufweisen, wenn

[3] Statt von 'Design-Variablen' wird meist von 'Dummy-Variablen' gesprochen. Da im folgenden von *Dummykodierung* bzw. *Effektkodierung* der Design-Variablen die Rede sein wird, verwende ich hier den neutraleren Ausdruck 'Design-Variable'.

[4] Da irgendeine der Ausprägungen einer Variable notwendigerweise realisiert wird, läßt sich das Auftreten einer Kategorie bei Kenntnis des Auftretens der übrigen Kategorien perfekt vorhersagen. Wird für jede Kategorie eine eigene Design-Variable erzeugt, enthält daher eine dieser Variablen keine zusätzlichen Informationen. Technisch gesprochen bestünde eine perfekte Multikollinearität unter der Prädiktoren.

die betreffende Kategorie bei einem Fall vorkommt. Die Kategorie, für die keine eigene 0/1-kodierte Design-Variable spezifiziert wird, wird als *Referenzkategorie* bezeichnet. Weist die nominalskalierte Ausgangsvariable bei einem Fall den Wert der Referenzkategorie auf, haben alle 0/1-kodierten Design-Variablen den Wert '0'. In der Varianzanalyse wird dagegen oft die *Effektkodierung* eingesetzt. Der Unterschied zur Dummykodierung liegt darin, daß die Referenzkategorie der Dummykodierung hier den Wert '-1' aufweist. Tabelle 2 zeigt die beiden Kodierungsmöglichkeiten für die Variablen aus Tabelle 1. Referenzkategorie bei der Einschätzung des Desinteresse von Politikern ist die Antwort "weiß nicht" (V20=3) und bei der Frage nach der Wahlbeteiligung die Antwort "nein" (V326=2).

Tabelle 2: Transformation der Modellvariablen in dummy- und effekt kodierte Design-Variablen

	Politiker sind desinteressiert (V20):			Wahlbeteiligung (V326)	
	ja (1)	nein (2)	weiß nicht (3)	ja (1)	nein (2)
Dummykodierung:					
Erste Design-Variable	1	0	0	1	0
Zweite Design-Variable	0	1	0		
Effektkodierung:					
Erste Design-Variable	1	0	-1	1	-1
Zweite Design-Variable	0	1	-1		

Mit Hilfe der Design-Variablen können nun die Vorhersagegleichungen für das log-lineare Modell formuliert werden. Zunächst soll ganz analog zur linearen Regression ein Modell geschätzt werden, bei dem die logarithmierten Häufigkeiten nur durch die Regressionskonstante und die Design-Variablen prognostiziert werden. Bei der Dummy-Kodierung ergeben sich dann folgende Vorhersagegleichungen:[5]

V326	V20	Y	≈	$\hat{\mathbf{Y}}$	=	$b_0 \cdot 1$	+	$b_1 \cdot D_1$	+	$b_2 \cdot D_2$	+	$b_3 \cdot D_3$
1	1	7.6917	≈	?	=	$b_0 \cdot 1$	+	$b_1 \cdot 1$	+	$b_2 \cdot 0$	+	$b_3 \cdot 1$
1	2	6.1115	≈	?	=	$b_0 \cdot 1$	+	$b_1 \cdot 0$	+	$b_2 \cdot 1$	+	$b_3 \cdot 1$
1	3	5.1030	≈	?	=	$b_0 \cdot 1$	+	$b_1 \cdot 0$	+	$b_2 \cdot 0$	+	$b_3 \cdot 1$
2	1	6.3044	≈	?	=	$b_0 \cdot 1$	+	$b_1 \cdot 1$	+	$b_2 \cdot 0$	+	$b_3 \cdot 0$
2	2	4.3175	≈	?	=	$b_0 \cdot 1$	+	$b_1 \cdot 0$	+	$b_2 \cdot 1$	+	$b_3 \cdot 0$
2	3	4.2195	≈	?	=	$b_0 \cdot 1$	+	$b_1 \cdot 0$	+	$b_2 \cdot 0$	+	$b_3 \cdot 0$

[5] Um die Ähnlichkeit zur linearen Regression zu verdeutlichen, habe ich die Regressionskoeffizienten durch den Buchstaben "b" gekennzeichnet. In der üblichen auf ***L.A. Goodman*** zurückgehenden Notation werden für die Regressionsgewichte üblicherweise kleine Lambdas (λ) verwendet.

Die in den Gleichungen vorkommenden Werte der Design-Variablen ergeben sich aus der Position der entsprechenden Zelle in der ursprünglichen Häufigkeitstabelle (Tabelle 1). Um dies deutlich zu machen, sind links jeweils die Werte der Ausgangsvariablen V326 und V20 angegeben. Die Design-Variable D_1 hat immer dann den Wert '1', wenn die erste Kategorie der Bürgernähe von Politikern auftritt (V20=1). Ansonsten ist ihr Wert stets '0'. Die nächste Design-Variable D_2 bezieht sich auf die zweite Kategorie der Bürgernähe (V20=2) und die letzte Design-Variable D_3 auf die erste Kategorie der Wahlbeteiligung (V326=1). Die Werte der abhängigen Variablen (Y) sind die logarithmierten Besetzungen der Zellen von Tabelle 1. Sie werden durch Vorhersagewerte ($\hat{Y}$) prognostiziert, die sich mit Hilfe der Regressionskoeffizienten aus den Werten der Design-Variablen berechnen lassen. Die zunächst unbekannten Koeffizienten sind die Parameter des log-linearen Modells. Die Realisierungen der Design-Variablen bilden die *Design-Matrix*, die der Datenmatrix der erklärenden Variablen in der linearen Regression entspricht. Jede Zeile der Design-Matrix gibt die Werte aller Design-Variablen für eine Zelle der Ausgangstabelle an. Jede Spalte enthält alle Werte einer Design-Variable. Um auch die Regressionskonstante berücksichtigen zu können, enthält die erste Spalte der Design-Matrix eine Konstante mit dem Wert '1'.

Die vier Regressionskoeffizienten b_0 bis b_3 werden nun so bestimmt, daß die Vorhersagewerte den logarithmierten Häufigkeiten möglichst ähnlich sind. Anders als bei der im linearen Regressionsmodell üblichen Schätzung nach dem Kleinstquadrate-Kriterium wird bei log-linearen Modellen das Kriterium der Maximum-Likelihood-Methode verwendet (ML-Schätzung). Bei der ML-Schätzung werden die Koeffizienten so festgelegt, daß eine maximale Wahrscheinlichkeit besteht, daß die tatsächlich beobachteten Werte der abhängigen Variable - hier also die logarithmierten Zellenhäufigkeiten - als Funktion der Koeffizienten realisiert werden können.

Die Dummy-Kodierung der Design-Variablen wird bei der Prozedur GENLOG in SPSS verwendet, die in SPSS für Windows bei der Spezifikation eines log-linearen Modells über das Pull-Down-Menue aktiviert wird (***Norusis***/SPSS Inc., 1994). Alternativ kann die Schätzung der Modellparameter auch im Syntax-Fenster angefordert werden:

```
genlog V326 V20
   /print design freq estim /plot none
   /criteria = delta(0)
   /design V20 V326 .
```

Nach dem Prozedurnamen GENLOG werden zunächst die Variablen aufgeführt, die die zu analysierende Tabelle aufspannen. Im Beispiel sind dies die beiden ALLBUS-Variablen V20 und V326. Die hinter dem Schlüsselwort "/print" angegebenen Spezifikationen steuern die Ausgabe der Prozedur. Angefordert wird die Ausgabe der Design-Matrix ("design"), die beobachteten und vorhergesagten Zellenhäufigkeiten ("freq") und die Schätzungen der Regressionskoeffizienten ("estim"). Mit der Option "/plot none" wird die Ausgabe von

Abbildung 1: Ergebnisse der Parameterschätzung bei Dummykodierung (SPSS-Prozedur: GENLOG)

```
Correspondence Between Parameters and Terms of the Design
Parameter   Aliased  Term
       1            Constant
       2            [V20 = 1]
       3            [V20 = 2]
       4       x    [V20 = 3]
       5            [V326 = 1]
       6       x    [V326 = 2]
Note: 'x' indicates an aliased (or a redundant) parameter.
      These parameters are set to zero.
Design Matrix
                          Cell          Parameter
Factor        Value  Structure     1     2     3     5
V326           ja
 V20            ja       1.000     1     1     0     1
 V20          nein       1.000     1     0     1     1
 V20    weiß nicht       1.000     1     0     0     1
V326          nein
 V20            ja       1.000     1     1     0     0
 V20          nein       1.000     1     0     1     0
 V20    weiß nicht       1.000     1     0     0     0
Table Information
                       Observed              Expected
Factor        Value       Count       %         Count       %
V326           ja
 V20            ja      2190.00 ( 62.38)      2199.11 ( 62.63)
 V20          nein       451.00 ( 12.85)       422.63 ( 12.04)
 V20    weiß nicht       180.00 (  5.13)       199.26 (  5.68)
V326          nein
 V20            ja       547.00 ( 15.58)       537.89 ( 15.32)
 V20          nein        75.00 (  2.14)       103.37 (  2.94)
 V20    weiß nicht        68.00 (  1.94)        48.74 (  1.39)
Goodness-of-fit Statistics
                    Chi-Square       DF       Sig.
Likelihood Ratio       19.3674        2     6.E-05
         Pearson       19.3584        2     6.E-05
Parameter Estimates
                                                 Asymptotic 95% CI
Parameter   Estimate         SE    Z-value      Lower      Upper
        1     3.8865      .0721      53.92       3.75       4.03
        2     2.4012      .0663      36.21       2.27       2.53
        3      .7519      .0770       9.76        .60        .90
        4      .0000      .             .          .          .
        5     1.4082      .0425      33.16       1.32       1.49
        6      .0000      .             .          .          .
```

Grafiken zur Beurteilung der Modellanpassung unterdrückt. In SPSS wird standardmäßig der Wert 0.5 zu allen Zellenhäufigkeiten addiert.[6] Über die Option "criteria = delta (0)" wird diese Voreinstellung ausgeschaltet. Die zu analysierenden Tabellenhäufigkeiten bleiben so unverändert. In der letzten Option "/design" wird das log-lineare Modell spezifiziert, dessen Parameter geschätzt werden sollen. Die Angabe der beiden Modellvariablen V20 und V326 führt dazu, daß die Prozedur temporär für diese beiden Variablen dummykodierte Design-Variablen erzeugt, die zur Prognose der logarithmierten Häufigkeiten herangezogen werden.

[6] Für die Zahl null ist ein Logarithmus nicht definiert. Unbesetzte Tabellenzellen können daher bei log- linearen Analysen zu Problemen führen.

In Abbildung 1 ist leicht gekürzt die Ausgabe der Prozedur GENLOG dokumentiert. Zunächst werden Informationen zur Bedeutung der geschätzten Parameter gegeben. Der erste Modellparameter ist die Regressionskonstante ('Constant'). Es folgt das Regressionsgewicht der Design-Variable für die erste Ausprägung der Bürgernähe von Politikern ('V20=1'), anschließend das Gewicht für die Design-Variable der zweiten Ausprägung ('V20=2'). Die dritte Ausprägung von V20 ist die Referenzkategorie. Für sie wird keine eigene Design-Variable erzeugt. Dies wird durch ein 'x' in der mit "Aliased" überschriebenen Spalte gekennzeichnet. Anschließend folgt der Parameter der Design-Variable für die erste Ausprägung der Wahlbeteiligung ('V326=1'). Die zweite Ausprägung ist wiederum durch ein 'x' als Referenzkategorie gekennzeichnet.

Es folgt die Wiedergabe der bei der Schätzung verwendeten Design-Matrix. Zur leichteren Identifizierung der Tabellenzellen sind links die Ausprägungen der die Tabelle definierenden Variablen wiedergegeben. Die mit "Cell Structure" überschriebene Spalte bezieht sich darauf, daß den Tabellenzellen Gewichte zugeordnet werden können, um beispielsweise durch ein Gewicht von 0.0 unbesetzte Zellen von der Analyse auszuschließen. Voreinstellung ist das Gewicht 1.0 für jede Zelle. Schließlich folgen die eigentlichen Spalten der Design-Matrix. Durch die in der Spaltenüberschrift angegebene Parameternummer kann erschlossen werden, auf welchen Modellparameter sich eine Spalte der Design-Matrix bezieht.

Am Ende der Ausgabe werden die geschätzten Regressionskoeffizienten ('Estimate'), deren Standardfehler ('SE'), die Quotienten aus Koeffizienten und Standardfehler ('Z- value') und die Grenzen der asymptotischen 95%-Konfidenzintervalle ausgedruckt. Die ML-Schätzung der Koeffizienten ergibt bei den Daten des ALLBUS 1996 aus Tabelle 1 eine Regressionskonstante von 3.8865. Für das Regressionsgewicht der ersten Design- Variable D_1 wird der Wert 2.4012, für das der zweiten Design-Variable D_2 der Wert 0.7519 und für das der letzten Design-Variablen D_3 der Wert 1.4082 geschätzt. Setzt man diese Werte in die Vorhersagegleichung ein, ergeben sich die Prognosen für die logarithmierten Häufigkeiten. In Tabelle 3 ist die Berechnung der logarithmierten Häufigkeiten exemplarisch durchgeführt.

Die letzte Spalte der Tabelle enthält zusätzlich die vorhergesagten absoluten Häufigkeiten. Diese ergeben sich, wenn der Antilogarithmus das ist die Umkehrfunktion des Logarithmierens der Vorhersagewerte aus der ersten Spalte berechnet werden. 2199.31 ist beispielsweise der Antilogarithmus von 7.6959 ($2199.31 \approx e^{7.6959}$).[7] Das Modell sagt somit für die erste Tabellenzelle eine Häufigkeit von 2199.31 voraus. Dies ist der geschätzte Erwartungswert, der sich im Durchschnitt über alle möglichen Zufallsstichproben ergeben würde, falls die geschätzten Regressionskoeffizienten die tatsächlichen Populationswerte sind. Die

[7] Als Folge von Rundungsfehlern stimmen die in Tabelle 3 berechneten erwarteten Häufigkeiten nicht genau mit den von SPSS ausgedruckten Werten in Abbildung 1 überein.

prognostizierten Häufigkeiten werden daher auch als erwartete Häufigkeiten ('expected counts') bezeichnet.

Der Vergleich mit den beobachteten Häufigkeiten weist bei den Kategorien '2' und '3' von V20 (Bürgernähe) deutliche Abweichungen auf.[8] Das Modell scheint also nicht mit den Daten vereinbar zu sein. Im Unterschied zur linearen Regression, wo Abweichungen zwischen den Vorhersagewerten und den tatsächlichen Werten der abhängigen Variablen auf nichterfaßte Größen zurückgeführt werden können, die in der Residualvariable zusammengefaßt sind, wird in log-linearen Modellen grundsätzlich unterstellt, daß die Design-Variablen bei Kenntnis der "wahren" Modellparameter auch stets die "wahren" erwarteten Häufigkeiten wiedergeben. Abweichungen zwischen den Vorhersagen ($\hat{Y}$) und den beobachteten Werten (Y) können dann nur Folge von zufälligen Stichprobenschwankungen sein.[9]

Tabelle 3: Berechnung der vorhergesagten logarithmierten Häufigkeiten

$\hat{Y}$ =	$b_0 \cong 1$	$+ b_1 \cong D_1$	$+ b_2 \cong D_2$	$+ b_3 \cong D_3$	$e^{\hat{Y}}$
7.6959 =	3.8865 ≅ 1	+ 2.4012 ≅ 1	+ 0.7519 ≅ 0	+ 1.4082 ≅ 1	2199.31
6.0466 =	3.8865 ≅ 1	+ 2.4012 ≅ 0	+ 0.7519 ≅ 1	+ 1.4082 ≅ 1	422.67
5.2947 =	3.8865 ≅ 1	+ 2.4012 ≅ 0	+ 0.7519 ≅ 0	+ 1.4082 ≅ 1	199.28
6.2877 =	3.8865 ≅ 1	+ 2.4012 ≅ 1	+ 0.7519 ≅ 0	+ 1.4082 ≅ 0	537.91
4.6384 =	3.8865 ≅ 1	+ 2.4012 ≅ 0	+ 0.7519 ≅ 1	+ 1.4082 ≅ 0	103.38
3.8865 =	3.8865 ≅ 1	+ 2.4012 ≅ 0	+ 0.7519 ≅ 0	+ 1.4082 ≅ 0	48.74

Zur Beurteilung der Übereinstimmung von Modell und Daten können Goodness-of-Fit Teststatistiken herangezogen werden. In der SPSS-Ausgabe werden sowohl die Statistik für den Likelihood-Ratio-Test als auch die für Pearsons Anpassungstest ausgegeben. Unter der (Null-) Hypothese, daß Abweichungen zwischen beobachteten und erwarteten Häufigkeiten tatsächlich nur durch zufällige Stichprobenschwankungen hervorgerufen werden,sind beide Teststatistiken asymptotisch chiquadratverteilt. Die Freiheitsgrade ergeben sich dabei aus der Differenz der Zeilenzahl (i.a. die Anzahl der Zellen der zu analysierenden Tabelle) und der Spaltenzahl (Zahl der geschätzten Modellparameter) der Design-Matrix. Beide Test-

[8] Bezogen auf die relativen Häufigkeiten sind die Abweichungen allerdings nicht sehr groß. Dies ist eine Folge der starken Schiefe der Variable V20. Bei schiefen Verteilungen führen Prozentsatzdifferenzen leicht in die Irre.

[9] Etwas anders sieht es aus, wenn - wie z.B. in der latenten Klassenanalyse (LCA) - die Ausgangsvariablen, die die Tabelle generieren, auf inhaltlich relevante Größen und Meßfehler zurückgeführt werden. Aber auch dann gilt, daß *nach* der Berücksichtigung von Meßfehlern alle weiteren Abweichungen zwischen beobachteten und erwarteten Häufigkeiten auf Stichprobenschwankungen zurückgeführt werden.

statistiken weisen Werte um 19.4 auf; bei zwei Freiheitsgraden ergibt sich ein empirisches Signifikanzniveau ("Sig.") kleiner 0.001. Die Wahrscheinlichkeit rein zufälliger Abweichungen ist also nahezu null. Es gibt somit gute Gründe, an der Übereinstimmung von Modell und Daten zu zweifeln.

Obwohl das Modell offensichtlich nicht zutrifft, soll es doch als ein einfaches Beispiel dafür herangezogen werden, wie die Koeffizienten eines log-linearen Modells interpretiert werden können. Aus der Design-Matrix (bzw. Tabelle 3) läßt sich erkennen, daß die Regressionskonstante den Logarithmus der geschätzten Häufigkeit für die Tabellenzelle angibt, bei der alle Design-Variablen den Wert null aufweisen. Im Beispiel ist dies die letzte Tabellenzelle von Tabelle 1 bzw. die letzte Zeile von Tabelle 3, die der Ausprägungskombination V20=3 (keine Meinung zum Desinteresse von Politikern) und V326=2 (keine Wahlbeteiligung) entspricht. Die erwartete Häufigkeit von 48.74 ergibt sich wieder über den Antilogarithmus der prognostizierten logarithmierten Häufigkeit ($48.74.e^{3.8865}$).

Das Regressionsgewicht von 2.4012 der ersten Design-Variable bezieht sich auf die erste Ausprägung der Bürgernähe von Politikern. Aus dem Vergleich der Vorhersagegleichungen in Tabelle 3 läßt sich erkennen, daß dieser Koeffizient besagt, um welchen Wert die loga-rithmierten Zellenbesetzungen im Durchschnitt jeweils ansteigen, wenn statt der Referenzkategorie 'weiß nicht' (V20=3) die erste Ausprägung 'ja' der Variable betrachtet wird (V20=1). Da in dem Modell die Effekte der zweiten erklärenden Variable Wahlbeteiligung über den Koeffizienten der Design-Variable D_3 kontrolliert werden, handelt es sich um einen partiellen Effekt bei Kontrolle der Ausprägungen der Wahlbeteiligung. Der Antilogarithmus dieses Koeffizienten gibt dann an, um welchen Faktor die Zellenhäufigkeit im Durchschnitt über der Häufigkeit der Referenzkategorie liegt. Es gibt im (geometrischen) Mittel ungefähr 11 mal mehr Personen, die der Ansicht 'Politiker sind desinteressiert' zustimmen, als es Personen gibt, die mit 'weiß nicht' antworten ($e^{2.4012}$. 11.03).

Die Interpretation der übrigen Koeffizienten folgt der gleichen Logik. Für die zweite Kategorie der Bürgernähe (V20=2) ergibt sich ein Koeffizient von 0.7519. Der Antilogarithmus dieser Zahl, 2.12 ($.e^{0.7519}$), gibt wiederum den Faktor an, mit dem diese Kategorie im Durchschnitt die Häufigkeit der Referenzkategorie übersteigt. Nach dem spezifizierten log-linearen Modell gibt es im Durchschnitt 2.12 mal so viele Befragte, die Politiker für interessiert halten (V20=2), wie es Befragte gibt, die hierzu keine Meinung haben (V20=3). Der letzte Koeffizient bezieht sich auf die erste Kategorie der Wahlbeteiligung (V326=1). Der positive Wert von 1.4082 weist darauf hin, daß die Zahl der Wähler (V326=1) im Durchschnitt um den Faktor 4.09 ($.e^{1.4082}$) höher ist als die Zahl der Nichtwähler (V326=2).

Die Regressionskoeffizienten der dummykodierten Design-Variablen des log-linearen Modells besagen also, wie die (logarithmierten) Häufigkeiten der Zellenbesetzungen mit den Ausprägungen der erklärenden Variablen relativ zur Referenzkategorie variieren. Die Inter-

pretationslogik entspricht derjenigen der linearen Regression mit nominalskalierten unabhängigen Variablen. Ein Unterschied besteht allerdings zwischen der üblichen linearen Regression und der log-linearen Vorhersage von Tabellenzellen. In der linearen Regression wird die Beziehung zwischen einer inhaltlich interessierenden abhängigen Variable und deren Prädiktoren modelliert. Die logarithmierten Häufigkeiten der Tabellenzellen dürften aber in den seltensten Fälle von theoretischem Interesse sein. Sie sind nur eine Hilfsgröße bei der Untersuchung des eigentlich interessierenden Zusammenhangs.

Tatsächlich impliziert das spezifizierte log-lineare Modell auch Aussagen über den Zusammenhang zwischen der Wahlbeteiligung (V326) und der Bürgernähe von Politikern (V20). Das Modell behauptet nämlich, daß diese beiden Größen statistisch unabhängig voneinander sind. Um dies zu erkennen, sei daran erinnert, daß zwei Ereignisse genau dann statistisch unabhängig voneinander sind, wenn ihre gemeinsame Auftretenswahrscheinlichkeit das Produkt der Auftretenswahrscheinlichkeiten der einzelnen Ereignisse ist. In der einfachen Tabellenanalyse wird dies genutzt, um die bei Unabhängigkeit erwarteten Häufigkeiten zu berechnen. Diese sind nämlich gerade das Produkt der über die relativen Häufigkeiten der Randverteilungen geschätzten Wahrscheinlichkeiten der Ausprägungen von Spalten- und Zeilenvariable sowie der Fallzahl.

Auf gleiche Weise berechnen sich die vorhergesagten Häufigkeiten des log-linearen Modells aus Abbildung 1. Tabelle 3 ist zu entnehmen, daß die logarithmierte Häufigkeit der ersten Tabellenzelle die Summe der Regressionskonstante und der Regressionsgewichte der ersten und dritten Design-Variable ist: 7.6959 = 3.8865 + 2.4012 + 1.4082. Die absoluten Häufigkeiten ergeben sich über die Berechnung der Antilogarithmen. Aus der Summe wird dabei ein Produkt: $2199.31 = e^{7.6959} = e^{3.8865} A e^{2.4012} A e^{1.4082}$. Wie bei statistischer Unabhängigkeit gefordert, ist die vorhergesagte Häufigkeit der ersten Tabellenzelle proportional zum Produkt des Effektes der ersten Kategorie der Zeilenvariable von Tabelle 1 ($e^{2.4012}$) und des Effekts der ersten Kategorie der Spaltenvariable ($e^{1.4082}$). Der so berechnete Vorhersagewert entspricht daher auch bis auf Rundungsfehler den erwarteten Häufigkeiten, die man bei der üblichen Berechnung des Chiquadrattests auf statistische Unabhängigkeit benötigt. Gleiches gilt für die übrigen fünf Tabellenzellen.

Ein statistischer Zusammenhang zwischen den beiden Variablen V20 und V326 wird erst zugelassen, wenn zusätzlich zu den Regressionsgewichten der Design-Variablen *Interaktionseffekte* spezifiziert werden. Das log-lineare Modell ist dazu um weitere Prädiktoren zu erweitern. Diese ergeben sich als Produkte der Design-Variablen des ersten Modells. Im Beispiel werden also die Design-Variablen der Ausprägungen von V326 (D_1 und D_2) mit der Design-Variable für die Ausprägungen von V326 (D_3) multipliziert. Um das erweiterte Modell mit der SPSS-Prozedur GENLOG zu schätzen, können in der Option "/design" die Interaktionseffekte durch die Spezifikation "V326*V20" oder "V326 by V20" angefordert werden.

Abbildung 2: Parameterschätzung im Modell mit Interaktionseffekten (SPSS-Prozedur: GENLOG)

```
-> genlog  V326 V20
->   /print design estim /plot none
->   /criteria = delta(0) /design V20 V326 V20*V326 .
Correspondence Between Parameters and Terms of the Design
Parameter   Aliased  Term
        1            Constant
        2            [V20 = 1]
        3            [V20 = 2]
        4       x    [V20 = 3]
        5            [V326 = 1]
        6       x    [V326 = 2]
        7            [V326 = 1]*[V20 = 1]
        8            [V326 = 1]*[V20 = 2]
        9       x    [V326 = 1]*[V20 = 3]
       10       x    [V326 = 2]*[V20 = 1]
       11       x    [V326 = 2]*[V20 = 2]
       12       x    [V326 = 2]*[V20 = 3]
Note: 'x' indicates an aliased (or a redundant) parameter.
      These parameters are set to zero.
Design Matrix
                          Cell                Parameter
Factor        Value  Structure     1     2     3     5     7     8
V326            ja
 V20            ja       1.000     1     1     0     1     1     0
 V20          nein       1.000     1     0     1     1     0     1
 V20    weiß nicht       1.000     1     0     0     1     0     0
V326         nein
 V20            ja       1.000     1     1     0     0     0     0
 V20          nein       1.000     1     0     1     0     0     0
 V20    weiß nicht       1.000     1     0     0     0     0     0
Goodness-of-fit Statistics
                   Chi-Square       DF       Sig.
Likelihood Ratio         .0000        0      .
         Pearson         .0000        0      .
Parameter Estimates
                                              Asymptotic 95% CI
Parameter   Estimate         SE    Z-value      Lower      Upper
        1     4.2195      .1213      34.79       3.98       4.46
        2     2.0849      .1286      16.21       1.83       2.34
        3      .0980      .1674        .59       -.23        .43
        4      .0000      .            .          .          .
        5      .9734      .1423       6.84        .69       1.25
        6      .0000      .            .          .          .
        7      .4138      .1502       2.76        .12        .71
        8      .8205      .1892       4.34        .45       1.19
        9      .0000      .            .          .          .
       10      .0000      .            .          .          .
       11      .0000      .            .          .          .
       12      .0000      .            .          .          .
```

Der Prozeduraufruf und Teile der SPSS-Ausgabe sind in Abbildung 2 wiedergegeben. Aus den Regressionskoeffizienten lassen sich wieder die Vorhersagen der logarithmierten Häufigkeiten berechnen (vgl. Tabelle 4). Der Vergleich der Regressionskonstante und der Koeffizienten für die Design-Variablen D_1, D_2 und D_3 mit den entsprechenden Koeffizienten des ersten Modells weist deutlich veränderte Werte auf. Änderungen der Koeffizientenschätzungen sind in der Regel zu beobachten, wenn die zusätzlich spezifizierten Interaktionseffekte die Vorhersagen der logarithmierten Häufigkeiten merklich verbessern können. Auch in der linearen Regression ändern sich oft die Regressionskoeffizienten, wenn zusätzliche erklärende Variablen in ein Modell aufgenommen werden.

Tabelle 4: Berechnung der vorhergesagten logarithmierten Häufigkeiten im Modell mit Interaktionseffekten

$\hat{Y}$ =	$b_0 \cong 1$	$+b_1 \cong D_1$	$+b_2 \cong D_2$	$+b_3 \cong D_3$	$+b_4 \cong (D_1 \cong D_3)$	$+b_5 \cong (D_2 \cong D_3)$
7.6116 =	4.2195 ≅ 1	+2.0849 ≅ 1	+0.0980 ≅ 0	+0.9734 ≅ 1	+0.4138 ≅ 1	+0.8205 ≅ 0
6.1114 =	4.2195 ≅ 1	+2.0849 ≅ 0	+0.0980 ≅ 1	+0.9734 ≅ 1	+0.4138 ≅ 0	+0.8205 ≅ 1
5.1929 =	4.2195 ≅ 1	+2.0849 ≅ 0	+0.0980 ≅ 0	+0.9734 ≅ 1	+0.4138 ≅ 0	+0.8205 ≅ 0
6.3044 =	4.2195 ≅ 1	+2.0849 ≅ 1	+0.0980 ≅ 0	+0.9734 ≅ 0	+0.4138 ≅ 0	+0.8205 ≅ 0
4.3175 =	4.2195 ≅ 1	+2.0849 ≅ 0	+0.0980 ≅ 1	+0.9734 ≅ 0	+0.4138 ≅ 0	+0.8205 ≅ 0
4.2195 =	4.2195 ≅ 1	+2.0849 ≅ 0	+0.0980 ≅ 0	+0.9734 ≅ 0	+0.4138 ≅ 0	+0.8205 ≅ 0

Die Goodness-of-Fit Teststatistiken weisen mit einem Wert von 0.0 (bei null Freiheitsgraden) darauf hin, daß eine perfekte Übereinstimmung zwischen den (nicht wiedergegebenen) erwarteten und beobachteten Häufigkeiten besteht. Dies ist jedoch kein empirischer Befund sondern eine logische Konsequenz der Tatsache, daß das Modell genau so viele Datenpunkte (=Zeilen der Design-Matrix) aufweist, wie unbekannte Modellparameter (=Spalten der Design-Matrix). Es ist bei solchen *saturierten Modellen* nicht möglich, anhand der Übereinstimmung der vorhergesagten mit den beobachteten Häufigkeiten die Güte des Modells zu beurteilen.

Welche Konsequenzen ergeben sich nun für die inhaltliche Interpretation? Diese Frage kann wiederum über die Berechnung der Vorhersagewerte beantwortet werden (Tabelle 4). Wie im ersten Modell besagt der positive Koeffizient der Design-Variable D_1, daß es verglichen mit der Referenzkategorien "weiß nicht" bei der wahrgenommenen Bürgernähe von Politikern im Durchschnitt deutlich mehr Personen gibt, die Politiker für desinteressiert halten. Der sehr geringe positive Wert von 0.098 für D_2 weist auf der anderen Seite darauf hin, daß verglichen mit den Meinungslosen bei Berücksichtigung von Interaktionseffekten nur mit geringfügig mehr Personen zu rechnen ist, die Politiker für interessiert halten. Dem positiven Effekt für die dritte Design-Variable ist schließlich zu entnehmen, daß es - wie im ersten Modell - im Durchschnitt deutlich mehr Wähler als Nichtwähler gibt.

Dieses Grundmuster wird nun durch die Interaktionseffekte modifiziert. Der erste Interaktionseffekt besagt, daß der Anstieg der Zellenhäufigkeit von der Referenzkategorie der Meinungslosen (V20=3) zu denjenigen, die Politiker für desinteressiert halten (V20=1), in der Gruppe der Wähler (V326=1) deutlich höher ist als in der Gruppe der Nichtwähler (V326=0). Bei den Wählern ist der Zuwachs der logarithmierten Häufigkeiten um den Wert des Interaktionseffekts höher als bei den Nichtwählern, also um 0.4138. Auf der Ebene der

absoluten Häufigkeiten übersteigt der Zuwachs der Wähler den der Nichtwähler daher um den Faktor 1.5126 (=$e^{0.4138}$). Alternativ kann der Interaktionseffekt auch so gelesen werden, daß der Anstieg der Zellenbesetzung beim Wechsel von den Nichtwählern (v326=2) zu den Wählern (V326=1) in der Gruppe derjenigen, die Politiker für desinteressiert halten (V20-=1), um den Faktor 1.5126 höher ist als in der Gruppe derjenigen, die keine Meinung haben (V20=3). Der zweite Interaktionseffekt ist noch höher: Die Besetzungszahlen steigen beim Wechsel von den Meinungslosen (V20=3) zu der Gruppe derjenigen, die Politiker für interessiert halten (V20=2), um den Faktor 2.2716 (=$e^{0.8205}$) stärker an, wenn es sich um Wähler (V326=1) und nicht um Nichtwähler (V326=2) handelt. Oder auf die Wahlbeteiligung bezogen: Beim Wechsel von Nichtwählern zu Wählern ist der Anstieg der Zellenbesetzung sehr viel stärker, wenn statt Meinungslosigkeit die Auffassung vertreten wird, daß Politiker interessiert sind. Zusammengenommen folgt also aus dem Modell, daß das Verhältnis von Wähler zu Nichtwählern bei den Meinungslosen am relativ geringsten und bei denjenigen, die Politiker für interessiert halten, am relativ höchsten ist.

Während die sogenannten *Haupteffekte* eines log-linearen Modells, das sind die Regressionsgewichte der Design-Variablen für die Kategorien der Modellvariablen, nur die unterschiedlichen Besetzungen der Kategorien dieser Variablen widerspiegeln, modellieren die Interaktionseffekte Zusammenhänge zwischen den inhaltlich interessierenden Variablen. Bei mehrdimensionalen Kreuztabellen können auch Interaktionseffekte höherer Ordnung spezifiziert werden. Dazu werden die Design-Variablen von drei oder mehr kategorialen Ausgangsvariablen miteinander multipliziert. Würde etwa bei den Daten aus Tabelle 1 zusätzlich zwischen Befragten aus den alten und den neuen Bundesländern unterschieden, könnte es sich zeigen, daß die Beziehung zwischen der Beurteilung der Interessiertheit von Politikern und der Wahlbeteiligung in den alten und neuen Bundesländern unterschiedlich ist. Im log-linearen Modell gäbe es dann deutliche Interaktionseffekte zwischen der Design-Variable der Wahlbeteiligung, den beiden Design-Variablen der Bürgernähe von Politikern und der Design-Variable für das Erhebungsgebiet. Die Interpretation eines log-linearen Modells mit solchen Interaktionseffekten zweiter oder noch höherer Ordnung wird allerdings schnell unübersichtlich. Wenn es die Daten zulassen, werden daher eher sparsame Modelle mit Interaktionseffekten möglichst geringer Ordnung bevorzugt.

2. Die Spezifikation benutzerdefinierter Design-Matrizen in SPSS

Die SPSS-Prozedur GENLOG verwendet grundsätzlich dummykodierte Design-Variablen bei der Spezifikation eines log-linearen Modells. Referenzkategorie ist die letzte Kategorie einer Ausgangsvariable, also deren numerisch höchster Wert. In der älteren SPSS-Prozedur LOGLINEAR wird dagegen als Voreinstellung Effektkodierung eingesetzt. Um die Unterschiede zwischen den beiden Kodierregeln zu verdeutlichen,soll das Modell aus Abbildung 2 auch mit effektkodierten Design-Variablen geschätzt werden. Um die Koeffizienten eines

log-linearen Modells mit Effektkodierung auch über die Prozedur GENLOG schätzen zu können, müssen die Spalten der Design-Matrix direkt vom Anwender generiert werden. Möglich wird dies durch die in der Prozedur implementierte Option der Schätzung von Effekten metrischer Prädiktoren, den sogenannten Kovariaten. Wird im Prozeduraufruf von GENLOG eine Variable als Kovariate aufgeführt, berechnet GENLOG für alle Zellen der zu analysierenden Tabelle die Mittelwerte der Kovariate über die jeweilige Anzahl der Fälle in den Zellen. Die resultierenden Mittelwerte können dann als zusätzliche Spalte in die Design-Matrix aufgenommen werden.[10] Über die Generierung geeigneter Kovariaten können somit benutzerdefinierte Spalten der Design-Matrix spezifiziert werden. Werden in der /DESIGN-Option ausschließlich benutzergenerierte Kovariaten aufgeführt, wird ein log-lineares Modell geschätzt, dessen Design-Matrix vom Benutzer frei gestaltbar ist. Die einzige Einschränkung besteht darin, daß die Regressionskonstante stets automatisch geschätzt wird.[11]

Für das Beispiel werden zunächst mit RECODE- und COMPUTE-Anweisungen aus den Ausgangsvariablen V20 (Bürgernähe) und V326 (Wahlbeteiligung) effektkodierte Design-Variablen für die Haupt- und Interaktionseffekte generiert:

```
recode V20 (1=1)(2=0)(3=-1) into D1.
recode V20 (1=0)(2=1)(3=-1) into D2.
recode V326 (1=1)(2=-1) into D3.
compute D1D3=D1*D3.
compute D2D3=D2*D3.
```

Die erste RECODE-Anweisung erzeugt für die erste Kategorie von V20 eine effektkodierte Design-Variable D1. Referenzkategorie ist die letzte Ausprägung der Ausgangsvariable, bei der die Design-Variable den Wert '-1' erhält. Auf analoge Weise werden für die zweite Kategorie von V20 und für die erste Kategorie von V326 effektkodierte Design-Variablen D2 und D3 gebildet. Schließlich werden mit den beiden COMPUTEAnweisungen durch einfaches Multiplizieren der gerade gebildeten Design-Variablen zusätzliche Produktvariablen D1D3 und D2D3 für die Schätzung von Interaktionseffekten generiert.

In der Modellspezifikation werden die so erzeugten Design-Variablen als Kovariaten hinter dem Schlüsselwort "WITH" direkt nach der Nennung der die Tabelle definierenden Variablen aufgeführt und in der /DESIGN-Option als Effekte angegeben. Der Prozeduraufruf und die Resultate der Modellschätzung sind in Abbildung 3 wiedergegeben. Der Vergleich mit den entsprechenden Werten des log-linearen Modells bei Dummykodierung (Abb. 2)

10 Eine mögliche Variation zwischen den Werten der Kovariaten in einer Tabellenzelle wird also ignoriert. Die Vorgehensweise ist daher keine Alternative zu Modellen, die - wie die logistische Regression - auf Individualdatenebene kategoriale abhängige Variablen durch metrische Prädiktoren erklären.

11 Außerdem werden redundante Spalten einer Designmatrix automatisch entfernt. Mathematisch ausgedrückt muß die Design-Matrix vollen Spaltenrang haben. Als eine technische Einschränkung ist schließlich die Zahl der möglichen Kovariaten auf 200 beschränkt.

weist auf gänzlich verschiedene Werte hin. Die (im Ausdruck nicht aufgeführten erwarteten Häufigkeiten) und die Goodness-of-Fit Teststatistiken sind jedoch identisch. Tatsächlich handelt es sich bei den beiden Modellen um *Reparametrisierungen* der gleichen Aussagen zur Struktur der analysierten Tabelle. Unterschiede bei Vorhersagewerten können erst dann auftreten, wenn zwei Modelle unterschiedliche Aussagen beinhalten, wie dies z.B. bei den beiden Modellen mit bzw. ohne Interaktionseffekten der Fall ist.

Abbildung 3: Parameterschätzung im Modell mit Dummykodierung

```
-> genlog V326 V20 with D1 D2 D3 D1D3 D2D3
->   /print des est /plot non /crit delta(0)
->   /des D1 D2 D3 D1D3 D2D3 .
Correspondence Between Parameters and Terms of the Design
Parameter   Aliased  Term
        1            Constant
        2            D1
        3            D2
        4            D3
        5            D1D3
        6            D2D3
Design Matrix
                          Cell                    Parameter
Factor        Value  Structure     1        2        3        4        5
V326           ja
 V20             ja       1.000     1    1.000     .000    1.000    1.000
 V20           nein       1.000     1     .000    1.000    1.000     .000
 V20    weiß nicht        1.000     1   -1.000   -1.000    1.000   -1.000
V326          nein
 V20             ja       1.000     1    1.000     .000   -1.000   -1.000
 V20           nein       1.000     1     .000    1.000   -1.000     .000
 V20    weiß nicht        1.000     1   -1.000   -1.000   -1.000    1.000
Design Matrix (continued)
                          Cell    Parameter
Factor        Value  Structure        6
V326           ja
 V20             ja       1.000     .000
 V20           nein       1.000    1.000
 V20    weiß nicht        1.000   -1.000
V326          nein
 V20             ja       1.000     .000
 V20           nein       1.000   -1.000
 V20    weiß nicht        1.000    1.000
Goodness-of-fit Statistics
                    Chi-Square       DF       Sig.
Likelihood Ratio         .0000        0       .
         Pearson         .0000        0       .

Parameter Estimates                              Asymptotic 95% CI
Parameter   Estimate         SE    Z-value      Lower      Upper
        1     5.6396      .0325     173.36       5.58       5.70
        2     1.3585      .0353      38.44       1.29       1.43
        3     -.4251      .0485      -8.76       -.52       -.33
        4      .6924      .0325      21.29        .63        .76
        5      .0012      .0353        .03       -.07        .07
        6      .2046      .0485       4.22        .11        .30
```

Die Regressionskonstante gibt bei Effektkodierung den Durchschnittswert der logaritmierten erwarteten Häufigkeiten wieder. Tabelle 5 verdeutlicht, warum dies so ist. Werden nämlich die Vorhersagegleichungen für alle Tabellenzellen aufsummiert, bleibt auf der rechten Seiten der Gleichung nur die Summe der Regressionskonstanten übrig. Alle anderen Spalten addieren sich stets zum Wert null. Da die Summe der sechs Vorhersagewerte also das sechsfache der Regressionskonstante ergibt, muß diese 1/6 dieser Summe sein.Aus der Tat-

sache, daß die Summe der Werte einer Design-Variable null ist, folgt weiter für die Interpretation des zugehörigen Koeffizienten, daß dieser die Abweichung vom Durchschnittswert erfaßt. Der Koeffizient 1.3585 des Effektes der Design-Variable D_1 für die erste Ausprägung von V20 besagt also, daß die logarithmierten Häufigkeiten der Tabellenzellen, bei der die Variable V20 den Wert '1' aufweist, bei Kontrolle von V326 und der Berücksichtigung der Interaktionseffekte im Mittel um den Wert 1.3585 vom Durchschnittswert aller Tabellenzellen abweichen. Entsprechend weist der nächste Koeffizient -0.4251 darauf hin, daß die Tabellenzellen, die sich auf die zweite Kategorie von V20 beziehen, im Schnitt eine um 0.4251 geringere logarithmierte Häufigkeit aufweisen. Da die Summe aller Abweichungen vom Durchschnitt null ergibt, ist die durchschnittliche Abweichung in der Referenzkategorie das Negative der Summe der übrigen Abweichungen. Die durchschnittliche Abweichung der 'Meinungslosen' bei der Bewertung der Bürgernähe (V20=3) ist also $-0.9334 = 1.3585 \cdot (-1) + -0.4251 \cdot (-1)$.

Tabelle 5: Berechnung der vorhergesagten logarithmierten Häufigkeiten im Modell mit Interaktionseffekten bei Effektkodierung

$\hat{y}$ =	$b_0 \cdot 1$	$+b_1 \cdot D_1$	$+b_2 \cdot D_2$	$+b_3 \cdot D_3$	$+b_4 \cdot (D_1 \cdot D_3)$	$+b_5 \cdot (D_2 \cdot D_3)$
7.6117=	5.6396 · 1	+1.3585 · 1	-0.4251 · 0	+0.6924 · 1	+0.0012 · 1	+0.2046 · 0
6.1115=	5.6396 · 1	+1.3585 · 0	-0.4251 · 1	+0.6924 · 1	+0.0012 · 0	+0.2046 · 1
5.1928=	5.6396 · 1	+1.3585 · -1	-0.4251 · -1	+0.6924 · 1	+0.0012 · -1	+0.2046 · -1
6.3045=	5.6396 · 1	+1.3585 · 1	-0.4251 · 0	+0.6924 · -1	+0.0012 · -1	+0.2046 · 0
4.3175=	5.6396 · 1	+1.3585 · 0	-0.4251 · 1	+0.6924 · -1	+0.0012 · 0	+0.2046 · -1
4.2196=	5.6396 · 1	+1.3585 · -1	-0.4251 · -1	+0.6924 · -1	+0.0012 · 1	+0.2046 · 1

Die beiden Haupteffekte der Bürgernähe von Politikern besagen also, daß es überdurchschnittliche viele Personen gibt, die Politiker für nicht interessiert halten, daß es dagegen unterdurchschnittlich viele Personen gibt, die Politiker für interessiert halten, und daß es noch weniger Personen gibt, die gar keine Meinung zu diesem Thema haben. Analog ergibt sich als Interpretation des Koeffizienten für die Wahlbeteiligung, daß in der analysierten Tabelle die logarithmierten Zellenhäufigkeiten um 0.6924 über dem Durchschnitt liegen, wenn es sich um Wähler handelt (V326=1). Umgekehrt liegen die logarithmierten Häufigkeiten bei Nichtwählern im Durchschnitt um -0.6924 unter dem Durchschnitt aller Zellen.

Die Interaktionseffekte modifizieren wiederum diese Grundaussage.Der sehr geringe Koeffizient von 0.0012 besagt, daß bei Zustimmung zur These der Desinteressiertheit von Politikern (V20=1) die Zahl der Wähler nur ganz geringfügig über der durchschnittlichen Zahl

von Wählern in den Tabellenzellen liegt und entsprechend die Zahl der Nichtwähler nur geringfügig unter dem Durchschnitt der Nichtwähler. Der letzte Koeffizient von 0.2046 besagt entsprechend, daß die Zahl der Wähler, die Politiker eher für interessiert halten (V20=2), überdurchschnittlich hoch ist und die Zahl der Nichtwähler, die Politiker für desinteressiert halten, dann deutlich unter dem Durchschnitt liegt. Für die Gruppe der Meinungslosen berechnet sich die Abweichung vom Durchschnitt der Wähler durch Aufsummieren der negativen Werte der beiden Interaktionseffekte: der negative Wert von -0.2058 (=-0.012 + -0.2046) weist darauf hin, daß es in dieser Gruppe besonders wenige Wähler und dafür umgekehrt besonders viele Nichtwähler gibt.

Bei Effektkodierung werden die Koeffizienten also ähnlich interpretiert wie bei der Dummykodierung. Der Unterschied liegt allein darin, daß bei der Effektkodierung jeweils Abweichungen vom Durchschnittswert gemessen werden, bei der Dummykodierung dagegen Abweichungen von einer Referenzgruppe. Das Endergebnis ist aber dasselbe. Beide Modelle besagen, daß es unter denjenigen, die die Politiker für nicht desinteressiert halten, besonders viele Wähler gibt und bei den Meinungslosen besonders wenige. Oder auch umgekehrt: unter den Wählern gibt es besonders viele Personen, die Politiker für nicht desinteressiert halten und besonders wenige, die keine Meinung zu diesem Thema haben.

Trotz der gleichen empirischen Behauptungen der Modelle gibt es bei den geschätzten Koeffizienten erhebliche Unterschiede. Im Modell mit Dummykodierung (Abbildung 2) sind beide Interaktionseffekte deutlich von null verschieden, der zweite Haupteffekt der Einschätzung des Desinteresses von Politikern aber praktisch vernachlässigbar. Beim Modell mit Effektkodierung sind dagegen alle Haupteffekte deutlich von null verschieden; dafür ist aber der erste Interaktionseffekt sehr klein. Ursache dieser Differenzen ist wiederum der unterschiedliche Bezugspunkt. Relativ zur Gruppe der Meinungslosen (Dummykodierung) gibt es bei Kontrolle der übrigen Effekte im Durchschnitt kaum mehr Personen, die Politiker nicht für desinteressiert halten. Verglichen zum Gesamtdurchschnitt (Effektkodierung) ist diese Zahl dagegen deutlich geringer. Umgekehrt ist es bei den Interaktionseffekten. Relativ zur Zahl der Wähler unter den Meinungslosen gibt es relativ mehr Wähler in der Gruppe, die Politiker für desinteressiert halten. Die Abweichung vom Durchschnitt aller Wähler ist dagegen nur gering. Inhaltliche Bedeutung bekommen die unterschiedlichen Sichtweisen erst, wenn versucht wird, sparsamere log-lineare Modelle zu schätzen, bei denen nicht signifikante Koeffizienten ausgelassen werden.

Ich erwähnte bereits, daß der erste Interaktionseffekt mit einem Wert von 0.0012 sehr klein ist. Der Koeffizient ist auch bei einer Irrtumswahrscheinlichkeit von 5% nicht signifikant von null verschieden. Sichtbar wird letzteres an den kleinen Z-Werten oder daran, daß die von SPSS ausgedruckten asymptotischen Konfidenzintervalle den Wert null einschließen. Bei der ML-Schätzung sind die geschätzten Koeffizienten asymptotisch normalverteilt. Der

Quotient aus Schätzung und Standardfehler (Z-Wert) kann daher als Teststatistik der Nullhypothese herangezogen werden, daß ein Koeffizient in der Grundgesamtheit null ist. Ist die Nullhypothese richtig, ist der Z-Wert bei größeren Fallzahlen in etwa normalverteilt. Ein Koeffizient ist also in einem zweiseitigen Test mit einer Irrtumswahrscheinlichkeit von etwa 5% signifikant von null verschieden, wenn sein Z-Wert größer +2 oder kleiner -2 ist.

Es liegt nun nahe, ein log-lineares Modell zu spezifizieren, bei dem nur die signifikanten Parameter berücksichtigt werden. Dazu wird einfach in der /DESIGN-Option die entsprechende Kovariate ausgelassen. Aufruf und Ergebnis der Modellschätzung sind in Abbildung 4 festgehalten. Die geringen Chiquadrat-Werte weisen darauf hin, daß die (nicht ausgedruckten) beobachteten und erwarteten Häufigkeiten sehr dicht beieinander liegen. Das Modell kann also die wesentlichen Aspekte der tabellierten Daten gut wiedergeben. Der Vergleich der Koeffizienten in den Abbildungen 3 und 4 zeigt weiter, daß sich die jeweils entsprechenden Koeffizienten kaum unterscheiden. Damit bleibt auch die Interpretation im wesentlich unverändert: Es gibt überdurchschnittlich viele Befragte (b_1=1.359), die Politiker für desinteressiert halten (V20=1) und unterdurchschnittlich viele Personen (b_2=-0.496), die Politiker für interessiert halten (V20=2). Die Zahl der Meinungslosen (V20=3) weicht im Durchschnitt noch stärker nach unten vom Gesamtmittel ab (- 0.863=-b_1-b_2). Bei der Wahlbeteiligung ist die Zahl der Wähler (V326=1) überdurchschnittlich (b_3=0.693), die Zahl der Nichtwähler (V326=2) entsprechend unterdurchschnittlich (-0.693=-b_3) hoch. Schließlich gibt es unter den Wählern überdurchschnittlich viele Befragte, die Politiker nicht für desinteressiert halten (b_5=0.205) und unterdurchschnittlich viele, die keine Meinung haben (-0.205=-b_5).

Das Interessante an dem Modell ist, daß der überdurchschnittliche Anstieg von Wählern in der Gruppe derjenigen, die Politiker nicht für desinteressiert halten, gerade genauso groß ist wie der überdurchschnittliche Rückgang bei den Meinungslosen. In gewisser Hinsicht wird dadurch für den Zusammenhang von Bürgernähe und Wahlbeteiligung eine Metrik definiert, bei der die Meinungslosen den einen Pol bilden und Personen, die Politiker für interessiert halten, den anderen Pol. Die mittlere Position wird von Personen eingenommen, die Politiker für desinteressiert halten.[12]

Die Rangordnung bzw. Metrik der Bürgernähe (V20) kann auch bei Effektkodierung modelliert werden. Dazu wird eine zusätzliche Design-Variable INT gebildet:

```
recode V20 (1=1)(2=2)(3=0) into INT.
if(V326=2) INT=0
```

12 Diese Art von Beziehung wird nach ***L. A. Goodmann*** als log-lineares Assoziationsmodell bezeichnet und für Modelle vorgeschlagen, bei denen eine Variable ordinales oder metrisches Skalenniveau aufweist. Das Modell kann gleichzeitig auch als log-lineare Darstellung des ordinalen Logitmodells der benachbarten Kategorien aufgefaßt werden.

Abbildung 4: Schätzung des Modells mit nur einem Interaktionseffekt bei Effektkodierung

```
-> genlog V326 V20 with D1 D2 D3 D1D3 D2D3
->   /print des est /plot non /crit delta(0)
->   /des D1 D2 D3 D2D3.

Correspondence Between Parameters and Terms of the Design
Parameter   Aliased  Term
        1            Constant
        2            D1
        3            D2
        4            D3
        5            D2D3

Design Matrix
                          Cell                    Parameter
Factor        Value  Structure     1        2        3        4        5
V326           ja
 V20            ja       1.000     1    1.000     .000    1.000     .000
 V20           nein      1.000     1     .000    1.000    1.000    1.000
 V20    weiß nicht       1.000     1   -1.000   -1.000    1.000   -1.000
V326          nein
 V20            ja       1.000     1    1.000     .000   -1.000     .000
 V20           nein      1.000     1     .000    1.000   -1.000   -1.000
 V20    weiß nicht       1.000     1   -1.000   -1.000   -1.000    1.000

Goodness-of-fit Statistics
                    Chi-Square       DF       Sig.
Likelihood Ratio         .0011        1      .9737
         Pearson         .0011        1      .9737

Parameter Estimates
                                                 Asymptotic 95% CI
Parameter   Estimate         SE    Z-value      Lower      Upper
        1     5.6391      .0295     191.32       5.58       5.70
        2     1.3591      .0290      46.93       1.30       1.42
        3     -.4255      .0469      -9.07       -.52       -.33
        4      .6932      .0213      32.50        .65        .74
        5      .2049      .0470       4.36        .11        .30
```

Den drei Kategorien der Bürgernähe (V20) werden hier entsprechend ihrer Rangfolge bei der Höhe der Wahlbeteiligung (V326) die Werte '1' (V20=1), '2' (V20=2) und '0' (V20=3) zugeordnet. Da es sich um einen Interaktionseffekt handelt, gelten diese Scores nur in der Gruppe der Wähler. Bei Nichtwählern weist die Variable INT den Wert '0' auf.

Die Modellspezifikation und Teile der SPSS-Ausgabe sind in Abbildung 5 wiedergegeben. Die Haupteffekte basieren auf dummykodierten Design-Variablen. Daher ist es bei diesem Modell nicht notwendig, die Design-Variablen für die Haupteffekte als Kovariaten zu spezifizieren. Die Gleichheit der Chiquadrat-Werte bzw. der (in der Abbildung nicht wiederge-gebenen) erwarteten Häufigkeiten weisen darauf hin, daß es sich wie bei den saturierten Modellen aus Abbildung 2 und 3 um eine Reparametrisierung des Modells mit Effektkodierung handelt. Tatsächlich läßt sich aus der Design-Matrix ablesen, daß das Modell postuliert, daß der (logarithmierte) Anstieg der Wähler beim Wechsel von den Meinungslosen zu denjenigen, die Politiker für desinteressiert halten, halb so groß ist, wie der Anstieg von den Meinungslosen zu denjenigen, die Politiker für nicht desinteressiert halten.Ausgehend von denjenigen, die Politiker für desinteressiert halten, entspricht also der Anstieg zu denjenigen,die Politiker nicht für desinteressiert halten,dem Rückgang bei den Meinungslosen.

Abbildung 5: Modellierung metrischer Interaktion bei Dummykodierung

```
-> genlog V326 V20 with INT
->   /print des est /plot non /crit delta(0)
->   /des V20 V326 INT.

Correspondence Between Parameters and Terms of the Design
Parameter   Aliased  Term
        1            Constant
        2            [V20 = 1.00]
        3            [V20 = 2.00]
        4       x    [V20 = 3.00]
        5            [V326 = 1.00]
        6       x    [V326 = 2.00]
        7            INT

Design Matrix
                          Cell                Parameter
Factor        Value  Structure     1     2     3     5        7
V326           ja
 V20            ja       1.000     1     1     0     1    1.000
 V20          nein       1.000     1     0     1     1    2.000
 V20    weiß nicht       1.000     1     0     0     1     .000
V326         nein
 V20            ja       1.000     1     1     0     0     .000
 V20          nein       1.000     1     0     1     0     .000
 V20    weiß nicht       1.000     1     0     0     0     .000

Goodness-of-fit Statistics
                   Chi-Square       DF       Sig.
Likelihood Ratio         .0011        1      .9737
         Pearson         .0011        1      .9737

Parameter Estimates
                                                 Asymptotic 95% CI
Parameter   Estimate         SE    Z-value      Lower      Upper
        1     4.2172      .0996      42.36       4.02       4.41
        2     2.0878      .0951      21.95       1.90       2.27
        3      .0982      .1675        .59       -.23        .43
        4      .0000          .          .          .          .
        5      .9766      .1056       9.25        .77       1.18
        6      .0000          .          .          .          .
        7      .4099      .0940       4.36        .23        .59
```

Die mögliche Einsparung eines Koeffizienten läßt sich bereits in Abbildung 2 erkennen. Der zweite Interaktionseffekt ist dort nämlich mit einem Wert von 0.8205 ziemlich genau das Doppelte des Wertes 0.4138 des ersten Interaktionseffekts. Wird nun in der Design-Matrix für das Modell aus Abbildung 2 das zweifache der letzten Spalte zur vorletzten Spalte addiert und die letzte Spalte anschließend gelöscht, ergibt sich gerade die Design-Matrix für das Modell aus Abbildung 5. Zur Verdeutlichung sind in Tabelle 6 noch einmal beide Design-Matrizen direkt untereinander geschrieben.

Die Design-Matrix für das Modell aus Abbildung 5 läßt sich daher auch als Spezifikation einer linearen Restriktion über die Koeffizienten der Design-Matrix aus Abbildung 2 verstehen, bei der der letzte Parameter aus dem Modell in Abbildung 2 auf das Doppelte des vorletzten Parameters festgesetzt wird. Tatsächlich lassen sich mit benutzerdefinierten Design-Matrizen beliebige lineare Restriktionen über die Parameter eines log-linearen Modells spezifizieren und schätzen. Die Logik entspricht der an anderer Stelle vorgestellten Logik der Spezifikation sparsamer logistischer Modelle (***Kühnel***, 1992).

Tabelle 6: Zusammenfassung von Spalten der Design-Matrix

a. Designmatrix des saturierten Ausgangsmodells (Abbildung 2)

V326	V20	'1'	D_1	D_2	D_3	$D_1 \cdot D_3$	$D_2 \cdot D_3$
1	1	1	1	0	1	1	0
1	2	1	0	1	1	0	1
1	3	1	0	0	1	0	0
2	1	1	1	0	0	0	0
2	2	1	0	1	0	0	0
2	3	1	0	0	0	0	0

b. Designmatrix des nicht saturierten Modells (Abbildung 5)

V326	V20	'1'	D_1	D_2	D_3	$(D_1 \cdot D_3) + 2 \cdot (D_2 \cdot D_3)$
1	1	1	1	0	1	1
1	2	1	0	1	1	2
1	3	1	0	0	1	0
0	1	1	1	0	0	0
0	2	1	0	1	0	0
0	3	1	0	0	0	0

Ausgehend von den Parameterschätzungen (Abbildung 5) läßt sich das Modell noch weiter vereinfachen. Der zweite Haupteffekt für V20 (D_2) ist nämlich wie bereits im saturierten Modell aus Abbildung 2 bei einer Irrtumswahrscheinlichkeit von 5% nicht signifikant von null verschieden. Es sollte daher möglich sein, auch diesen Koeffizienten auf null zu setzen. Dies kann über die Spezifikation einer benutzerdefinierten Design-Matrix erfolgen, bei der auch die Design-Variablen für die Haupteffekte als Kovariaten eingehen. Der verbleibende erste Haupteffekt von V20 läßt sich aber auch in der /DESIGN-Option durch die Angabe der Kategoriennummer in einer Klammer nach dem Variablennamen ansprechen:

```
genlog V326 V20 with INT
    /print freq des est /plot none /crit delta (0)
    /design V20(1) V326 INT .
```

Das Modell postuliert gegenüber dem Modell aus Abbildung 5 zusätzlich, daß im Durchschnitt bei Kontrolle der unterschiedlichen Häufigkeiten der Wähler und Nichtwähler und des Zusammenhangs zwischen Wahlbeteiligung und Bürgernähe von Politikern die Häufigkeiten der Meinungslosen (V20=3) sich nicht von den Häufigkeiten derjenigen unterscheiden, die Politiker nicht für desinteressiert halten (V20=2). Abweichungen gegenüber diesen beiden Kategorien weist nur die erste Kategorie auf, für die daher eine dummykodierte De-

sign-Variable spezifiziert ist. Die Schätzung des Modells gibt bei zwei Freiheitsgraden einen Chiquadrat-Wert von 0.34 für die Likelihood-Ratio Teststatistik und einen Wert von 0.35 für Pearsons Teststatistik. Auch dieses Modell paßt also gut zu den Daten. Da sich die zusätzliche Einsparung eines Koeffizienten aber nur auf die Randverteilung einer Variable (V20) bezieht und nicht auf die Beziehung zwischen den Variablen, ist das Einsparen eines Parameters inhaltlich kaum interessant. Auf eine Wiedergabe der geschätzten Koeffizienten sei daher hier verzichtet.

3. Schlußbemerkung

Oftmals steht der Anwendung log-linearer Modelle bei der Analyse kategorialer Daten die Vorstellung entgegen, daß diese Modelle recht kompliziert und kaum zu interpretieren seien. Die Konzeption log-linearer Modelle als Regressionsmodelle für logarithmierte Häufigkeiten kann m.E. die Interpretation erleichtern. Der wesentliche Unterschied zur linearen Regression mit nominalskalierten unabhängigen Variablen bzw. zur Varianzanalyse besteht dann allein darin, daß Zusammenhänge zwischen zwei Variablen nicht durch Haupteffekte, sondern durch Interaktionseffekte modelliert werden. Die Sichtweise als Regressionsmodell erleichtert auch das Verständnis der Design-Matrix eines log-linearen Modells, die dann der üblichen Datenmatrix der Prädiktoren in der linearen Regression entspricht.

Durch die Verwendung benutzerdefinierter Design-Matrizen ergibt sich eine hohe Flexibilität bei der Modellspezifikation. Es ist z.B. möglich, sehr sparsame Modelle zu schätzen, die keine "überflüssigen" Parameter enthalten. Inferenzstatistisch gesehen erhöht dies die Teststärke bei Hypothesenprüfungen. Wichtiger ist aber, daß die Möglichkeiten benutzerdefinierter Design-Matrizen dazu genutzt werden können, Modelle zu spezifizieren, die für die Untersuchung der jeweils interessierenden inhaltlichen Fragestellung angemessener sind als Standardmodelle. Als einfaches Beispiel wurde oben die Beziehung zwischen der Bürgernähe von Politikern und der Wahlbeteiligung als eine quasi-metrische Beziehung aufgefaßt, die über einen einzigen Koeffizienten (Interaktionseffekt) erfaßt werden kann.

Voraussetzung für die Nutzung dieser Möglichkeiten log-linearer Analysen ist eine Software, die den Anwendern den direkten Zugriff auf die Design-Matrix erlaubt. In der SPSS-Prozedur GENLOG ist dies über die Spezifikation von Kovariaten möglich. Die gleiche Technik läßt sich aber auch bei Nutzung der älteren Prozedur LOGLINEAR anwenden, bei der außerdem benutzerspezifische Design-Matrizen direkt spezifiziert werden können (siehe Anhang). Die einzige Einschränkung in SPSS besteht darin, daß es in beiden Prozeduren nicht möglich ist, Modelle ohne Konstante zu schätzen. Ansonsten stehen aber auch mit diesen Standardprozeduren alle Möglichkeiten der log-linearen Analyse mit benutzerdefinierten Design-Matrizen zur Verfügung.

Literatur:

***Andreß*, *H.-J*., *Hagenaars*, *J*.** und ***Kühnel*, *S.M*.** (1997)
Analyse von Tabellen und kategorialen Daten. Berlin u.a.: Springer.

Kühnel*, *S.M. (1992)
Sparsame Modellierung mit logistischen Zufallsnutzenmodellen. ZA- Information 31, S. 70-92.

***Norusis*, *M.J*.** und SPSS Inc. (1994)
SPSS Advanced Statistics 6.1. Chicago: SPSS Inc.

Zentralarchiv (1996): Allgemeine Bevölkerungsumfrage der Sozialwissenschaften. ALLBUS 1996. Codebuch, ZA-Nr. 2800. Köln, ZA.

ANHANG:

Das Einlesen von Tabellendaten in SPSS und die Spezifikation benutzerdefinierter Design-Matrizen in der Prozedur LOGLINEAR

Log-lineare Modelle basieren auf den Besetzungen von Tabellenzellen. Solche Informationen können ohne großen Aufwand in SPSS eingelesen werden. Um mit der Prozedur GENLOG oder LOGLINEAR die Daten aus Tabelle 1 zu analysieren, können folgende SPSS-Anweisungen verwendet werden:

```
data list free / V326 V20 N.
begin data
1 1 2190   1 2 451   1 3 180   2 1 547   2 2 75   2 3 68
end data.
formats V20 V326 (f1.0).
var lab V20 'Politiker sind nicht an Problemen interessiert'
  /V326'Wahlbeteiligung letzte Bundestagswahl'.
val lab V20 V326 1'ja' 2'nein' 3'weiß nicht'.
weight by N.
```

In der Anweisung "data list" werden die beiden Modellvariablen V326 und V20 aufgeführt, die die zu analysierende Tabelle aufspannen. Die zusätzliche Variable N wird für die Eingabe der Besetzungszahlen benötigt. Die Dateneingabe erfolgt zwischen den Schlüsselwörtern "begin data" und "end data.". Für jede Tabellenzelle werden die jeweiligen Werte der Modellvariablen sowie die Besetzungzahl aufgeführt. Die drei nachfolgenden Anweisungen "formats", "var lab" und "val lab" definieren das Ausgabeformat für die Modellvariablen und Variablen- und Wertetiketten. Sie sind nicht notwendig, sondern dienen allein der Übersichtlichkeit der späteren SPSS-Ausgaben. Wichtig ist die Anweisung "weight by N.", die dafür sorgt, daß die Datenanalyse tatsächlich auf den korrekten Fallzahlen beruht.

Für die Schätzung der log-linearen Modelle muß nicht notwendigerweise die Prozedur GENLOG verwendet werden. Alle Analysen können auch mit der Prozedur LOGLINEAR ausgeführt werden, die bereits in älteren SPSS-Versionen und in SPSS/PC verfügbar ist. In

den neueren Versionen von SPSS für Windows kann die Prozedur LOGLINEAR ausschließlich über ein Syntax-Fenster angefordert werden. Der wesentliche Unterschied zu GENLOG besteht darin, daß LOGLINEAR als Voreinstellung effektkodierte Design-Variablen verwendet. Soll beispielsweise das in Abbildung 3 wiedergegebene saturierte Modell mit LOGLINEAR geschätzt werden, kann dies folgendermaßen realisiert werden:

```
loglinear V326(1 2) V20(1 3)
  /print design freq est
  /crit delta(0)
  /design V20 V326 V326*V20 .
```

Wie bei der Prozedur GENLOG folgt nach dem Prozedurnamen die Angabe der Variablen, die die zu analysierende mehrdimensionale Tabelle definieren. Im Unterschied zur neueren Prozedur muß hinter den Variablennamen der kleinste und der größte zu berücksichtigende Ausprägungswert angegeben werden. Die Optionen "/PRINT", "/CRIT" und "/DESIGN" entsprechen den gleichnamigen Optionen von GENLOG. Die Druckerausgabe ist etwas unübersichtlicher, enthält aber im wesentlichen die gleichen Informationen wie die Ausgabe von GENLOG. Um die Zuordnung der geschätzten Koeffizienten zu den Design-Variablen zu erleichtern, empfiehlt sich stets die Angabe des Schlüsselworts "design" in der /PRINT-Option.

Es ist zu beachten, daß die Prozedur LOGLINEAR sowohl bei der Wiedergabe der Design-Matrix wie auch bei der Ausgabe der geschätzten Koeffizienten die Konstante ignoriert, obwohl sie tatsächlich aus den Daten berechnet wird. Die Ursache für das Fehlen der Regressionskonstante liegt darin, daß die ML-Schätzung in der Prozedur LOGLINEAR von einer Multinomialverteilung der Häufigkeiten in den Zellen der Ausgangstabelle ausgeht. Dies hat die Konsequenz, daß die Fallzahl der zu analysierenden Tabelle ein fest vorgegebener Modellparameter ist, der nicht mehr aus den Daten geschätzt werden braucht. Dieser vorgegebene Modellparameter wird bei der Effektkodierung des log-linearen Modells durch die Regressionskonstante modelliert. Die Prozedur GENLOG unterstellt dagegen als Voreinstellung Poisson-Verteilungen für die einzelnen Tabellenzellen und betrachtet daher die Regressionskonstante nicht als vorgegebene Größe, sondern als zu schätzenden Parameter.

Wird bei der Nutzung von LOGLINEAR der Wert der Konstanten benötigt, muß dieser per Hand aus den erwarteten Häufigkeiten berechnet werden. Über die Design-Matrix läßt sich die jeweilige Rechenformel ableiten. Sind alle Design-Variablen effektkodiert, ist der Wert der Regressionskonstante stets das arithmetische Mittel der logarithmierten erwarteten Häufigkeiten. Bei dummykodierten Design-Variablen ist die Konstante der Logarithmus der erwarteten Häufigkeit der Zelle, bei der alle übrigen Design-Variablen den Wert null aufweisen.

Sollen benutzerdefinierte Design-Matrizen spezifiziert werden, ist die gleiche Vorgehensweise möglich, die auch bei der Prozedur GENLOG angewendet wird. Die gewünschten Design-Variablen sind explizit über COMPUTE-, RECODE- und/oder IF-Anweisungen zu generieren und anschließend als Kovariaten in das Modell aufzunehmen. Als Beispiel habe ich im folgenden die SPSS-Anweisungen für die Schätzung der log-linearen Modelle mit Dummykodierung aus den Abbildungen 1, 2 und 5 über die Prozedur LOGLINEAR aufgelistet. Mit den ersten drei RECODE-Anweisungen werden zunächst die drei dummykodierten Design-Variablen D1, D2 und D3 für die Haupteffekte von V20 und V326 gebildet. Die beiden nachfolgenden COMPUTE-Anweisungen generieren die Interaktionseffekte. Die letzte COMPUTE-Anweisung erzeugt die Variable INT für den Interaktionseffekt aus dem in Abbildung 5 wiedergegebenen Modell. Anschließend folgt die Spezifikation der einzelnen Modelle:

```
recode V20(1=1)(2,3=0) into D1.
recode V20(1,3=0)(2=1) into D2.
recode V326(1=1)(2=0) into D3.
compute D1D3=D1*D3.
compute D2D3=D2*D3.
compute INT=D1D3+2*D2D3.
* Modell aus Abbildung 1.
loglinear V326(1 2)  V20( 1 3) with D1 D2 D3
  /print des est /crit delta(0)
  /des D1 D2 D3.
* Modell aus Abbildung 2.
loglinear V326(1 2)  V20( 1 3) with D1 D2 D3 D1D3 D2D3
  /print des est /crit delta(0)
  /des D1 D2 D3 D1D3 D2D3.
* Modell aus Abbildung  5.
loglinear V326(1 2) V20(1 3) with D1 D2 D3 INT
  /print des est /crit delta(0)
  /des D1 D2 D3 INT.
```

Neben der Realisierung über Kovariaten erlaubt die Prozedur LOGLINEAR auch die direkte Spezifikation von Design-Matrizen in der in GENLOG nicht verfügbaren Option "/contrast". Mit dieser Option können u.a. für jede Modellvariable benutzerspezifizierte Design-Variablen gebildet werden. Ausgangspunkt ist jeweils eine quadratische Matrix, die so viele Zeilen und Spalten enthält,wie die Variable,für die DesignVariablen spezifiziert werden, Ausprägungen hat. Die Einschätzung des Desinteresses von Politikern hat drei Ausprägungen. Die Matrix für benutzerspezifizierte Design-Variablen hat entsprechend drei Zeilen mit jeweils drei Spalten. Jede Zeile steht für eine zu spezifizierende Design-Variable und jede Spalte für eine Ausprägung der Ausgangsvariable. Der Aufbau der Matrix entspricht also der Definition von Design-Variablen in Tabelle 2. Es wird allerdings eine Zeile mehr spezifiziert. Da bei einer Variable mit K Ausprägungen aber nur maximale K-1 unabhängige Design-Variablen berücksichtigt werden, sind tatsächlich nur die letzten K-1 Zeilen der Matrix relevant. Die erste Zeile steht gewissermaßen für die Regressionskonstante und

sollte daher die Werte '1' aufweisen. Sollen also für die drei Kategorien der Variable V20 zwei effektkodierte Design-Variablen gebildet werden, geschieht das über die Option "/contrast" durch folgende Spezifikation:

```
/contrast (V20)=special ( 1  1  1
                          1  0  0
                          0  1  0  )
```

Nach dem Schlüsselwort "/CONTRAST" folgt in Klammern der Name der Variable, für die benutzerspezifizierte Design-Variablen erzeugt werden sollen. Nach einem Gleichheitszeichen wird durch das Schlüsselwort 'SPECIAL' angezeigt, daß - wiederum in Klammern - die Angabe einer quadratischen Matrix mit der Definition der Design-Variablen folgt. Die erste Zeile der Matrix enthält die drei Einsen für die Regressionskonstante. Es folgen die Werte der ersten Design-Variable. Da das erste Element der Zeile den Wert '1' aufweist und die beiden übrigen Elemente jeweils den Wert '0', bedeutet dies, daß die Design-Variable bei der ersten Ausprägung von V20 den Wert '1' hat und bei den übrigen Ausprägungen den Wert null. Die Zeile definiert also eine dummykodierte Designvariable für V20=1. Die letzte Zeile enthält die Definition für die zweite Design-Variable. Das Muster "0 1 0" bewirkt, daß die Design-Variable bei der zweiten Ausprägung von V20 den Wert '1' aufweist, ansonsten den Wert '0'.

Die Elemente der quadratischen Matrix können auch hintereinander in eine Reihe geschrieben werden. Außerdem können aufeinanderfolgende Wiederholungen abgekürzt werden. Anstelle der Zeichenfolge '1 1 1' kann also die Abkürzung '3*1' geschrieben werden. Wie in SPSS üblich, können die Schlüsselwörter auch bis auf drei Zeichen abgekürzt werden. Für jede Variable kann ein eigenes /CONTRAST-Statement spezifiziert werden. Soll also das ursprünglich mit der Prozedur GENLOG geschätzte log-lineare Modell ohne Interaktionseffekte aus Abbildung 1 mit der Prozedur LOGLINEAR und der Definition von Kontrasten geschätzt werden, kann dies im Syntax- Fenster mit folgender Anweisung realisiert werden:

```
loglinear V326(1 2) V20(1 3)
  /print freq des est /crit delta(0)
  /cont (V20)=spec(1 1 1  1 0 0  0 1 0)
  /cont (V326)=spec(1 1  1 0)
  /design V20 V326 .
```

Die /CONTRAST-Option der Prozedur LOGLINEAR erlaubt zunächst nur die Definition von Design-Matrizen jeweils einer einzigen Variable, aber nicht die Definition von Interaktionseffekten zwischen mehreren Variablen. Mit einem kleinen Trick kann diese Einschränkung aber umgangen werden. Der Trick besteht darin, alle Ausprägungskombinationen der zu analysierenden Tabelle als Kategorien einer einzigen Modellvariablen zu definieren. Dann lassen sich alle Effekte einschließlich der Interaktionseffekte in einer einzigen /CONTRAST-

Anweisung spezifizieren. Um die sechs Zellen von Tabelle 1 als Ausprägungen einer neuen Variable X zu definieren, können folgende SPSS-Anweisungen formuliert werden:

```
compute X=V326*10+V20.
recode X(11=1)(12=2)(13=3)(21=4)(22=5)(23=6).
var lab X'Kombination von V326 und V20'.
val lab x 1'1 1' 2'1 2' 3'1 3' 4'2 1' 5'2 2' 6'2 3'.
```

Es ist anschließend möglich, mit der SPSS-Prozedur LOGLINEAR Modelle für die Variable X zu schätzen. Mit der Option "/CONTRAST" können dabei für X beliebige Design-Variablen spezifiziert werden. Da X sechs Ausprägungen hat, hat die benutzerspezifizierte quadratische Matrix 6 Zeilen mit je sechs Spalten. Diese Matrix ist dann aber gerade die transponierte, d.h. um 90° gekippte, Design-Matrix für ein saturiertes log-lineares Modell. Durch Nullsetzen von Zeilen dieser Matrix können sparsamere Modelle spezifiziert werden. Um beispielsweise für X das Modell aus Abbildung 1 zu schätzen, wird folgende /CONTRAST-Anweisung benutzt:

```
/contrast (X)=special   ( 1  1  1  1  1  1
                          1  0  0  1  0  0
                          0  1  0  0  1  0
                          1  1  1  0  0  0
                          0  0  0  0  0  0
                          0  0  0  0  0  0)
```

Die erste Zeile der Matrix steht für die Regressionskonstante. In den nächsten drei Zeilen werden die Design-Variablen spezifiziert. Die letzten beiden Zeilen sind auf null gesetzt. Wird diese /CONTRAST-Option beim Aufruf von LOGLINEAR verwendet, wird wiederum das Modell aus Abbildung 1 geschätzt:

```
loglinear X(1 6)
  /print freq des est
  /cont (X)=spec(6*1  1 0 0 1 0 0  0 1 0 0 1 0  1 1 1 0 0 0 12*0 )
  /crit delta(0) /design X .
```

Die letzten beiden Spalten der Design-Matrix enthalten nur Nullen. Die diesen Spalten zugeordneten Koeffizienten erhalten dann automatisch ebenfalls den Wert null. Deren Standardfehler und darauf aufbauende Statistiken werden nicht berechnet. Außerdem werden die Spalten bei der Berechnung der Freiheitsgrade für die Goodness-of-Fit Statistiken nicht berücksichtigt.