Modellierung von Responsefunktionen mit Hilfe der SAS-Prozedur CATMOD

Umsetzung der odds ratio und Kappa als abhängige Variable im GSK-Ansatz (lineare kategoriale Regression)[1]

Wolfgang Rüffer

Institut für Soziologie • Universität Heidelberg

Die statistischen Maßzahlen odds ratio und Kappa eignen sich besonders gut für die Quantifizierung randverteilungsunabhängiger Aussagen über die Stärke der Assoziation von kategorialen Merkmalen. Eine weitere nützliche Eigenschaft dieser Maßzahlen ist ihre (inhaltliche) Interpretationsmöglichkeit (z.B. im Fall der odds ratio als Barriere, respektive Chance, im Fall von Kappa als Maß an Übereinstimmung). Es besteht die Möglichkeit, diese Assoziationsmaße selbst als abhängige Variable in einer linearen kategorialen Regression zu modellieren. Die SAS-Prozedur CATMOD läßt neben den inplementierten Standardresponsefunktionen auch das Modellieren eigener Responsefunktionen zu. Am Beispiel der odds ratio und Kappa soll die Umsetzung in eine Responsefunktion (Matrixschreibweise) gezeigt werden.

1. Einleitung

Innerhalb des STAT-Moduls bietet SAS die zur Analyse kategorialer Daten besonders gut geeignete Prozedur CATMOD an. Das Problem der adäquaten Analyse kategorialer Daten stellt sich gerade im sozialwissenschaftlichen Forschungskontext besonders häufig, da die dort untersuchten Merkmale bzw. ihre Ausprägungen oftmals nur nominal skaliert sind. Besondere Schätzprobleme ergeben sich vor allem, wenn die Zielvariablen selbst nur in kategorialer Form vorliegen. Bekannte statistische Verfahren zur Analyse solcher Zielvariablen sind zum Beispiel die logistische Regression, log-lineare Modelle und die lineare kategoriale Regression. Einzig im Fall der kategorialen Regression ist das Vorhandensein einer kategorialen Zielvariable Anwendungsbedingung. Der Ansatz der linearen kategorialen Regression, welcher in wesentlichen Teilen auf Grizzle/Starmer/Koch zurückgeht und deshalb auch oft als GSK-Ansatz bezeichnet wird, steht im Mittelpunkt dieses Beitrags (Grizzle et al. 1969). Genaugenommen basiert der GSK-Ansatz auf einer Anwendung des allgemeinen linearen Modells auf kategoriale Daten (vgl. Andreß et al. 1997: 55; Fahrmeir et al. 1996). Wesentliche Vorteile der linearen kategorialen Regression sind die Modellierfähigkeit der Responsefunktion und die leichte Interpretierbarkeit der Koeffizienten[2] (Andreß et al. 1997; Hamerle et al. 1984).
Im Rahmen dieses Beitrags können die mathematischen Grundlagen der linearen kategorialen Regression nur sehr verkürzt dargestellt werden. Ziel des Beitrags ist es, dem interessierten Leser zu vermitteln, wie er selbst Responsefunktionen entwickeln kann und diese innerhalb der Prozedur CATMOD umsetzen kann. Die Umsetzung von Responsefunktionen wird anhand konkreter statistischer Maßzahlen (odds ratio und Kappa) erläutert.

2. Die Prozedur CATMOD

Zunächst folgt ein kurzer Überblick über die Syntax der CATMOD-Prozedur (siehe Abb. 1) und die dort implementierten Standardresponsefunktionen bzw. die möglichen Transforma-

[1] überarbeitete Fassung des Vortrags.

[2] Eine übersichtliche Darstellung zu den Analysemöglichenkeiten kategorialer Daten befindet sich in Andreß/Hagenaars/Kühnel (1997: 20).

tionen (siehe Tabelle 1). Bei den in Tabelle 1 wiedergegebenen Standardresponsefunktionen handelt es sich z.B. im Fall:

• von **logits** um verallgemeinerte Logits der Randwahrscheinlichkeiten,
• von **joints (=marginals)** um Randwahrscheinlichkeiten,
• von **means** um Mittelwerte der abhängigen Variablen (numerisch).

Welches statistische Verfahren in CATMOD zur Analyse der Daten herangezogen werden soll, wird im wesentlichen durch die Bezugnahme auf diese verschiedenen Standardresponsefunktionen, das Model-Statement bzw. ergänzende Statements (z.B. loglin) gesteuert. Das verwendete Schätzverfahren (weighted least square bzw. maximum-likelihood-Schätzer) hängt von der gewählten Responsefunktion ab (siehe: Stokes et al. 1995: 8). Eine gute Übersicht hierzu ist dem SAS/STAT- User's Guide Volume 1 zu entnehmen (siehe auch: SAS Institute 1988; SAS Institute 1990; Nagl 1992; Falk et al. 1995; Stokes et al. 1995; Graf/Ortseifen 1995). Der Beitrag konzentriert sich, wie auch in Abbildung 1 deutlich wird, auf die Möglichkeit der Umsetzung "eigener" Responsefunktionen innerhalb des Responsestatements.

```
PROC CATMOD DATA= SAS-data-set
ORDER= DATA;
       DIRECT variable-list;
       MODEL response_effect= design_effects / options; /*required*/;
       CONTRAST 'label' row_description, row_description,...;
       BY variable-list;
       FACTORS factor_description,... / options;
       LOGLIN effects / option;
       POPULATION variable-list;
       REPEATED factor_description,... / options;
       RESPONSE function / options;
```

Thema: Umsetzung eigener Responsefunktionen
am Beispiel: Odds Ratio und Kappa

```
       RESTRICT parameter=value<...parameter=value>;
       WEIGHT variable;
```

Abbildung 1: Prozedur CATMOD
(Quelle: PROC CATMOD, SAS 6.12 Hilfe)

Die in Tabelle 1 angegeben Transformationsmöglichkeiten sind zur Umsetzung "eigener" Responsefunktionen von großer Wichtigkeit (siehe Punkt 6: Umsetzung der Maßzahlen in Matrizen).

Tabelle 1: Standardresponsefunktionen und mögliche Transformationen

Standardresponsefunktionen:		mögliche Transformationen:
• alogit/s • clogit/s • logit	• joint • marginal/s • mean/s	• Addieren einer Konstante • Linear-Kombinationen • Logarithmieren • Exponieren und die Kombination dieser vier Transformationen

Quelle: PROC CATMOD (SAS 6.12 Hilfe)

Zum besseren Verständnis des Gesamtzusammenhangs werden im nächsten Abschnitt die mathematischen Grundlagen der kategorialen Regression wiedergegeben.

3. Kurzdarstellung der mathematischen Grundlagen der kategorialen Regression (GSK-Ansatz)

Der Ansatz der kategorialen Regression ist zur Analyse von tabellierten Daten (multivariaten Kreuztabellen) geeignet. Innerhalb des Ansatzes der kategorialen Regression gibt es keine Beschränkung bezüglich der analysierbaren Subpopulationen und der Anzahl der Kategorien der Responsevariablen. Dabei ist allerdings Bedingung, daß die Stichprobe groß genug ist, um eine ausreichende Zellbesetzung zu liefern (vgl. Andreß et al. 1997: 58). Die kategoriale Regression[3] kann folgendermaßen beschrieben werden:

Gegeben seien für **s** Subpopulationen (unabhängige Variablen) die Wahrscheinlichkeiten (π_{sr}) für **r** Responsekategorien (abhängige Variablen):

	1	**2**	**...**	**r**
1	π_{11}	π_{12}	...	π_{1r}
2	π_{21}	π_{22}	...	π_{2r}
...	...	π_{ij}	...	...
s	π_{s1}	π_{s2}	...	π_{sr}

$$= \begin{pmatrix} \pi_1' \\ \pi_2' \\ \pi_{...}' \\ \pi_s' \end{pmatrix}$$

wobei für alle i gilt:

$$\sum_{j=1}^{r} \pi_{ij} = 1,$$

(d.h. die Anteilswerte einer Subpopulation addieren sich zu 1).

Dies führt dazu, daß sich der letzte (r-te) Anteilswert immer aus den anderen ergibt. Rein datentechnisch gesehen bestehen die Ausgangsdaten der kategorialen Regression also immer aus einer r x s -Tabelle. Die entsprechenden Häufigkeiten n_{ij} sind für jede Subpopulation im Fall dichotomer Variablen binomial und im Fall mehrerer Ausprägungen multinomial verteilt. Die Subpopulationen müssen voneinander unabhängig sein. Sind die Responsevariablen

[3] Die mathematischen Grundlagen orientieren sich im wesentlichen an einem unveröffentlichten Skript von Dr. Willi Nagl (Uni-Konstanz), welches er mir dankenswerterweise zur Verfügung stellte. Zudem orientiert sich die Darstellungsweise in etwa an den Ausführung von dem Referenzguide (SAS Institute 1990). Herleitungen und genauere Spezifikationen einzelner mathematischer Grundbedingungen sind der Literatur zu entnehmen (vor allem: Hamerle et al. 1984: 211-233).

voneinander abhängig, muß ein kombinierter Responsefaktor gebildet werden (vgl. Andreß et al. 1997: 59). Die Werte werden in der sogenannten Designmatrix gespeichert.
Die beobachteten Anteilswerte (p_{ij}) für die Responsekategorie j sind wie folgt definiert:

$p_{ij} = \frac{n_{ij}}{n_i}$, wobei n_i der jeweilige Umfang der Subpopulation ist.

Man betrachtet also die Wahrscheinlichkeit der Responsevariable (abhängige Variable) unter der Bedingung, daß die Untersuchungseinheit aus der Subpopulation (unabhängige Variable) stammt.

Die Anteile $p_i^T = \left(\frac{n_{ij}}{n_i}, \ldots \frac{n_{ir}}{n_i} \right)$ für die Populationskategorie[4] i werden in einem (sxr,1)-Vektor für alle Populationen gespeichert: $p^T = (p_1, p_2 \ldots p_s)$.

Ein konsistenter Schätzer für die Varianz-Kovarianz-Matrix der p ist:

$$V(p) = \begin{bmatrix} V_{p1} & 0 & 0 & 0 \\ 0 & V_{p2} & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & V_{ps} \end{bmatrix} = DIAG(V_{p1}, V_{p2}, \cdots V_{ps}) \text{ mit}$$

$$V(p_i) = \frac{\left(DIAG(p_i) - p_i p_i^T\right)}{n_i}.$$

Seien nun m Funktionen (Responsefunktionen) von p gegeben:

$$F(p) = \begin{bmatrix} F_1(p) \\ F_2(p) \\ \vdots \\ F_m(p) \end{bmatrix}.$$

Ein konsistenter Schätzer für die Varianz-Kovarianzmatrix für F ist dann gegeben durch:
$V(F) = HV(p)H^T$, wobei $H = \frac{\partial F(p)}{\partial p^{T'}}$ (die Ableitung von jedem F an der Stelle p).

Die Funktionen können nun mit Hilfe eines linearen Modells $F(\pi) = X\beta$ untersucht werden, wobei π den Vektor der Populationswahrscheinlichkeiten für alle Populationen beschreibt. Der Vektor F enthält die Responsefunktionen und β (Parameter) beschreibt die Variation über die Responsefunktionen. Im Prinzip können beliebige Funktionen gewählt werden (siehe hierzu Forthofer/Koch 1973; Andreß et al. 1997). Die **Tests für den Fit** (Modellanpassung) werden mit Hilfe der Wald-Statistik durchgeführt. Die zugrundeliegende Verteilung ist annähernd χ^2-verteilt (Andreß et al. 1997; Hamerle et al. 1984). Die Berechnung des Modelltests erfolgt nach:

$$Q = Q(X, F) = (RF)^T \left(RV_F R^T\right)^{-1} RF$$

[4] Für die Populationskategorie i würde gelten: $p_i^T = (p_{i1}, p_{i2} \ldots p_{ir})$;(r,1)-Vektor.

bei einem (u,t-Vektor) X mit u-t Freiheitsgraden. Das Modell ("Goodness-of-Fit"-Test) wird mit Hilfe der Nullhypothese: "Variation der Residuen = 0" getestet. Die Anzahl der Freiheitsgrade entspricht der Anzahl der Responsefunktionen minus der Zahl der Parameter der unabhängigen Variablen. Bei signifikanter Residualkategorie paßt das Modell nicht, bei nicht signifikanter Residualkategorie paßt das getestete Modell und die Effektparameter können interpretiert werden (SAS Institute 1988: 182).

Die **Hypothesentests** beruhen ebenfalls auf der Wald-Statistik und werden folgendermaßen berechnet:

$$Q_C = (Cb)^T \left[C\left(X^T V_F^{-1} X\right)^{-1} C^T \right]^{-1} Cb$$

mit c Freiheitsgraden, wobei die Nullhypothese folgendermaßen lautet:

$$H_0 : C_{(c,t)} \beta = 0 .$$

Der Effekttest bzw. die statistischen Tests zur Variation der unabhängigen Variablen in der gewählten Responsefunktion wird durch die Nullhypothese: "Parameter = 0" getestet. Ist ein Effekt signifikant, so stellt er eine Quelle der Variation auf die abhängigen Variablen dar. Ein nicht signifikanter Effekt kann aus dem Modell entfernt werden, er wird dann dem Residualraum zugerechnet (SAS Institute 1988: 181).

Die **Prädiktionswerte** bei der Parameterschätzer (WLS-Schätzung) sind folgendermaßen definiert:

$$\hat{F} = Xb = X\left(X^T V_F^{-1} X\right)^{-1} X^T V_F^{-1} F$$

mit der Varianz-Kovarianzmatrix:

$$V_{\hat{F}} = X\left(X^T V_F^{-1} X\right)^{-1} X^T .$$

Der Vektor b für die Parameter ist dann:

$$b = \left(X^T V_F^{-1} X\right)^{-1} X^T V_F^{-1} F$$

(vgl. auch Küchler 1978: 352).

3.1 WLS-Schätzverfahren

Der GSK-Ansatz verwendet gewichtete Kleinst-Quadrate-Schätzungen (weighted least squares). Die WLS-Methode minimiert die gewichtete Residuenquadratsumme, wodurch die Voraussetzung der Streuungsgleichheit sichergestellt ist (vgl. Küchler 1978: 352; SAS Institute 1990: 413). Um robuste WLS-Schätzer zu erhalten, werden an den Stichprobenumfang gewisse Mindestanforderungen gestellt[5]. Den log-linearen Modellen liegt ein Maximum Likelihood-Schätzverfahren zugrunde. Beide Schätzverfahren liefern bei großen Stichproben

[5] Es wird in der Literatur von etwa 20-30 Fällen pro Subpopulation gesprochen (vgl. Küchler 1979: 251). Laut Forthofer darf diese Forderung bei einem Viertel der Subpopulation verletzt sein, wenn keine weniger als 10 Fälle hat (vgl. Andreß et al. 1997: 58; Forthofer/Lehnen 1981).

vergleichbare Ergebnisse[6] (vgl. Andreß et al. 1997: 21 u. 39; Langeheine 1986: 161). Der kategorialen Regression kann ein additives wie auch ein multipikatives Regressionsmodell zugrunde liegen. Im Fall eines additiven Regressionsmodells kommt ein wesentlicher Nachteil des WLS-Schätzverfahrens zum Tragen: Es dürfen sich in der zu analysierenden Tabelle keine Nullzellen befinden. Des weiteren sind unzulässige Prognosen, d.h. Wahrscheinlichkeiten unter 0 bzw. über 1 nicht ausgeschlossen. Zusätzlich kommen bei sehr schief verteilten Zielvariablen (abhängige Variablen) die Unterschiede in den Subgruppen kaum zur Geltung (vgl. Andreß et al. 1997: 21).

3.2 Restriktionen

Der zur kategorialen Regression benötigte Stichprobenumfang hängt von der Anzahl der kategorialen Variablen ab, die man in einem Modell untersuchen will. Je mehr unabhängige Variablen, desto mehr Dimensionen hat die der Analyse zugrundeliegende Tabelle und desto höher ist die Wahrscheinlichkeit, daß manche Zellen der Tabelle nicht besetzt sind. Dies zwingt dazu, die Menge der potentiell in Frage kommenden unabhängigen Variablen auf wenige aussagekräftige zu beschränken (vgl. Andreß et al. 1997: 132). Stichproben-Nullen[7] bilden aufgrund der WLS-Schätzung[8] ein großes Problem bei der kategorialen Regression. Nullzellen können unter Umständen durch das Zusammenfassen von Kategorien vermieden werden, jedoch gehen dadurch möglicherweise wertvolle Information verloren. Wenn trotz sorgfältigem Datenmanagement Stichproben-Nullen auftreten, wird in der Literatur empfohlen, die Null durch eine kleine Zahl zwischen 0 und 1 zu ersetzen oder wie Grizzle (1969) vorschlagen hat, das Ergebnis der Division "1/ Anzahl der Responsekategorien" zu nehmen (vgl. Andreß et al. 1997: 128). Innerhalb des Statistik-Analyse-Systems SAS wird beim Vorhandensein nicht besetzter Zellen zu *allen* vorkommenden Zellen der zugrundeliegenden Tabelle der Wert 0,5 addiert[9]. Allen Korrekturen gemeinsam ist die Auswirkung auf die Schätzergebnisse. Der Einfluß des verwendeten Korrekturverfahrens kann durch das Einsetzen verschiedener Korrekturwerte abgeschätzt bzw. erkannt werden. Durch die gewählten Responsefunktionen (odds ratio und Kappa) machen sich Nullzellen in ähnlicher Weise auf seiten der abhängigen Variablen bemerkbar.

4. Allgemeines zur Designmatrix und Responsefunktion

4.1 Designmatrix:

Zum besseren Verständnis der matrizenorientierten Schreibweise wird anhand eines einfachen Beispiels (dichotome Zielvariable[10]) das Zustandekommen einer Designmatrix demonstriert.
Der klassische Ansatz der univariaten linearen Regression lautet:

$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \varepsilon_i$, i=1,...I.

y_i sind dabei die Beobachtungen der abhängigen Variablen y und der Vektor $x_i^T = (1, x_{i1}, \ldots, x_{ip})$ enthält die Werte für die unabhängigen Variablen.

---

[6] Dies gilt, wenn die Anteile bei der linearen kategorialen Regression ungefähr zwischen 0,20 und 0,80 liegen.
[7] Eine zulässige Merkmalskombination ist in der Stichprobe nicht erfaßt.
[8] Jede Subpopulation wird mit der inversen Varianz gewichtet, bei polytomen abhängigen Variablen werden zusätzlich auch die Kovarianzen berücksichtigt (vgl. Andreß et al. 1997: 127ff).
[9] Dabei handelt es sich um eine Standardfunktion, die individuell angepaßt werden kann.
[10] Im Fall einer mehrkategorialen Zielvariablen muß ein allgemeinerer Ansatz gewählt werden (siehe: Hamerle et al. 1984: 221-233).

$\beta^T = (\beta_0, \ldots \beta_p)$ ist der unbekannte Parametervektor (Regressionskoeffizienten) und ε_i die nicht beobachtbare Fehlervariable (vgl. Hamerle et al. 1984: 213). Die Variablen bei der kategorialen Regression sind nicht metrisch, d.h. sie müssen entsprechend kodiert werden[11]. Am Beispiel einer Effekt-Kodierung (1, falls Kategorie i der Variablen A vorliegt, -1, falls Kategorie I der Variablen A vorliegt, sonst 0; bei i=1,...,I-1) und einer Dummy-Kodierung (1, falls Kategorie i der Variablen A vorliegt, sonst 0; bei i=1,...,I-1) wird dies aufgezeigt. Bezeichnet man alle Merkmalskombinationen (Datenvektoren) mit $x_1, \ldots, x_I$, läßt sich $x_1, \ldots, x_I$ zur Designmatrix

$X = \begin{pmatrix} x_1^T \\ \vdots \\ x_I^T \end{pmatrix}$ zusammenfassen. Die x_i erhalten dabei jeweils als erste Komponente eine 1 für den Koeffizienten β_0. Das heißt, die Regressionsformel läßt sich unter Weglassung der Störgrößen wie folgt ausdrücken:

$\pi = X\beta$.

Die Vektoren $x_1, \ldots, x_I$ sind durch die Komplexität des zugrundeliegenden Modells festgelegt. Die Anzahl der Zeilen der Designmatrix ergibt sich aus der Anzahl der Kombinationen der Kategorien der unabhängigen Variablen. Für jede beteiligte unabhängige Variable gibt es i-1 Koeffizienten (hypothetisches Beispiel für die unabhängigen Variablen: Alter (B) mit 2 Kategorien; Konfession (C) mit 3 Kategorien) I=2*3=6 Zeilen der Designmatrix und z.B. für die Variable (C) 3-1= 2 Koeffizienten). Die Anzahl der Responsefunktionen (abhängige Variable) richtet sich wegen der linearen Abhängigkeit der jeweils letzten Kategorie nach der Kategorienzahl der Responsevariable minus 1 (i-1). Im Fall einer dichotomen Responsevariable ergibt sich genau eine Responsefunktion. In einem (saturierten) Modell ergeben sich auf das Beispiel oben bezogen folgende Designmatrix:

Designmatrix	Effekt-Kodierung	Designmatrix	Dummy-Kodierung
für die unabhängigen Variablen B und C (siehe oben)	B: 1, -1 C: 1, 0 0, 1 -1, -1	für die unabhängigen Variablen B und C (siehe oben)	B: 1, 0 C: 1, 0 0, 1 0, 0
$X = \begin{bmatrix} 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & 0 & -1 & 0 \\ 1 & -1 & 0 & 1 & 0 & -1 \\ 1 & -1 & -1 & -1 & 1 & 1 \end{bmatrix}$	β_0 β_{B1} β_{C1} β_{C2} β_{B1*C1} β_{B1*C2}	$X = \begin{bmatrix} 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$	β_0 β_{B1} β_{C1} β_{C2} β_{B1*C1} β_{B1*C2}

4.2 Responsefunktion

Die Responsefunktion läßt sich nahezu beliebig modellieren[12], es kann z.B. entweder auf die gemeinsamen oder die zeilenbedingten Anteile einer vorliegenden Kreuztabelle Bezug genommen werden. Des weiteren können Logits die Grundlage der Responsefunktion sein (logarithmierte Verhältnisse, siehe hierzu (SAS Institute 1990: 435ff; Hamerle et al. 1984: 219ff)).

[11] Entweder die klassische Dummy-Kodierung oder mit einer Effekt-Kodierung.

[12] Die Funktion muß eindeutig und hinreichend oft differenzierbar sein (vgl. Hamerle et al. 1984: 218).

Die Responsefunktion bedeutet nichts anderes, als daß für jedes π_i aus der Matrix π eine bestimmte Funktion $g(\pi_i)$ zugrunde liegt. Das heißt, es gilt im allgemeinsten Fall:

$h(\bar{\pi}) = Z\beta + \varepsilon$ (vgl. Hamerle et al. 1984: 223).
Zur Umsetzung der odds ratio und Kappa als abhängige Variable im GSK-Ansatz muß die Berechnungsformel der jeweiligen Maßzahl als Responsefunktion benutzt werden (zur Beschreibung der beiden Maßzahlen siehe weiter unten).

5. Inhaltliche und methodische Vorbemerkungen

An einem konkreten Beispiel aus der Familiensoziologie (Partnerwahl: Bildungshomogamie) wird ein möglicher inhaltlicher Bezug der Modellierung der odds ratio und Kappa als abhängige Variable klar. Bei der Analyse von Partnerwahl wurde festgestellt, daß Ehepartner überzufällig oft die gleiche bzw. ähnliche Bildung besitzen. Dieses Phänomen wird als Bildungshomogamie bezeichnet. Will man herausfinden, ob z.B. die soziale Herkunft der Partner, das Alter der Partner und andere soziale Merkmale die Bildungshomogamie beeinflussen, stößt man auf das Problem der Randverteilungsabhängigkeit der abhängigen Variablen "Bildung". Das Problem verschärft sich bei einer ländervergleichenden Perspektive zusätzlich.
Diese Randverteilungsabhängigkeit kann man durch den Bezug auf randverteilungsunabhängige Maßzahlen odds ratio und Kappa (zu den Maßzahlen siehe auch: Hildebrand et al. 1977; Liebetrau 1983; Reynolds 1977; Rudas 1998) und durch die Aufnahme dieser Maßzahlen als abhängige Variable in einem Regressionsmodel adäquat berücksichtigen. Durch die Interpretation der odds ratio als bildungsbezogene Heiratsbarrieren und Kappa als Maß für Übereinstimmung (also Homogamie) wird zudem die inhaltliche Dimension der benutzten Maßzahlen deutlich (siehe hierzu auch Punkt 5.1 und Punkt 5.2).

5.1 Odds ratio

Odds ratios werden in der Soziologie häufig zur Analyse von Mobilitätstabellen verwendet (z.B. Alba 1987; Cobalti 1988; Cobalti 1989; Goodman 1969; Goodman 1970; Kaufman/Schervish 1987) und haben sich auch bei der Analyse von Heiratsmustern etabliert (Klein 1998; Klein 1997). Ausgehend von einer 2x2-Tabelle mit den Feldern a, b, c und d (Diagonalzellen a, d) ist die odds ratio (OR) folgendermaßen definiert:

$$OR = \frac{(a * d)}{(b * c)}.$$

Odds ratios lassen sich auch als Relation Relativer Risiken interpretieren, die im Gegensatz zu dem Relativen Risiko selbst randverteilungsunabhängig sind. Odds ratios können Werte von 0 bis unendlich annehmen. Ob ein Wert kleiner 1 oder größer 1 ist, hängt dabei von der Konzeption der Kreuztabelle und von der Wirkungsrichtung des untersuchten Zusammenhangs ab. Berechnet man die odds ratio aus einer Kreuztabelle, die sich aus den Ausprägungen der dichotomisierten Bildungsausprägungen Verheirateter ergibt, läßt sich diese Maßzahl als bildungsbezogene Heiratsbarriere interpretieren. Wird die odds ratio genau 1, dann besagt der berechnete Wert, daß unter Kontrolle der Randverteilungen keine bildungsbezogene Heiratsbarrieren zwischen den Geschlechtern bestehen (vgl. Klein/Rüffer 1999).

5.2 Kappa

Kappa ist ein von Cohen entwickeltes Maß, das ursprünglich als Reliabilitätsmaß konzipiert wurde, um speziell die auf Urteilsübereinstimmung zielende Kontingenz zu messen, die sogenannte Inter-Rater-Reliabilität. Der Koeffizient ist auf Kategorialdaten andwendbar und resultiert aus der Differenz der tatsächlichen beobachteten Besetzungszahlen der Diagonalfelder einer Kreuztabelle und der aus den Randverteilungen (bei statistischer Unabhängigkeit) erwartbaren Besetzungszahlen und ist folgendermaßen definiert:

$$Kappa = \frac{\theta_1 - \theta_2}{1 - \theta_2}$$

Der Koeffizient liegt zwischen -1 und +1. Positive Werte zeigen an, daß überzufällige Homogamie vorliegt. Negative Werte indizieren hingegen, daß soziale Mechanismen auf Heterogamie angelegt sind. Den Wert Null erreicht Kappa bei vollständiger Übereinstimmung der zufällig erwarteten mit den tatsächlichen Anteilen in den Diagonalfeldern. Werte nahe Null besagen somit, daß kaum soziale Steuerungsmechanismen wirksam sind (vgl.Klein 1999).

6. Umsetzung der statistischen Maßzahlen in Matrizenschreibweise

Zur Umsetzung von Formel in Matrizenschreibweise lassen sich folgende "Tips" formulieren:

1. Tabellenstruktur anhand des in CATMOD ausgegebenen Response-Profils überprüfen (wegen der Anteilsdefinition).
2. Formel als Resultat der Verrechnung dieser Anteilswerte reformulieren.
3. Benötigte Matrizen konzipieren.
4. Matrizen richtungsverkehrt in SAS eingeben.
5. Überprüfung der Ergebnisse der Berechnungen (Responsefunktion).

6.1 Umsetzung der odds ratio

Nun zur Umsetzung der odds ratio als abhängige Variable in einem linearen kategorialen Regressionsmodell. Im folgenden wird der Weg von der Kreuztabelle zur Responsefunktion am Beispiel der odds ratio aufgezeigt.

Ausgehend von einer Kreuztabelle mit 2x2 Feldern:

Variable 1	Variable 2		
	Ausprägung I	Ausprägung II	Summe
Ausprägung I	a	b	a+b
Ausprägung II	c	d	c+d
Summe	a+c	b+d	a+b+c+d

ist die odds ratio (OR) folgendermaßen definiert:

$$OR = \frac{(a * d)}{(b * c)}.$$

Die Formel muß mit Hilfe von linearen Operatoren (log- und exp- Transformationen) formuliert werden. Hierzu werden folgende (Hilfs-)Transformationen zur Berechnung innerhalb der Matrix benötigt:

zur Durchführung einer **Multiplikation:** $\ln(x)+\ln(y)=\ln(x*y)$
zur Durchführung einer **Division:** $\ln(x)-\ln(y)=\ln(x/y)$

Die 2x2-Kreuztabelle entsteht (innerhalb CATMOD) durch folgende Spezifikation im Model-Statement:

```
proc catmod ...;
...
model abhängige Variable 1 x abhängige Variable 2 = unabhängige Variable(n);
...
```

(Im Fall der Bildungsvariablen als abhängige Variablen also "Bildung des Mannes" mal "Bildung der Frau").

Aufgrund der 2x2 Ausprägungen der abhängigen Variablen ergeben sich im GSK-Ansatz 4 Responsekategorien mit den Anteilen: p‘ =(a,b,c,d) siehe Kreuztabelle.
Benötigt werden "a∗d" und "b∗c" als Komponenten der odds ratio-Formel, sowie später die Division der beiden Produkte.

1. Schritt

Man braucht eine Matrize, in der die Anteile a, b, c, d “eingelesen” werden:

	a	b	c	d
1.	1	0	0	0
2.	0	1	0	0
3.	0	0	1	0
4.	0	0	0	1

sozusagen: 1*a+0*b+0*c+0*d
entsprechend: 0*a+1*b+0*c+0*d
usw.

Diese Matrix enspricht dem Inhalt der Zellen einer 2x2-Tabelle. Da man die Produkte “a∗d” und “b∗c” braucht, muß man zunächst die Matrix logarithmieren (wegen der Umwandlung einer Multiplikation in eine Addition: $\ln(x)+\ln(y)=\ln(x*y)$).

2. Schritt

Man “addiert” dann also:

	ln(a)	ln(b)	ln(c)	ln(d)
1.	1	0	0	1
2.	0	1	1	0

sozusagen: $1*\ln(a)+0*\ln(b)+0*\ln(c)+1*\ln(d)$
entsprechend: $0*\ln(a)+1*\ln(b)+1*\ln(c)+0*\ln(d)$

durch anschließendes Delogarithmieren erhält man die gewünschten Produkte “a∗d” und “b∗c” (Nenner und Zähler der odds ratio!). Dadurch erhält man folgende neue Matrix:

3. Schritt

	(a∗d)	(b∗c)
1.	1	0
2.	0	1

sozusagen: $1*(a*d)+0*(b*c)$
entsprechend: $0*(a*d)+1*(b*c)$

Nun muß nochmals logarithmiert werden, um sich dieses mal gemäß der Transformation $\ln(x)-\ln(y)=\ln(x/y)$ zunutze machen zu können:

4. Schritt

d.h. konkret:

	ln(a∗d)	ln(b∗c)		
1.	1	-1	sozusagen:	1∗ln(a∗d)-1∗ln(b∗c)

durch die Subtraktion beider logarithmierten Bestandteile der Matrize wird durch erneutes Delogarithmieren die odds ratio berechnet (die Division durchgeführt).

6.1.1 Umsetzung in CATMOD

Die oben entwickelte Matrizenfolge muß nun in umgekehrter Reihenfolge (SAS arbeitet die Matrizen "rückwärts" ab) innerhalb des Response-Statements eingegeben werden. Des weiteren gilt es zu beachten, daß in SAS log=ln entspricht und die Zeilen der Matrizen durch Kommata getrennt werden. Die Übertragung ins SAS-Programm (Abbildung 2) geschieht nun folgendermaßen:

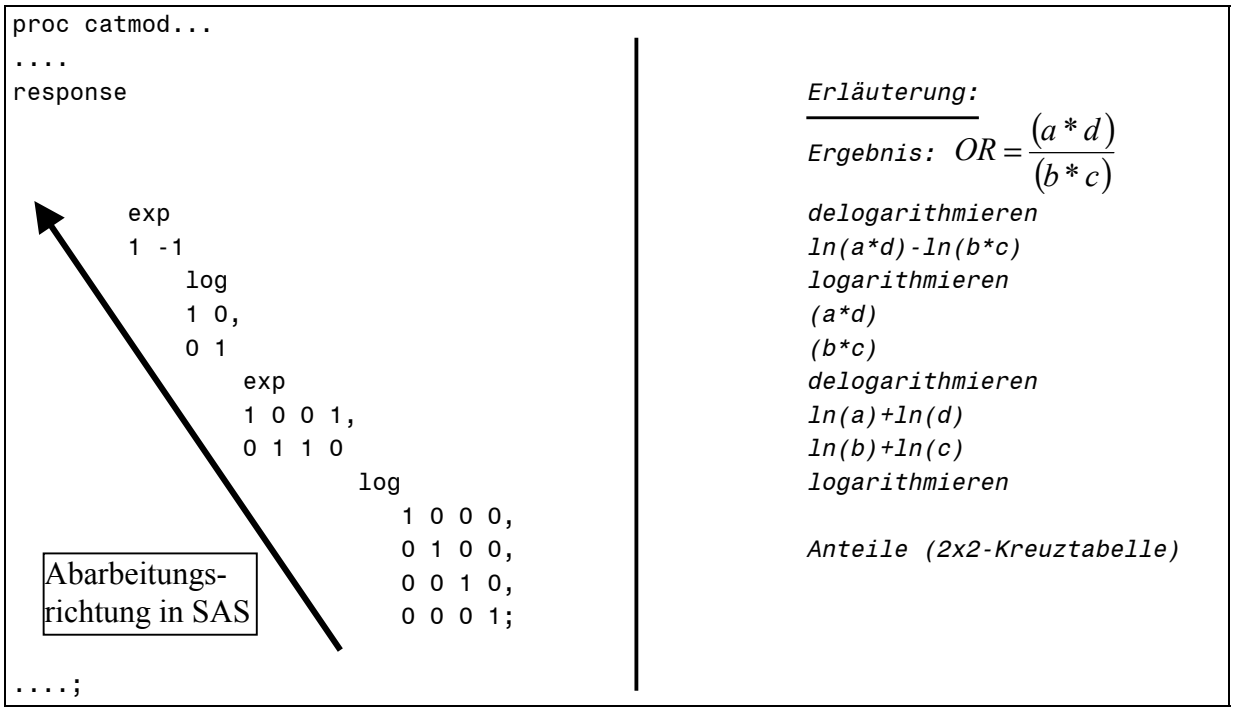

Abbildung 2: Umsetzung der odds ratio als Responsefunktion in SAS CATMOD

6.2 Umsetzung von Kappa

Die "Kappa"-Formel (siehe Punkt 5.2) wird zunächst in einer an die 2x2-Tabelle angepaßte Schreibweise reformuliert (Kappa kann auch für n-dimensionale Tabellen berechnet werden, hierzu muß einfach eine entsprechend größere Matrize formuliert werden).
Zunächst folgen die Einzelbestandteile der "Kappa"-Formel und deren Reformulierung:

$$\theta_1 = \sum_{i=1}^{I} p_{ii} \Rightarrow a + d$$

$$\theta_2 = \sum_{i=1}^{I} p_{i+} p_{+i} \Rightarrow ((a+b)\times(a+c)+(b+d)\times(c+d))$$

$$Kappa = \frac{\theta_1 - \theta_2}{1 - \theta_2}$$ (siehe Liebetrau 1983:S. 32)

Formel: (an 2x2-Kreuztabelle angepaßte Schreibweise)

$$Kappa = \frac{(a+d) - ((a+b)\times(a+c) + (b+d)\times(c+d))}{1 - ((a+b)\times(a+c) + (b+d)\times(c+d))}.$$

Gegeben sind wieder die 4 Responsekategorien mit den Anteilen der oben genannten 2x2-Kreuztabelle: p‘=(a,b,c,d). Als Komponenten der Kappa-Formel werden benötigt: "(a+d)" und "((a+b)∗(a+c)+(b+d)∗(c+d))" sowie "(a+b+c+d)=1".

1. Schritt

	a	b	c	d		
1.	1	1	0	0	a+b	1. Zeilenrandanteil
2.	0	0	1	1	c+d	2. Zeilenrandanteil
3.	1	0	1	0	a+c	1. Spaltenrandanteil
4.	0	1	0	1	b+d	2. Spaltenrandanteil
5.	1	0	0	1	a+d	Summe der Diagonale (Übereinstimmung)
6.	1	1	1	1	a+b+c+d	Gesamtsumme entspricht 1

Da man die Produkte “(a+b)∗(a+c)” und “(b+d)∗(c+d)” erhalten will, muß wieder logarithmiert und entsprechend addiert werden (wegen ln(x)+ln(y)=ln(x∗y)).

2. Schritt

	ln(a+b)	ln(c+d)	ln(a+c)	ln(b+d)	ln(a+d)	ln(a+b+c+d)	
1.	1	0	1	0	0	0	ln(a+b)+ln(a+c)
2.	0	1	0	1	0	0	ln(c+d)+ln(b+d)
3.	0	0	0	0	1	0	ln(a+d)
4.	0	0	0	0	0	1	ln(a+b+c+d)

Danach wird delogarithmiert, und man hat die gewünschten Produkte, die anderen Komponenten ((a+d) und (a+b+c+d)) bleiben dadurch unberührt. Nach dem Delogarithmieren hat man dann also:
"(a+b)∗(a+c)" und "(c+d)∗(b+d)" sowie "(a+d)" und "(a+b+c+d)" zur Verfügung.

3. Schritt

	(a+b)∗(a+c)	(c+d)∗(b+d)	(a+d)	(a+b+c+d)	
1.	-1	-1	1	0	(a+d)-((a+b)∗(a+c)+(c+d)∗(b+d)) (Zähler der Kappa-Formel)
2.	-1	-1	0	1	(a+b+c+d)-((a+b)∗(a+c)+(c+d)∗(b+d)) (Nenner der Kappa-Formel)

Danach wieder logarithmieren, um die Division durchführen zu können.

4. Schritt

	ln (Zähler Kappa-Formel)	ln(Nenner Kappa-Formel)	
1.	1	-1	ln(Zähler K...)-ln(Nenner K...)

Durch abschließendes Delogarithmieren wird die Division erreicht (wegen ln(x)-ln(y)=ln(x/y).

6.2.1 Umsetzung in CATMOD

Die Vorgehensweise bei der Umsetzung in SAS (Abbildung 3) entspricht der bei dem vorangegangenen Beispiel (odds ratio).

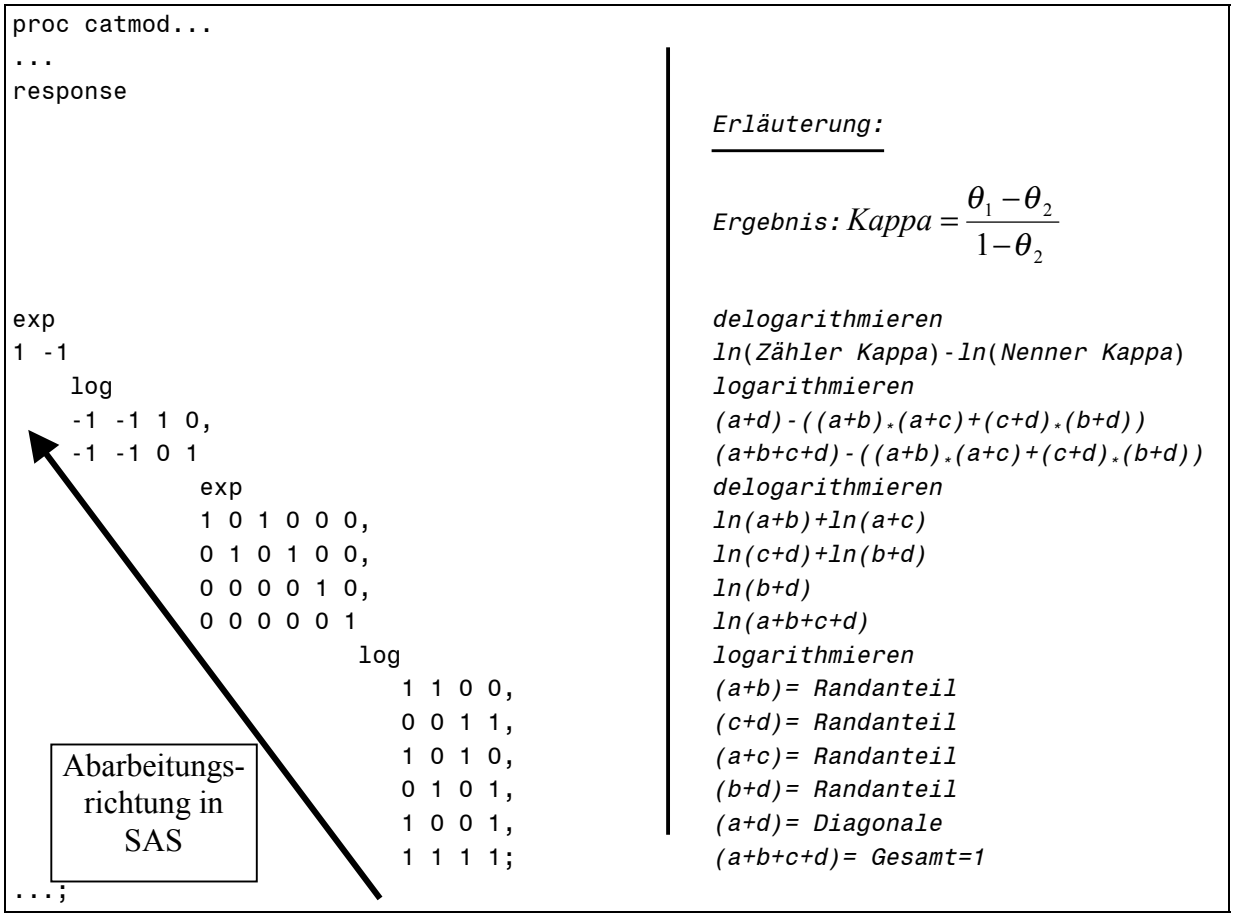

Abbildung 3: Umsetzung von Kappa als Responsefunktion in SAS CATMOD

7. Ausblick

Natürlich lassen sich auch andere Formeln mit Hilfe dieser Vorgehensweise umsetzen (z.B. G-K-Tau, Lambda usw.). Die Wahl der Responsefunktion hängt dabei im wesentlichen von dem "gewünschten" inhaltlichen Bezug und von den statistisch bzw. mathematischen Voraussetzungen ab. Eine nützliche Anwendungsmöglichkeit der vorgestellten Berechnungsweise ist die relativ einfache Berechnung der Subgruppen-odds ratios bzw. -Kappas im Fall der Verwendung von saturierten Modellen.

Die Interpretation der Koeffizienten unterscheidet sich je nach Kodierung der unabhängigen Variabeln (Dummy- bzw. Effektkodierung). Im Fall von dummy-kodierten unabhängigen Variablen ist der Parameter direkt als “Erhöhung” bzw. “Senkung” der entsprechenden odds ratio bzw. Kappa bezüglich der jeweils gewählten Referenzgruppe interpretierbar. Im Fall von effekt-kodierten unabhängigen Variablen sind die Parameter als Abweichung von einem "mittleren" odds ratio bzw. Kappa interpretierbar. Welche Kodierung gewählt wird, hängt dabei wiederum von der "gewünschten" inhaltlichen Bedeutung ab. Abschließend sei noch einmal erwähnt, daß eine sinnvolle Anwendung der oben beschriebenen Vorgehensweise eine ausreichend große Stichprobe voraussetzt.

Literatur

Alba, R.D., 1987: Interpreting the Parameters of Log-Linear Models. Sociological Methods & Research, 16: 45-77.

Andreß, H.-J./Hagenaars, J.A./Kühnel, S., 1997: Analyse von Tabellen und kategorialen Daten. Log-lineare Modelle, latente Klassenanalyse, logistische Regression und GSK-Ansatz. Berlin: Springer.

Cobalti, A., 1988: Alternative Conceptual Frameworks for the Analysis of Mobility Tables and the Log-Linear Models. Quality & Quantity, 22: 31-47.

Cobalti, A., 1989: A Relative Mobility Table: A Modest Proposal. Quality and Quantity, 23: 205-220.

Cohen, J., 1960: A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement, 20: 37-46.

Fahrmeir, L./Hamlere, A./Tutz, G., 1996: Multivariate statistische Verfahren. Berlin: Walter de Gruyter.

Falk, M./Becker, R./Marohn, F., 1995: Angewandte Statistik mit SAS. Eine Einführung. Berlin/Heidelberg: Springer.

Forthofer, R.N./Koch, G.G., 1973: An Analysis for Compounded Functions of Categorical Data. Biometrics: 143-157.

Forthofer, R.N./Lehnen, R.G., 1981: Public Program Analysis: A New Categorical Data Approach. Belmont: Wadsworth.

Goodman, L.A., 1969: On the Measurement of Social Mobility: An Index of Status Persistance. American Sociological Review, 34: 831-850.

Goodman, L.A., 1970: How to Ransack Social Mobility Tables and Other Kinds of Cross-Classification Tables. American Journal of Sociology, 75: 1-40.

Graf, A./Ortseifen, C., 1995: Statistische und grafische Analyse mit SAS: Spektrum Akademischer Verlag.

Grizzle, J.E./Starmer, C.F./Koch, G.G., 1969: Analysis of Categorical Data by Linear Models. Biometrics, 25: 489-504.

Hamerle, A./Kemény, P./Tutz, G., 1984: Kategoriale Regression. S. 210-256. In: L. Fahrmeir/A. Hamerle (Hg.), Multivariate statistische Verfahren. Berlin: Walter de Gruyter.

Hildebrand, D.K./Laing, J.D./Rosenthal, H., 1977: Analyisis of Ordinal Data. Beverly Hills: Sage.

Kaufman, R.L./Schervish, P.G., 1987: Variations on a Theme: More Uses of Odds Ratios to Interpret Log-Linear Parameters. Sociological Methods and Research, 16: 218-255.

Klein, T., 1997: Intergenerationale und intragenerationale Heiratsmobilität von Frauen. S. 41-64. In: R. Becker (Hg.), Generationen und sozialer Wandel. Generationsdynamik, Generationenbeziehungen und Differenzierung von Generationen. Opladen: Leske + Budrich.

Klein, T., 1998: Entwicklung und Determinanten der bildungsbezogenen Partnerwahl. Zeitschrift für Bevölkerungswissenschaft, 23: 123-149.

Klein, T., 1999: Partnerwahl in Ehen und Nichtehelichen Lebensgemeinschaften. In: T. Klein/W. Lauterbach (Hg.), Nichteheliche Lebensgemeinschaften aus soziologischer Perspektive. Opladen: Leske + Budrich.

Klein, T./Rüffer, W., 1999: Bildungshomogamie im internationalen Vergleich: Empirische Untersuchungen für die USA, Österreich, Ungarn und Deutschland. Zeitschrift für Familienforschung, zur Publikation angenommen.

Küchler, M., 1978: Alternativen in der Kreuztabellenanalyse - Ein Vergleich zwischen Goodmans "General Model" (ECTA) und dem Verfahren gewichteter Regression nach Grizzle et al. (NOMMET II). Zeitschrift für Soziologie, 7: 347-365.

Küchler, M., 1979: Multivariate Analyseverfahren. Stuttgart: Teubner.

Langeheine, R., 1986: Log-lineare Modelle. S. 122-195. In: J. v. Koolwijk/M. Wieken-Mayser (Hg.), Techniken der empirischen Sozialforschung. Kausalanalyse. München: Oldenburg.

Liebetrau, A.M., 1983: Measures of Association. Beverly Hills: Sage.

Nagl, W., 1992: Statistische Datenanalyse mit SAS. Frankfurt: Campus.

Reynolds, H.T., 1977: The Analysis of Cross-Classifications. New York: The Free Press, A Division of Macmillan Publishing Co., Inc.

Rudas, T., 1998: Odds Ratios In The Analysis Of Contingency Tables. Iowa City: Sage Publications, Inc.

SAS Institute, 1988: Categorical Data Analysis, Course Notes. Cary: SAS Institute Inc.

SAS Institute, 1990: SAS/STAT User's Guide Version 6, Fourth Edition Volume 1. Cary: SAS Institute Inc.

Stokes, M.E./Davis, C.S./Koch, G.G., 1995: Categorical Data Analysis Using the SAS System. Cary, NC., USA: SAS Campus Drive.