Zeitschrift: Schweizerische Zeitschrift für Soziologie = Revue suisse de sociologie = Swiss journal of sociology

Herausgeber: Schweizerische Gesellschaft für Soziologie

Band: 5 (1979)

Heft: 3

Artikel: Multidimensionale Analyse von kategorialen Daten : log-lineare Modelle : eine Einführung und ein Beispiel aus der Epidemiologie des Drogenkonsums

Autor: Binder, Johann

DOI: https://doi.org/10.5169/seals-814088

Schweiz.Z.Soziol./Rev.suisse sociol., 5 (1979) 365-380

MULTIDIMENSIONALE ANALYSE VON KATEGORIALEN DATEN : LOG-LINEARE MODELLE

Eine Einführung und ein Beispiel aus der Epidemiologie des Drogenkonsums

Johann Binder
Psychiatrische Universitätsklinik Zürich

ZUSAMMENFASSUNG

Das von Leo Goodman entwickelte Verfahren der Analyse log-linearer Modelle erweist sich als geeignet, bei der Analyse mehrdimensionaler Kreuztabellen Korrelationen und Interaktionen höherer Ordnung zu identifizieren und ihre relative Stärke abzuschätzen. In vorliegendem Aufstatz wird versucht, das Verfahren in möglichst einfacher Weise darzustellen. Dabei wird auf strukturelle Ähnlichkeiten zur Mehrweg-Varianzanalyse hingewiesen und das Verfahren wird anhand einer sechsdimensionalen Tabelle aus der Epidemiologie des Drogenkonsums vordemonstriert.

RÉSUMÉ

Pour analyser des données distribuées sur plusieurs dimensions, la méthode dite d'"analyse des modèles log-linéaires" mise au point par Leo Goodman s'est avérée efficiente dans l'identification de corrélations et d'interactions d'un niveau plus élevé, ainsi que dans la détermination du poids fonctionnel relatif de ces dernières. L'article expose cette méthode de manière très simple. Dans ce but ont été mises en évidence les analogies avec les analyses multifactorielles de la variance et la méthode a été appliquée à des données empruntées à l'épidémiologie de la toxicomanie et distribuées sur un tableau à six entrées.

1. PROBLEMSTELLUNG UND ZIELSETZUNG

In den Sozialwissenschaften fallen häufig mehrdimensionale Kreuztabellen von kategorialen oder ordinalen Daten an. Solche Tabellen mit drei und mehr Dimensionen werden sehr rasch unübersichtlich und können ohne komplexe statische Methoden nicht zuverlässig interpretiert werden. Die sequentielle Analyse aller bivariaten Zusammenhänge erweist sich als untaugliches Vorgehen, weil damit Scheinkorrelationen oder Interaktionseffekte zwischen mehreren Variablen nicht zuverlässig aufgedeckt werden können. Bei der Interpretation solcher mehrdimensionaler Tabellen sind vor allem die folgenden Fragen von Bedeutung :

(1) Zwischen welchen Variablen bestehen Zusammenhänge ? oder :

(1 a) Falls eine der Variablen als abhängige Variable betrachtet wird : wie hängt die abhängige Variable von den übrigen, als unabhängig betrachteten Variablen ab ?

(2) Wie stark sind die Zusammenhänge, die zwischen den einzelnen Variablen bestehen ?

In den letzten Jahren sind mehrere solcher multivariater Verfahren für die Analyse von Nominal- und Ordinaldaten entwickelt worden, so etwa gewichtete Regression (Grizzle, Starmer, Koch, 1969), multivariate nominal scale analysis (Andrews, Messenger, 1973), das DIEC-Verfahren von Küchler (1976) sowie die

von Leo Goodman entwickelte Methode der log-linearen Modelle. Mit der zunehmenden Berücksichtigung dieser Verfahren in allgemein verfügbaren Statistikprogrammpaketen wird die Verwendung dieser Verfahren stark erleichtert.

Die zuletzt genannte Methode der log-linearen Modelle erfreut sich seit einigen Jahren zunehmender Beliebtheit, wenn man die Anzahl von Publikationen in vorwiegend amerikanischen Fachzeitschriften betrachtet, in denen sie angewendet worden ist. Hierzulande scheint die Methode jedoch noch wenig beachtet zu werden. Aus diesem Grunde möchte ich mit diesem Aufsatz die Methode der log-linearen Modelle (und den wichtigen Spezialfall des Logit-Modells) in möglichst "untechnischer" Sprache bekanntmachen und einige elementare Grundbegriffe erklären, soweit sie zu einem intuitiven Verständnis der Methode notwendig sind. Nicht beabsichtigt is eine statistisch fundierte Diskussion oder eine kritische Würdigung dieses Verfahrens im Vergleich zu anderen Methoden. Dazu sei auf die reichlich vorhandene Spezialliteratur verwiesen: Bischop et al., 1975; Goodman, 1972, 1973, 1976; Davis, 1974 und als Einführung Fienberg, 1977. Das Verständnis der Methode soll weiter dadurch erleichtert werden, dass an mehreren Stellen auf die den log-linearen Modellen eigene strukturelle Ähnlichkeit zur Mehrweg-Varianzanalyse hingewiesen wird und indem in Abschnitt 5 ein Beispiel für die Anwendung dargestellt wird.

2. GRUNDBEGRIFFE

Die zum Verständnis der Methode notwendigen Grundbegriffe sollen anhand der folgenden dreidimensionalen Kontingenztafel dargestellt werden; es fällt nicht schwer, diese Begriffe auf den vier- oder mehrdimensionalen Fall zu verallgemeinern. Wir betrachten im folgenden drei dichotome Variablen:

A Einkommen der Eltern,
B Geschlecht,
C Drogenerfahrung.

Tabelle 1.

A Einkommen den Eltern	A = 1 (tief)		A = 2 (hoch)	
B Geschlecht	B = 1	B = 2	B = 1	B = 2
C Drogenerfahrung	(Mann)	(Frau)	(Mann)	(Frau)
LC = 1 (ja)	1 036	160	699	248
LC = 2 (nein)	3 659	1 141	1 978	952

Die Zahlen in den einzelnen Zellen sind die beobachteten Häufigkeiten der jeweiligen Konfiguration von Merkmalsausprägungen in den drei Variablen. F_{122} bezeichnet die Anzahl der Beobachtungen mit der Merkmalsausprägung A = 1, B = 2, C = 2, d.h. der Frauen ohne Drogenerfahrung, deren Eltern ein tiefes Einkommen haben. Im Beispiel ist $F_{122} = 1141$.

Unter einer *Randverteilung* (marginal distribution) versteht man diejenige Kreuztabelle, die sich ergibt, wenn man über eine oder mehrere Variablen in der vollständigen mehrdimensionalen Kreuztabelle summiert. Die Randverteilung von B und C im obigen Beispiel erhält man etwa, wenn man die Werte für A = 1 und A = 2 zusammenzählt. Dies ist der Zusammenhang zwischen Geschlecht und Drogenerfahrung, wenn man das Einkommen der Eltern nicht berücksichtigt.

Tabelle 2.

B Geschlecht C Drogenerfahrung	B = 1 (Mann)	B = 2 (Frau)	Total
C = 1 (ja)	1 735	408	2 143
C = 2 (nein)	5 637	2 093	7 730

Die Randverteilung von C erhält man, wenn man zusätzlich über die Werte von B summiert. Dies ist die einfache Häufigkeitsauszählung der Variable C, Drogenerfahrung (s. rechten Tabellenrand).

Odds ratios.

Unter einem odds ratio versteht man das Verhältnis der Häufigkeiten der beiden Kategorien einer dichotomen Variablen (bzw. einer bestimmten Kategorie zu allen übrigen Kategorien bei polytomen Variablen). Das odds ratio für die Variable Drogenkonsum ist 2143/7730 = 0.277. Die Angabe der Verteilung einer dichotomen Variablen als Verhältniszahl der beiden Kategorien zueinander scheint zunächst etwas unüblich; sie kann aber bei Bedarf immer in eine entsprechende relative Häufigkeit übergeführt werden (21,7 % der Stichprobe haben Drogenerfahrung).

Ein bedingtes odds ratio ist ein odds ratio bei einer bestimmten Bedingung, beispielsweise das odds ratio für C unter der Bedingung B = 1 : 1735/5637 = 0.308.

Ein odds ratio zweiter Ordnung ist der Quotient zweier bedingter odds ratios, z.B. das odds ratio von C für Männer (B = 1) dividiert durch das odds ratio von C für Frauen (B = 2): (1735/5637)/(408/2093) = 1.579.

Ein odds ratio zweiter Ordnung lässt sich als Korrelation interpretieren. Wenn das odds ratio zweiter Ordnung = 1 ist, dann besteht keine Korrelation zwischen den beiden involvierten Variablen. Ein odds ratio dritter Ordnung ist der Quotient zweier odds ratios zweiter Ordnung usf.

Modelle

Es lässt sich zeigen, dass die Häufigkeit einer Zellenbesetzung F_{ijk} für A = i, B = j, C = k immer durch einen Produktausdruck der folgenden Art dargestellt werden kann

$$F_{ijk} = \eta \cdot \tau^{A}_{i} \cdot \tau^{B}_{j} \cdot \tau^{C}_{k} \cdot \tau^{AB}_{ij} \cdot \tau^{AC}_{ik} \cdot \tau^{BC}_{jk} \cdot \tau^{ABC}_{ijk}.$$ [1]

[1] Bei den Koeffizienten τ, log τ, λ, bezeichnet das Superskript $^{(A,B,C)}$ die Variable, das Subskript $_{(i,j,k)}$ die Merkmalsausprägung.

Dabei ist η das geometrische Mittel der Zellenhäufigkeiten, τ_i^A das odds ratio von A = i, τ_{ij}^{AB} das odds ratio zweiter Ordnung von A = i, B = j, usw.

Die obige Formel kann auch anders geschrieben werden, wenn man logarithmiert; aus dem Produkt ergibt sich nun eine Summe von Logarithmen:

$$\log F_{ijk} = \log\eta + \log\tau_i^A + \log\tau_j^B + \log\tau_k^C + \log\tau_{ij}^{AB} + \log\tau_{ik}^{AC} + \log\tau_{jk}^{BC} + \log\tau_{ijk}^{ABC}.$$

Verwendet man nun noch Abkürzungen $\theta = \log\eta$, $\lambda = \log\tau$ für die Logarithmenausdrücke, so kann man die Formel für den Logarithmus einer Zellenhäufigkeit als eine gewöhnliche Summe schreiben:

$$\log F_{ijk} = \theta + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{ik}^{AC} + \lambda_{jk}^{BC} + \lambda_{ijk}^{ABC}$$

Ein solcher Ausdruck für den Logarithmus der Zellenhäufigkeit wird als log-lineares Modell für die Zellenhäufigkeit bezeichnet. Im Prinzip besteht für jede einzelne Zelle F_{ijk} ein solches Modell. Da aber $\Sigma\,\lambda_i^A = 0$ für alle i, usw. gilt, vereinfachen sich die verschiedenen Modelle für die einzelnen Zellenhäufigkeiten. Insbesondere gilt für dichotome Variablen: $\lambda_1^A = -\lambda_2^A$.

Effekte.

Die einzelnen Summanden in der obigen Gleichung bezeichnen wir als Effektparameter. Als Haupteffekte bezeichnen wir die Effekte mit einem Subskript, z.B. λ_i^A, als Interaktionseffekte zweiter Ordnung jene mit zwei Subskripten, z.B. λ_{jk}^{BC}, und als Interaktioneffekte dritter Ordnung jene mit drei Subskripten: λ_{ijk}^{ABC}.

Ein Interaktionseffekt zweiter Ordnung λ_{jk}^{BC} lässt sich als Beziehung (Korrelation) zwischen B and C interpretieren. Ein Interaktionseffekt dritter Ordnung λ_{ijk}^{ABC} ist zu interpretieren als ein Zusammenhang zwischen B und C, dessen Stärke von der Ausprägung der Variable A abhängig ist.

Ein log-lineares Modell für eine n-dimensionale Tabelle, das alle möglichen Effekte erster, zweiter bis n-ter Ordnung enthält nennen wir ein *saturiertes Modell.* Wie bereits erwähnt, lassen sich die Effektparameter des saturierten Modells aus den beobachteten Werten einer beliebigen mehrdimensionalen Kreuztabelle direkt berechnen. Dabei besteht immer eine perfekte Übereinstimmung zwischen den Daten in der Krauztabelle und den Werten F_{ijk}, die sich durch das log-lineare Modell berechnen lassen; das saturierte Modell ist mithin nur eine andere Darstellung der mehrdimensionalen Kreuztabelle. Da das saturierte Modell dieselbe Information enthält wie die Kreuztabelle, eignet es sich allerdings ebenso wenig zur Interpretation. Vor allem die Interaktionen höherer Ordnung machen das Modell rasch unübersichtlich.

Die Suche nach einfacheren Modellen.

Das Ziel der Analyse von Kreuztabellen mit Hilfe log-linearer Modell ist deshalb, durch Weglassen möglichst vieler Effekte des saturierten Modells ein einfacheres (sparsameres) Modell zu bilden, das noch hinreichend gut mit den beobachteten Daten übereinstimmt. Ein solches Modell könnte etwa so aussehen:

$$\log F_{ijk} = \theta + \lambda^{A}_{i} + \lambda^{B}_{j} + \lambda^{C}_{k} + \lambda^{AB}_{ij} + \lambda^{BC}_{jk}.$$

In diesem Modell sind gegenüber dem saturierten Modell die Effekte

$$\lambda^{ABC}_{ijk}, \lambda^{AC}_{ik}$$

weggefallen. Dies bedeutet: es besteht ein Zusammenhang zwischen B und C (Geschlecht und Drogenerfahrung) und einer zwischen B und A (Geschlecht und Einkommen der Eltern). Hingegen besteht kein Zusammenhang zwischen A und C (Einkommen der Eltern und Drogenerfahrung) und der Zusammenhang zwischen A und B ist auch nicht abhängig von C (Drogenerfahrung). Die Techniken der Analyse log-linearer Modelle und die entsprechenden Computerprogramme (Dixon, 1976) dienen dazu, solche einfacheren Modelle zu finden und sie auf ihre Übereinstimmung mit den Daten zu prüfen.

Zusammenhang zwischen angepassten Randverteilungen und berücksichtigten Effekten.

Es besteht eine eineindeutige Zuordnung zwischen jenen Effekten, die in eieem Modell berücksichtigt worden sind und jenen Randverteilungen, die durch das Modell exakt wiedergegeben werden (*fitted marginals*). Im obigen Modell, das die Effekte $\lambda^{AB}_{ij}, \lambda^{BC}_{jk}$ enthält, sind die Randverteilungen AB und BC exakt enthalten. Wegen dieser eineindeutigen Zuordnung kann man ein Modell statt durch die darin enthaltenen Effekte auch dadurch beschreiben, dass man angibt, welche Randverteilungen im Modell exakt angepasst sind. Im obigen Modell sind dies: A, B, C, AB, BC.

Das hierarchische Prinzip.

Eine grundlegende Eigenschaft der log-linearen Modelle ist ihre hierarchische Struktur. Wenn ein Modell einen Effekt *n*-ter Ordnung enthält, so impliziert dies, dass alle darin enthaltenen Effekte geringerer Ordnung im Modell ebenfalls vorkommen müssen. Kommt der Interaktionseffekt dritter Ordnung ABC vor, so heisst das, dass die Interaktionen AB, AC und BC im Modell ebenfalls enthalten sein müssen. Der Interaktionseffekt ABC enthält also nur jenen Effekt, der über das hinausgeht, was die Interaktionseffekte AB, AC und BC zusammen erklären. Dieses hierarchische Prinzip ist unmittelbar einsichtig, wenn man an die eineindeutige Zuordnung von Effekten und angepassten Randverteilungen denkt: aus der Randverteilung der drei Variablen A B C kann man alle darin enthaltenen Randverteilungen berechnen; aus diesem Grunde impliziert das Vorhandensein des Effektes λ^{ABC}_{ijk} auch das Vorhandensein folgender Effekte:

$$\lambda^{A}_{i}, \lambda^{B}_{j}, \lambda^{C}_{k}, \lambda^{AB}_{ij}, \lambda^{AC}_{ik}, \lambda^{BC}_{jk}.$$

Die Existenz des hierarchischen Prinzips bei den log-linearen Modellen erlaubt es übrigens, die in einem Modell enthaltenen Effekte abgekürzt durch die angepassten Randverteilungen höchster Ordnung zu beschreiben: ein Modell, bei dem die Randverteilungen ABC und CD angepasst sind, enthält folgende Effekte A,B,C,D, AB,AC,BC,CD, ABC.

3. DIE ZWEI HAUPTSCHRITTE IN DER ANALYSE

Das Verfahren der log-linearen Modelle impliziert zwei Analyseschritte, die nacheinander auszuführen sind:

(1) Die Suche nach dem optimalen Modell, d.h. nach einem Modell, das eine ausreichende Erklärungskraft bei möglichst einfacher Form und möglichst guter theoretischer Interpretierbarkeit bietet. Das optimale Modell ist bestimmt, wenn man weiss, welche Effekte in diesem Modell enthalten sind und welche nicht.

Die Schätzung der Modellparameter, d.h. der im Modell aufgenommenen Effekte. Bei nicht saturierten Modellen lassen sich die Effektparameter nicht direkt durch Umformung der Ausgangsdaten gewinnen. Vielmehr sind maximum-likelihood-Schätzmethoden notwendig (Fienberg, 1977).

Aus den Effektparametern kann auch auf die relative Stärke der einzelnen im Modell vorhandenen Effekte geschlossen werden. Ausserdem können die Resultate in einer Weise dargestellt werden, die es erlaubt, die relative Erklärungskraft eines Modells oder einzelner Effekte eines Modells in Beziehung zu setzen mit einem Basismodell. Das Basismodell entspricht der Nullhypothese und ist in der Regel jenes Modell, das das Fehlen von Zusammenhängen zwischen den Variablen postuliert.

3.1 Identifikation eines Modells

Die Anpassung (fit) eines Modells an die Daten wird durch die Abweichung der auf Grund des Modelles geschätzten Zellenwerte von den beobachteten Zellenwerten mittels eines verallgemeinerten Chiquadrattests geprüft; es handelt sich um das likelihood ratio-Chiquadrat. Besteht eine signifikante Abweichung zwischen Modell und Daten, so ist das Modell nicht adäquat.[2]

Meist geht es nicht darum, ein theoretisch erwartetes Modell anhand von Daten zu überprüfen, sondern man möchte aus den vorhandenen Daten ein Modell erschliessen, das die Daten optimal erklärt. Wie erwähnt heisst optimal: möglichst einfach bei hinreichend guter Übereinstimmung mit den Daten. Es besteht keine eindeutige Suchstrategie für das optimale Modell. In jedem Fall wird die Suchstrategie in einer schrittweisen Analyse bestehen, indem ein bestimmtes Modell mit verschiedenen "benachbarten" Modellen verglichen wird. Dabei sind Vorwärts- und Rückwärtsstrategien möglich, d.h. ein Vergleich mit Modellen, die jeweils einen Effekt mehr bzw. weniger enthalten. Falls zwischen zwei Modellen keine

[2] In der Literatur werden keine minimalen Zellenbesetzungen für die mehrdimensionalen Kreuztabellen gefordert. Besondere Aufmerksamkeit ist jedoch dem Vorhandensein von leeren Zellen zu schenken. Dabei ist zu unterscheiden zwischen definitionsgemäss leeren Zellen (structural zeros – beispielsweise militärischer Grad von Frauen) und zufällig leeren Zellen (sampling zeros), die deshalb zustande kommen, weil von einer seltenen Merkmalskombination keine Einkeit in der Stichprobe enthalten ist. Die zufällig leeren Zellen sind solange nicht problematisch, als sie nicht die Berechnung von erwarteten Zellenwerten von Null erzwingen. In diesem Falle sowie bei definitionsgemäss leeren Zellen muss eine Korrektur der Freiheitsgrade angebracht werden. Für eine eingehende Diskussion s. Fienberg, 1977, Kap. 8.

signifikante Differenz in der Erklärungskraft besteht, wird das einfachere Modell bevorzugt. Falls in Modell signifikant mehr Abweichung erklärt, so wird jenes Modell mit der grösseren Erklärungskraft gewählt (genaues Vorgehen: es werden die Differenzen zwischen zwei likelihood ratio-Chiquadraten auf ihre Signifikanz geprüft).

Aus praktischen Erwägungen ist es nicht möglich, alle theoretisch existierenden Modelle miteinander zu vergleichen. Zwei Strategien sind geeignet, das optimale Modell zielstrebig zu identifizieren.

3.1.1 Methode der Effekte gleicher Ordnung

Zunächst werden nur jene Modelle miteinander verglichen, bei denen *alle* Effekte gleicher Ordnung enthalten sind. Im obigen Beispiel mit drei Variablen würde das bedeuten, dass die folgenden Modelle miteinander verglichen werden:

$$(1)\ A,B,C \quad : \log F_{ijk} = \theta + \lambda_i^A + \lambda_j^B + \lambda_k^C$$

$$(2)\ AB,AC,BC \quad : \log F_{ijk} = \theta + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{ik}^{AC} + \lambda_{jk}^{BC}$$

$$(3)\ ABC \quad : \log F_{ijk} = \theta + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{ik}^{AC} + \lambda_{jk}^{BC} + \lambda_{ijk}^{ABC}.$$

Durch diesen Vergleich wird man das Modell höchster Ordnung, das die Abweichung noch ungenügend erklärt, identifizieren, und das Modell niedrigster Ordnung, das für eine genügende Anpassung ausreicht. Es kann nun angenommen werden, dass das optimale Modell zwischen diesen beiden Modellen liegt. Durch systematische Hinzunahme bzw. Elimination von Effekten wird man das optimale Modell finden.

3.1.2. Methode der standardisierten Effekte

Ausgehend vom saturierten Modell berechnet man alle standardisierten Effektparameter (Effektparameter dividiert durch den Standardfehler). Diese geben einen Hinweis auf den relativen Einfluss jedes einzelnen Effekts. Als erste Näherung wird man nun ein Modell wählen, in dem nur die grössten Effekte enthalten sind. Von diesem ersten Modell aus wird man durch Vergleich mit "benachbarten" Modellen eine Annäherung ans optimale Modell anstreben.

3.2. Vergleich der Stärke der einzelnen Effekte im Modell

In multivariaten Verfahren ist man daran interessiert, die relative Stärke der einzelnen Zusammenhänge abzuschätzen. Bei den log-linearen Modellen bieten sich hierzu zwei Möglichkeiten.

3.2.1. Vergleich der Effektparameter

Die eine besteht darin, die im Modell vorhandenen Effektparameter λ miteinander zu vergleichen. Dabei empfiehlt es sich, die standardisierten Effektparameter zu verwenden (Effektparameter dividiert durch den Standardfehler). Je grössen ein Effektparameter ist, desto grösser auch der Einfluss des entsprechen-

den Effektes im Modell. Die Effektparameter können nur innerhalb eines Modells in bezug auf ihre Stärke miteinander verglichen werden. Ein weiterer Nachtteil bei der Betrachtung von Effektparametern liegt darin, dass für die Variablen mit mehr als zwei Merkmalsausprägungen kein einheitlicher Effektparameter berechnet wird, sondern je einer für jede Merkmalsausprägung bzw. für jede Kombination von Markmalsausprägungen bei Interaktionen höherer Ordnung, was ziemlich rasch zu unübersichtlichen Resultaten führt. Bei dichotomen Merkmalen unterscheiden sich die Effektparameter für die beiden Kategorien nur durch das Vorzeichen, deshalb genügt die Angabe des Effektparameters für eine Kategorie.

Die unstandardisierten Effektparameter lassen sich in Analogie zu den adjustierten Mittelwertsabweichungen der Mehrweg-Varianzanalyse leicht interpretieren: Sie geben an, um wieviel sich der Logarithmus einer Zellenhäufigkeit F_{ijk} ändert durch den entsprechenden Effekt, da ja die Summe *aller* Effekte gerade gleich dem geschätzten Wert für den Logarithmus der Zellenhäufigkeit F_{ijk} ist.

3.2.2. Determinationskoeffizienten

Wie bereits erwähnt, misst das likelihood ratio Chiquadrat die Abweichung eines bestimmten Modells von den beobachteten Daten. Das likelihood-ratio Chiquadrat eines bestimmten Modelles kann in mehrere additive Komponenten aufgeteilt werden, die den im Modell enthaltenen Effekten zugeordet werden können. Auf diese Weise ist es möglich, die "Abweichung" eines bestimmten Modells von den Daten in ähnlicher Weise wie bei der Varianzanalyse auf einzelne Komponenten aufzuteilen. Tabelle 3 gibt ein Beispiel hierfür, das sich auf verschiedene Modelle für die in Tabelle 1 gezeigte dreidimensionale Kreuztabelle bezieht.

Goodman (1972, S. 42) schlägt ausserdem die Berechnung eines Determinationskoeffizienten als Mass für die Güte eines Modells vor (goodness of fit). Der Determinationskoeffizient hat eines strukturelle Ähnlichkeit zum quadrierten multiplen Korrelationskoeffizienten (Mass für die erklärte Varianz) in der varianzanalytischen Statistik. Anstelle des Begriffs der Varianz tritt hier allerdings der Begriff der Abweichung des Modells von den beobachteten Daten.

Während in der varianzanalytischen Statistik die totale Varianz einer Variablen als Bezugsgrösse für die erkläre Varianz dient, verwendet man beim Determinationskoeffizienten in der Analyse log-linearer Modelle die Abweichung des Modells von den Daten bei einer Nullhypothese. Die Nullhypothese wird meist ein Modell sein, bei dem das Fehlen von Beziehungen zwischen den Variablen postuliert wird. In unserem Drei-Variablen-Beispiel also die Abweichung: X^2 (Modell 1). Das Modell M 4 (AB, AC, BC) bewirkt eine geringere Abweichung von den Daten. Die Differenz zwischen den beiden Chiquadrats

$$X^2 \text{ (Modell 1)} - X^2 \text{ (Modell 4)}$$

kann als Verbesserung der Anpassung des Modells M 4 gegenüber dem Modell M 1 gelten. Diese Verbesserung in der Uebereinstimmung kann in Form eines Quotien-

Tabelle 3.

Nr.	Modell	likelihood ratio X^2	df	p
1	A, B, C,	632.71	4	0.0
2	A, BC	346.10	3	0.0
3	AC, BC	125.15	2	0.0000
4	AB, AC, BC	10.35	1	0.0000
5	ABC	0	0	1.0

Analyse der "Abweichung" in der dreidimensionalen Kreuztabelle ABC (Tabelle 1)

Quelle der Abweichung	Nr. d. Modells	df	likelihood-ratio X^2
1. Abweichung aufgrund aller Interaktionseffekte	(1)	4	632.71
2. durch Modell (3) nicht erklärte Abweichung	(3)	2	125.15
3. durch Modell (3) erklärte Abweichung	(1) – (3)	2	507.56
Aufteilung von 2.			
2a. durch Modell (4) nicht erklärte Abweichung	(4)	1	10.35
2b. durch Effekt AB in Modell (4) erklärte Abweichung	(3) – (4)	1	114.80

ten in Beziehung gesetzt werden zur Abweichung unter der Nullhypothese. Diese Grösse bezeichnet Goodman als Determinationskoeffizient:

$$\frac{X^2\,(M\,1) - X^2\,(M\,4)}{X^2\,(M\,1)}$$

Der Begriff der erklärten Abweichung kann in analoger Weise verwendet werden, um die Erklärungskraft eines einzelnen Effekts in einem Modell zu bestimmen: *Der partielle Determinationskoeffizient* ist ein Mass dafür, wieviel ein einzelner Effekt erklärt, wenn man alle übrigen Effekte des Modells mit gleicher oder geringerer Ordnung konstant hält. Im Zähler steht die Differenz in der Abweichung zwischen zwei Modellen, die sich nur durch den interessierenden Effekt unterscheiden. Im Nenner steht wiederum die totale zu erklärende Abweichung, d.h. die Abweichung unter der Nullhypothese[3]. So ist der partielle Determinations-

[3] Die Bezugsgrösse im Nenner wird bei Goodman (1972) anders definiert. Die hier verwendete Definition wird aber von verschiedenen Autoren in der Literatur verwendet (Duncan-Jones, 1976; Hauser *et al.*, 1975; Stolzenberg *et al.*, 1977). Die Verwendung der Abweichung unter der Nullhypothese im Nenner weist eine grössere Analogie zum partiellen Korrelationskoeffizienten auf.

koeffizient von ABC:

$$\frac{X^2\,(M4) - X^2\,(M5)}{X^2\,(M1)}$$

4. LOGIT-MODELLE UND ODDS RATIOS

Bis hierin haben wir in unseren Überlegungen nicht zwischen abhängigen und unabhängigen Variablen unterschieden, sondern die Beziehungen zwischen allen Variablen in einer mehrdimensionalen Kreuztabelle untersucht. Ein wichtiger Spezialfall der log-linearen Modelle bezieht sich nun auf den Fall einer beliebigen Anzahl unabhängiger Variablen und einer abhängigen Variable, die nur zwei Ausprägungen kennt, etwa C in unserem Beispiel: Drogenerfahrung ja/Drogenerfahrung nein. Dieser Spezialfall, der als Logit-Modell bezeichnet wird, vereinfacht das Arbeiten mit log-linearen Modellen.

Während in der bisherigen Betrachtung Beziehungen zwischen beliebigen Variablen interessiert haben, werden beim Logit-Modell nur Beziehungen zwischen den unabhängigen Variablen einerseits und der abhängigen Variable andererseits betrachtet. Die Beziehungen unter den unabhängigen Variablen interessierten nicht. Aus diesem Grund wird beim Logit-Modell nicht versucht, ihre Struktur aufzuklären, sondern die Beziehungen unter den unabhängigen Variablen werden als fix angenommen und durch das saturierte Submodell für die unabhängigen Variablen dargestellt.

Anstelle der relativen Häufigkeit des Auftretens eines dichotomen Merkmals (z.B. 21,7 % Personen mit Drogenerfahrung in Tabelle 2) kann äquivalent das odds ratio für diese Kategorie angegeben werden (2143/7730 = 0,28 ist das odds ratio, Drogenerfahrung zu haben). Beim Logit-Modell geht es darum, ein Modell für das odds ratio der Zellenfrequenzen der abhängigen Variable zu gewinnen – und nicht wie bisher ein Modell für die einzelnen Zellenfrequenzen. Es resultiert folgende Gleichung:

$$\log \frac{F_{ij1}}{F_{ij2}} = \log F_{ij1} - \log F_{ij2} = \theta + \lambda_i^A + \lambda_j^B + \lambda_l^C + \lambda_{ij}^{AB} + \lambda_{il}^{AC} + \lambda_{jl}^{BC} + \lambda_{ijl}^{ABC} - (\theta + \lambda_i^A + \lambda_j^B + \lambda_2^C + \lambda_{ij}^{AB} + \lambda_{i2}^{AC} + \lambda_{j2}^{BC} + \lambda_{ij2}^{ABC}$$

Da bei dichotomen Variablen die Parameter für die beiden Kategorien sich nur durch das Vorzeichen unterscheiden, kann wie folgt vereinfacht werden:

$$\log \frac{F_{ij1}}{F_{ij2}} = \lambda_1^C + 2 \cdot \lambda_{il}^{AC} + 2 \cdot \lambda_{jl}^{BC} + 2 \cdot \lambda_{ij1}^{ABC}$$

Ersetzt man $2\,\lambda = \beta$, so vereinfacht sich die Gleichung für Logit-Modell zu

$$\log \frac{F_{ij1}}{F_{ij2}} = \beta_1^C + \beta_{i1}^{AC} + \beta_{j1}^{BC} + \beta_{ij1}^{ABC}.$$

Tabelle 4

YEAR Y	SEX S	ORTGR U	BILDUNG B	EINKVAT E	DROGKONT(D) JA	NEIN
71.	MANN	<10000	TIEF	TIEF	104	717
				HOCH	21	58
			HOCH	TIEF	104	458
				HOCH	74	135
		>10000	TIEF	TIEF	89	371
				HOCH	20	44
			HOCH	TIEF	101	324
				HOCH	75	118
		ZUERICH	TIEF	TIEF	132	377
				HOCH	18	41
			HOCH	TIEF	222	408
				HOCH	92	145
	FRAU	<10000	TIEF	TIEF	5	39
				HOCH	0	0
			HOCH	TIEF	11	119
				HOCH	10	49
		>10000	TIEF	TIEF	2	32
				HOCH	0	4
			HOCH	TIEF	10	77
				HOCH	7	37
		ZUERICH	TIEF	TIEF	7	56
				HOCH	1	5
			HOCH	TIEF	29	138
				HOCH	19	52
78.	MANN	<10000	TIEF	TIEF	43	259
				HOCH	48	206
			HOCH	TIEF	39	154
				HOCH	104	347
		>10000	TIEF	TIEF	58	207
				HOCH	39	186
			HOCH	TIEF	40	135
				HOCH	92	338
		ZUERICH	TIEF	TIEF	62	152
				HOCH	50	119
			HOCH	TIEF	42	97
				HOCH	66	241
	FRAU	<10000	TIEF	TIEF	11	132
				HOCH	14	68
			HOCH	TIEF	12	95
				HOCH	65	172
		>10000	TIEF	TIEF	14	115
				HOCH	13	83
			HOCH	TIEF	18	97
				HOCH	33	197
		ZUERICH	TIEF	TIEF	13	99
				HOCH	17	65
			HOCH	TIEF	28	142
				HOCH	69	220

Dieser Spezialfall des Logit-Modells weist die grösste Analogie zur multiplen Regression auf, wie Goodman (1972) eingehend dargelegt hat.

5. BEISPIEL FÜR EIN LOGIT-MODELL: DIE VERÄNDERUNG DES DROGENKONSUMS IM KANTON ZÜRICH 1971/1978

Die in Tabelle 4 dargestellten Daten sind der Untersuchung von Binder *et al.* (1979) entnommen. 1971 und 1978 wurden teils in Vollerhebungen, teils anhand von repräsentativen Stichproben, 19- bis 20-jährige Männer und Frauen im Kanton Zürich mit einem schriftlichen Fragebogen u.a. über ihren Konsum von illegalen Drogen befragt[4]. Im folgenden wird als Drogenerfahrung (DROGKONT) die mindestens einmalige Einnahme eines der folgenden Stoffe verstanden: Haschisch, Halluzinogene, Weckamine, Opiate. Die Drogenerfahrung ist in Tabelle 4 wie folgt aufgegliedert:

Y – YEAR (Erhebungsjahr): 1971/1978;
S – SEX (Geschlecht): männlich/weiblich
U – ORTGR (Urbanisierung des Wohnorts): bis 10 000 Einwohner/
10 001-100 000/über 100 000 Einwohner = Stadt Zürich;
B – BILDUNG (Schulbildung): tief = Ober- und Realschule
hoch = Sekundar- u. Mittelschule;
E – EINKVAT (Einkommen der Eltern): tief = bis Fr. 2 000.–,
hoch = über Fr. 2 000.–;
D – DROGKONT (Drogenerfahrung): ja/nein.

Die Fragen, die bei der Analyse dieser Tabelle zu beantworten sind, lauten:
1. Hat sich die Häufigkeit von Drogenerfahrungen von 1971 bis 1978 verändert?
2. Hat sich der Zusammenhang zwischen Drogenerfahrung und sozialen Hintergrundsvariablen im selben Zeitraum verändert?

Die Untersuchung von Binder *et al.* (1979) konnte durch konventionellen Vergleich von Kreuztabellen in verschiedenen Untergruppen zeigen, dass sich weniger die Verbreitung der Drogenerfahrung insgesamt als vielmehr der soziale Hintergrund der Personen mit Drogenerfahrungen im Laufe der Untersuchungsperiode verändert hat. Eine exakte Beurteilung erlaubt aber erst die mehrdimensionale Kreuztabellenanalyse wie sie im folgenden durchgeführt wird.

Tabelle 5

Nr.	Modell	likelihood-ratio X^2	df	p
(1)	YSUBE, YSD, YED, SED, UED, BD	41.24	36	0.25
(2)	YSUBE, YSED, UED, YBD	31.58	34	0.586
(3)	YSUBE, D	348.67	47	0.0
(4)	YSUBE, YSD, YED, SED, UED, YBD	36.80	35	0.38

[4] Die Studien wurden durch den Schweizerischen Nationalfonds unterstützt.

Wir gehen aus vom Anfangsmodell (1), das wir mit dem Screening-Test von Brown als erstes Näherungsmodell identifiziert haben. Dieses Modell enthält nur Zwei-Weg-Interaktionen.[5] Durch systematisches Hinzufügen signifikanter und Eliminieren nicht-signifikanter Effekte erreichen wir schliesslich das optimale Modell (2). Der Determinationskoeffizient für dieses Modell ergibt sich aus der Differenz zwischen den Abweichungen des Modells ohne Prädiktoren (3) und des optimalen Modells (2) dividiert durch die Abweichung beim Modell ohne Prädiktoren (3):

Determinationskoeffizient Modell (2) = 0.909

Der partielle Determinationskoeffizient für den Effekt YSED ergibt sich als Differenz zwischen der Abweichung des Modells (4), das den Effekt nicht enthält und der Abweichung des optimalen Modells (2) dividiert durch das Nullmodell (3):

partieller Determinationskoeffizient YSED = 0.014.

Tabelle 6. Effekte im optimalen Modell: YSUBE, YSED, UED, YBD

Effekt	β	stand. β	part. Determ.-koeffizient
Grand mean (β_1^C)	− 1.436	− 43.914	
Y (78)	− 0.036	1.116	0.033
S (Mann)	0.316	8.576	0.200
U (− 10 000)	− 0.192	− 3.710	
(10-100 000)	− 0.074	− 1.594	0.225
(Zürich)	0.266	6.702	
B (hoch)	0.138	4.212	0.082
E (hoch)	0.216	6.588	0.105
Y (78) S (Mann)	− 0.102	− 3.100	0.020
Y (78) (hoch)	− 0.054	1.660	0.012
Y (78) E (hoch)	− 0.080	2.422	0.060
S (Mann) E (hoch)	− 0.090	− 2.744	0.038
E (hoch) U (− 10 000)	0.204	3.934	
(10-100 000)	− 0.066	− 1.414	0.100
(Zürich)	− 0.138	− 3.456	
Y (78) S (Mann) E (hoch)	− 0.080	2.476	0.014

Bei der Beurteilung der Grösse der einzelnen Effekte im Modell halten wir uns an die Rangfolge der Determinationskoeffizienten. Diese zeigt, dass der Urbanisierungsgrad den grössten Einfluss hat auf die Drogenerfahrung, es folgt das Geschlecht, der soziale Status der Eltern, dann bereits der erste Interaktionseffekt zwischen Urbanisierung und Einkommen der Eltern usw. Auffallend ist der relativ geringe Effekt des Erhebungsjahres, d.h. der gesamte Drogenkonsum hat sich in

[5] Im Logit-Modell, wo eine Variable als abhängige angesehen wird, bezeichnen wir Effekt zwischen einer unabhängigen und der abhängigen Variable als Haupteffekte, Interaktionen zwischen zwei unabhängigen und der abhängigen Variable als Interaktionseffekte zweiter Ordnung usw.

der Vergleichsperiode nur wenig geändert. Es ist noch zu bemerken, dass die Rangfolge der Determinationskoeffizienten nicht vollstandig mit der Rangreihenfolge der standardisierten Effektparameter übereinstimmt. Dies ist darauf zurückzuführen, dass bei Variablen mit mehr als zwei Ausprägungen je nach Kategorie verschiedene Effektparameter berechnet werden, die keine eindeutige Aussage über den Gesamteffekt der Variablen erlauben.

Die einzelnen Effekte des Modells können wie folgt interpretiert werden. Der Anteil der 19-jährigen mit Drogenerfahrung ist im Jahr 1978 um einen minimen Betrag zurückgegangen. Männer haben wesentlich häufiger Drogenerfahrungen. Drogenerfahrungen sind in der Stadt am verbreitesten, in Gemeinden mit weniger als 10 000 Einwohner am seltensten. Höhere Schulbildung führt generell zu mehr Drogenerfahrung, und Kinder von Eltern mit höherem Einkommen haben ebenfalls häufiger Drogenerfahrung. Bei all diesen Beziehungen ist anzumerken, dass sie unter Kontrolle aller übrigen Beziehungen berechnet sind, d.h. z.B. dass hohe Schulbildung und hoher Status der Eltern einen unabhängigen Effekt auf den den Drogenkonsum der Jugendlichen haben. Die bisher besprochenen Beziehungen sind für beide Jahre 1971 und 1978 gültig. Die im folgenden zu besprechenden Interaktionen zwischen dem Erhebungsjahr, sozialen Daten und Drogenkonsum zeigt hingegen Änderungen in den Beziehungen zwischen sozialem Hintergrund und Drogenerfahrung im untersuchten Zeitintervall an. Das Ueberwiegen von Männern bei Jugendlichen mit Drogenerfahrung hat im Jahre 1978 abgenommen (Effekt YS). Ebenso hat der Zusammenhang zwischen Drogenerfahrung und Schulbildung bzw. sozialem Status der Eltern sich 1978 abgeschwächt (Interaktionseffekte YB bzw. YE sind negativ).

Die Variable "Einkommen der Eltern" ist in zwei weiteren Interaktionseffekten enthalten: der Effekt SE lässt sich dahingehend interpretieren, dass der Zusammenhang zwischen hohem sozio-ökonomischen Status der Eltern und der Drogenerfahrung für Männer schwächer ist als für Frauen. Betrachtet man auch noch den Interaktionseffekt YSE, so zeigt sich, dass die Abschwächung dieses Zusammenhangs bei den Männern 1978 besonders deutlich war. Addiert man die Effekte SE und YSE, so zeigt sich, dass bei den Männern 1978 der Effekt des elterlichen Status durch die beiden genannten Interaktionseffekte beinahe aufgehoben wird. Dies stimmt überein mit dem Befund in Binder *et al.* (1979), wo mit einfacheren statistischen Methoden ebenfalls festgestellt worden ist, dass 1978 bei den Männern kaum mehr ein Zusammenhang zwischen Drogenerfahrung und sozialem Status der Eltern besteht. Der Interaktionseffekt EU kann dahingehend interpretiert werden, dass in den wenig urbanisierten Gemeinden vor allem die Jugendlichen aus höheren sozialen Schichten Drogenerfahrungen haben, während in der Stadt Drogenerfahrung eher ein Verhaltensmuster der Jugendlichen aus unteren sozialen Schichten ist.

Gesamthaft führt die Interpretation der sechsdimensionalen Kreuztabelle bezüglich sozialer Variablen und Drogenkonsum für zwei Erhebungszeitpunkte zur Schlussfolgerung, dass in der beobachteten Erhebungsperiode praktisch kein Rückgang der Drogenerfahrung stattgefunden hat. Hingegen haben sich die sozialen

Merkmale der Jugendlichen mit Drogenerfahrungen stark verändert: Jugendliche aus höheren sozialen Schichten haben keinen Vorsprung mehr bezüglich Drogenerfahrungen: tendenziell kann also von einer Nivellierung des Drogenkonsums bezüglich des Geschlechts und auch bezüglich des sozio-ökonomischen Status gesprochen werden.

6. ABSCHLIESSENDE BEMERKUNGEN

In dieser Arbeit ist es darum gegangen, die praktische Anwendung des Verfahrens der log-linearen Modelle zur Analyse mehrdimensionaler Kreuztabelle zu demonstrieren. Absichtlich wurde darauf verzichtet, einen Vergleich mit anderen multivariaten Analyseverfahren für kategoriale Daten durchzuführen (vgl. dazu Küchler, 1978; Kershner *et al.*, 1976; Goodman, 1976), oder auf Verbesserungen, Erweiterungen und spezielle Anwendungen des Verfahrens einzugehen. Mit dem Hinweis auf einige dieser neueren Entwicklungen soll jedoch dem Leser gezeigt werden, dass mit den hier demonstrierten Anwendungen die Möglichkeiten der Analyse von Kreuztabellen mit log-linearen Modellen noch keineswegs erschöpft sind:

(1) Das Verfahren ist auch anwendbar für mehrdimensionale Kreuztabellen mit strukturellen leeren Zellen (Fienberg, 1977, Kap. 8).

(2) Das Verfahren erlaubt nicht nur die Analyse von kategorialen sondern auch von ordinalen Daten. Dabei entsteht ein Informationsgewinn gegenüber der Behnadlung von ordinalen Variablen als kategorialen (Fienberg, 1977, Kap. 4).

(3) Das Verfahren der log-linearen Modelle eignet sich auch zur Kausalanalyse in einer Form, die rein äusserlich der konventionellen Pfadanalyse sehr ähnlich ist (Goodman, 1973).

BIBLIOGRAPHIE

ANDREWS, F.J., and MESSENGER, C. (1973), "Multivariate nominal scale analysis. A report on a new analysis technique and a computer program" (Ann Arbor, Michigan).

BINDER, J., SIEBER, M. und ANGST, J. (1979), "Entwicklung des Suchtmittelkonsums bei 19/20 jährigen Jugendlichen. Ein Vergleich im Kanton Zürich 1971, 1974 und 1978", *Schweiz. med. Wschr.* 109 (1979). 1298-1305, 1331-1335.

BISHOP, Y.M.M., FIENBERG, S.E. and HOLLAND, P.W. (1975), "Discrete multivariate analysis: theory and practice" (MIT Press, Cambridge, Massachusetts).

BROWN, M.B. (1976), Screening effects in multidimensional contingency tables, *Appl. Statist.*, 25 (1976) 37-45.

DAVIS, J.A. (1974), Hierarchical models for significance tests in multivariate contingency tables: an exegesis of Goodman's recent papers, *Sociol. Methodol.* (Costner, H.L., Ed.) 189-231.

DIXON, W.J. (Ed.) (1977), BMDP-77, Biomedical computer programs, P-series, *Program: BMDP3F* (University of California Press, Berkeley).

DUNCAN-JONES, P. (1976), "Studying the odds: simple presentation of binary outcomes" (sociological association of Australia and New Zealand. Annual Conference 1976).

FIENBERG, S.E. (1977), "The analysis of cross-classified categorical data" (MIT Press, Cambridge, Massachusetts).

GOODMAN, L.A. (1972), A modified multiple regression approach to the analysis of dichotomous variables, *Am. Sociol. Rev.* 37 (1972) 28-46.

GOODMAN, L.A. (1973), The analysis of multidimensional contingency tables when some variables are posterior to others: a modified path analysis approach, *Biometrika* (1973) 179-192.

GOODMAN, L.A. (1976), The relationship between modified and usual multiple regression approaches to the analysis of dichotomous variables, *Sociol. Methodol.* (Heise, D., Ed.) (1976) 83-100.

GRIZZLE, J.E., STARMER, C.F. and KOCH, G.S. (1969), Analysis of categorical data by linear models, *Biometrics*, **25** (1969) 489-584.

HAUSER, R.M., KOFFEL, J.N., TRAVIS, H.P. and DICKINSON, P.J. (1975), Temporal change in occupational mobility: evidence for men in the United States, *Am. Sociol. Rev.*, **40** (1975) 279-297.

KERSHNER, R.P. and CHAO, G.C. (1976), A comparison of some categorical analysis programs, *Proc. Stat. Comp. Sec., Am. Stat. Assoc.* (1976) 178-184.

KÜCHLER, M. (1976), Multivariate Analyse nominal-skalierter Daten, *Z. f. Soziol.*, **5** (1976) 237-255.

KÜCHLER, M. (1978), Alternativen in der Kreuztabellenanalyse – Ein Vergleich zwischen Goodmans "General Model" (ECTA) und dem Verfahren gewichteter Regression nach Grizzle et al. (Nonmet II), *Z. f. Soziol.*, 7 (1978) 347-365.

STOLZENBERG, R.M. and AMICO, R.J.D. (1977), City differences and nondifferences in the effect of race and sex on occupational distribution, *Am. Sociol. Rev.*, **42** (1977) 937-950.