In: Allgemeines Statistisches Archiv, 1979, 63.

A. Abhandlungen

Datenanalyse und multiplikative Modelle

Von NANNY WERMUTH, Mainz

I. Einleitung

Eine Analyse von Assoziationen zwischen mehreren Variablen wird häufig gewünscht. In der Praxis beschränkt man sich jedoch entweder auf das bloße, gleichzeitige Betrachten einfacher Assoziationsmaße oder aber man setzt relativ komplexe Methoden wie die Hauptkomponenten- oder Faktorenanalyse ein. Bei dem zuletzt genannten Vorgehen versucht man, wenige neue Variable aus den zahlreichen, beobachteten Variablen zu definieren. Wünscht man stattdessen auf der Ebene der beobachteten Variablen zu bleiben, so bieten sich multiplikative Modelle als relativ einfache Zusammenhangstrukturen zur Datenbeschreibung an. Diese Modelle basieren auf dem statistischen Begriff der bedingten Unabhängigkeit und sie führen dazu, daß man Variablenuntergruppen ausweisen kann, die durch zusammengehörige Variablen gekennzeichnet sind.

Multiplikative Modelle beschreiben bestimmte Eigenschaften der Verteilungsfunktion eines Zufallsvektors. Diese Modelle werden erst seit einigen Jahren in der statistischen Fachliteratur beschrieben. So wurde etwa die Bezeichnung „multiplikatives Modell" im Jahr 1970 von Goodman für Multinomialverteilungen eingeführt und von verschiedenen Autoren (Bishop 1971, Bishop, Fienberg und Holland 1975, Wermuth 1976, 1978) weiter verwendet. Unter dem Namen „auflösbare Modelle" (decomposable models) sind dieselben Modelle von anderen Autoren (Kellerer 1964, Haberman 1974, Anderson 1974, Sundberg 1975, Darroch, Speed and Lauritzen 1977) diskutiert worden. Ziel dieser Arbeit ist zweierlei. Es soll veranschaulicht werden, wie multiplikative Modelle bei der Datenanalyse nutzen können (Abschnitt II) und es sollen einige der bisher bekannte Eigenschaften multiplikativer Modelle zusammengestellt werden (Abschnitt III). Zunächst führen wir nur kurz die Begriffe ein, die zum Verständnis der Datenbeispiele unerläßlich sind.

Definition: Ein multiplikatives Modell liegt vor, wenn sich die Dichte (oder die Wahrscheinlichkeitsfunktion) eines Zufallsvektors in nicht-trivialer Art in Randdichten (oder

-Wahrscheinlichkeitsfunktionen) faktorisieren läßt, und zwar dergestalt, daß es genügt, die Verteilungen von bestimmten Teilen des Zufallsvektors zu kennen, um die Verteilung des gesammten Vektors zu erhalten. Nehmen wir als Beispiel einen Vektor mit fünf Variablen $(X_1, X_2, X_3, X_4, X_5)$ an, dessen Dichte wie folgt faktorisiert werden kann:

$$f(x_1, x_2, x_3, x_4, x_5) = \frac{f_{1,4,5}(x_1, x_4, x_5)\, f_{2,4,5}(x_2, x_4, x_5)\, f_{3,4,5}(x_3, x_4, x_5)}{f_{4,5}(x_4, x_5)\, f_{4,5}(x_4, x_5)}.$$

Hier reicht es aus, die Dichten von drei Teilen des Zufallsvektors, von (X_1, X_4, X_5), (X_2, X_4, X_5) und (X_3, X_4, X_5), zu kennen, um sich daraus die Dichte des gesamten Vektors ableiten zu können. Indexgruppen, die diese wichtigen Teile des Gesamtvektors kennzeichnen, verwenden wir zur Modellbezeichnung. So sprechen wir in diesem Beispiel vom multiplikativen Modell 145/245/345.

Aus der Faktorisierungseigenschaft einer Verteilung ergeben sich Aussagen darüber, wie die einzelnen Variablen zusammenhängen. Einerseits kann man die Modellbezeichnung selbst als eine Liste von zusammengehörigen Variablengruppen interpretieren, andererseits wird mit der Modellbezeichnung auf Unabhängigkeitseigenschaften von Variablenpaaren und von Variablengruppen hingewiesen. Indexpaare, die in der Modellbezeichnung nicht gemeinsam vorkommen, kennzeichnen die bedingt unabhängigen Variablenpaare. In diesem Sinne beschreiben multiplikative Modelle bestimmte Zusammenhangsstrukturen. Solche Strukturen sind für die Datenanalyse attraktiv, weil sie zu vereinfachenden Beschreibungen und Zusammenfassungen von an sich komplexen Zusammenhängen führen.

II. Anwendungsbeispiele für multiplikative Modelle

Wir beschreiben drei unterschiedliche Ziele, die mit der Anpassung eines multiplikativen Modells an Beobachtungswerte verfolgt werden können. Damit viele Anwender statistischer Methodik sich angesprochen fühlen, verwenden wir Daten aus verschiedenen Sachgebieten, aus einer medizinischen, einer soziologischen und einer volkswirtschaftlichen Untersuchung. Es handelt sich jeweils um nichtexperimentell gewonnene Daten.

Die Beobachtungen liegen in der Form einer mehrdimensionalen Kontigenztafel und – bei den quantitativen Daten – als Korrelationsmatrix zusammengefaßt vor. Es wird jeweils unterstellt, daß die Daten als Stichprobe eines multinomialverteilten oder eines multivariat-normalverteilten Zufallsvektors angesehen werden können.

A. Urteil über zusammengehörige Variable

Die medizinischen Daten entstammen der prospektiven Studie „Schwangerschaftsverlauf und Kindesentwicklung", die 1964 begonnen wurde und von der Deutschen Forschungsgemeinschaft getragen wird (DFG-Forschungsbericht 1978). Aus der Gesamtuntersuchung betrachten wir an einer Teilfrage und zwar an der, ob starker Zigarettenkonsum während der Schwangerschaft die perinatale Mortalität der Neugeborenen erheblich erhöht, inwiefern ein Urteil über zusammengehörige Variablen nötig und möglich ist.

Sowohl die Einfluß- wie auch die Zielgröße erfassen relativ seltene Ereignisse. Etwa 10% der Schwangeren gaben an, mehr als 5 Zigaretten pro Tag zu rauchen und bei etwa 3% der ausgetragenen Schwangerschaften wird das Kind tot geboren oder stirbt innerhalb der ersten sieben Tage. Es muß mit zahlreichen Stör- oder Hintergrundfaktoren gerechnet werden, weil die Daten – aus offensichtlichen Gründen – nicht aus einem Experiment, sondern aus einer Erhebung stammen.

Bei dem Versuch, die Ausgangsfrage zu beantworten, sind zwei Risiken gegeneinander abzuwägen. Das Risiko, die Art der Abhängigkeit der Zielgröße von der Einflußgröße verfälscht wiederzugeben, weil wichtige Hintergrundfaktoren vernachlässigt werden, gegen das Risiko, durch eine zu detaillierte Darstellung reine Zufallsschwankungen als inhaltlich wichtige Beziehungen zu interpretieren. Selbst bei einer großen Zahl von Patientinnen, hier fast 6000, ist daher eine Entscheidung vonnöten, welche und wie viele der potentiellen Störfaktoren unbedingt berücksichtigt werden müssen. Wir schlagen vor, mit einem Suchverfahren (Wermuth 1976b) herauszufinden, welches multiplikative Modell sich den Beobachtungen gut anpassen läßt, und das Ergebnis der genannten Entscheidung zugrunde zu legen.

Aus inhaltlichen Überlegungen werden zunächst vier potentiell wichtige Hintergrundfaktoren ausgewählt und eine sechsdimensionale Kontingenztafel erstellt (siehe Tab. 1, 6).

Kurz zusammengefaßt werden bei diesen Variablen folgende Abhängigkeiten der Zielgröße erwartet: eine höhere Mortalitätsrate bei starkem Zigarettenrauchen, bei kurzer Schwangerschaftsdauer, bei älteren oder untergewichtigen Schwangeren und bei langen Anfahrtswegen zur Klinik. Die erwarteten Zusammenhänge der Einflußgröße Zigarettenrauchen mit den Hintergrundfaktoren waren: höherer Zigarettenkonsum bei kürzerer Schwangerschaftsdauer, bei jüngeren oder untergewichtigen Schwangeren und beim Wohnen in einer Großstadt.

Bei diesen sechs Variablen zeigt sich – als Ergebnis der Modellsuche – daß das multiplikative Modell 1234/456 ausgezeichnet mit den Beobachtungen zu vereinbaren ist, das heißt, daß sich nur geringe Abweichungen zwischen den beobachteten und den

Tabelle 1: Medizinische Variablen (Wermuth 1976c)

Variablennummer	Variablennamen	Klassenanzahl
1	Perinatale Mortalität	2
2	Schwangerschaftsdauer	3
3	Alter, Mutter	3
4	Zigarettenrauchen, Mutter	3
5	Wohnort in der Großstadt	2
6	Untergewichtigkeit, Mutter	2

Beobachtungszahl: 5945.
Gut passendes multiplikatives Modell: 1234/456 (LQ – χ^2 = 151,61 bei 161 Freiheitsgraden).

unter den Modellannahmen geschätzten Häufigkeiten in der mehrdimensionalen Kontingenztafel ergeben.

Das Ausmaß der Abweichungen wird anhand einer Likelihood-Quotienten-Prüfgröße beurteilt (vgl. Abschnitt III). Das Modell weist Variablen in zwei Gruppen, in Gruppe 1234 und in Gruppe 456 als zusammengehörig aus. Ferner sind die Variablen 5,6 bedingt unabhängig von den Variablen 1,2 und 3, gegeben die Variable 4. Damit ist begründbar, daß die beiden Hintergrundfaktoren Untergewichtigkeit (Variable 5) und Wohnortgröße (Variable 6) nicht mehr berücksichtigt werden, wenn die Abhängigkeit der perinatalen Mortalität (Variable 1) vom Zigarettenrauchen (Variable 4) dargestellt wird (vgl. Wermuth 1976c).

B. Interpretation als Zusammenhangsstruktur

An den folgenden soziologischen Daten wird die inhaltliche Interpretation eines gut passenden multiplikativen Modells als Zusammenhangsstruktur dargestellt. Die Beobachtungen stammen aus einer Querschnittsuntersuchung (Goldberg 1971) zur Frage von soziologischen Bestimmungsgründen für das Wahlverhalten. Für 625 Wähler bei der Präsidentschaftswahl in den Vereinigten Staaten von Amerika im Jahr 1954 liegen Informationen über sechs Variable vor (siehe Tabelle 2): über die Wahlentscheidung und über die Zufriedenheit des Wählers mit der Regierungspolitik vor der Wahl; darüberhinaus ist über den Wähler selbst wie auch über dessen Vater bekannt, welcher politischen Partei er nahesteht und welcher sozioökonomischen Schicht er angehört.

Der Autor verwendet die Daten in der Form einer Korrelationsmatrix und versucht zu beurteilen, welche alternativen Kausalmodelle für das Wahlverhalten von den Beob-

Tabelle 2: Soziologische Variablen (Goldberg 1971)

Variablennummer	Variablennamen
1	Wahlentscheidung
2	Zufriedenheit mit Regierungspolitik
3	Parteiidentifikation, Wähler
4	Sozialstatus, Wähler
5	Parteiidentifikation, Vater
6	Sozialstatus, Vater

Beobachtungszahl: 645.
Gut passendes multiplikatives Modell: 123/3456 (LQ – χ^2 = 10,48 bei 6 Freiheitsgraden).

achtungen eher gestützt, welche eher widerlegt werden. Wir versuchen wiederum herauszufinden, ob ein einfaches multiplikatives Modell die Daten hinreichend gut beschreibt, so daß nicht mehr alle paarweisen Korrelationen als wichtig angesehen werden müssen, sondern nur noch ein Teil derselben. Es zeigt sich dabei, daß ein gut passendes Modell auch Aussagen über die Bestimmungsgründe des Wahlverhaltens mit einschließt.

Tabelle 3: Korrelationen der soziologischen Variablen

Variablennummer	1	2	3	4	5	6
1	1,00	0,74	0,72	0,27 (0,29)	0,47 (0,43)	0,28 (0,29)
2	0,47	1,00	0,71	0,29 (0,29)	0,45 (0,43)	0,32 (0,30)
3	0,36	0,33	1,00	0,41	0,60	0,40
4	0,02	0,05	0,14	1,00	0,42	0,81
5	0,07	0,01	0,33	0,03	1,00	0,45
6	0,03	0,09	0,02	0,12	0,16	1,00

Obere Hälfte: r_{ij}, einfache beobachtete Korrelationen.
In Klammern: r^*_{ij}, einfache, durch Modell 123/3456 implizierte Korrelationen.
Untere Hälfte: $r_{ij.klrs}$, partielle Korrelationen.

Als ein gut passendes Modell ergibt die Modellsuche dasjenige mit der Bezeichnung 123/3456. Gut passend bedeutet hier, daß die beobachtete Korrelationsmatrix nur geringfügig von der unter den Modellannahmen geschätzten Korrelationsmatrix abweicht (vgl. Tabelle 3): Das Modell beinhaltet, daß sechs Variablenpaare bedingt unabhängig sind, gegeben alle übrigen Variablen, und zwar die Paare (1,4), (1,5), (1,6) und (2,4), (2,5), (2,6). Daraus, daß diese sechs Bedingungen gleichzeitig erfüllt sind, folgt daß jedes dieser Variablenpaare auch bedingt unabhängig ist, gegeben nur die Variable 3. Das heißt, gegeben die Variable 3, sind zwei Variablengruppen unabhängig, die Gruppe mit den zwei Variablen 1 und 2 einerseits und die Gruppe mit den drei Variablen 4,5 und 6 andererseits. Das bedeutet formal, daß die Korrelationen zwischen den sechs genannten Variablenpaaren verschwinden, wenn man Variable 3 konstant hält. Das bedeutet inhaltlich, daß für Wählergruppen mit derselben Parteiidentifikation (3) aus der Kenntnis des sozioökonomischen Status des Wählers (4) oder seines Vaters (5) oder der Parteiidentifikation des Vaters (6) kein verbesserter Rückschluß auf die Wahlentscheidung (1) oder die Zufriedenheit mit der Regierungspolitik (2) möglich ist. Dies scheint eine relativ kurze, inhaltlich sinnvolle Beschreibung der fünfzehn Korrelationskoeffizienten zu sein.

C. Interpretation als System von Abhängigkeitsbeziehungen

Anhand der volkswirtschaftlichen Daten zeigen wir, wie sich ein multiplikatives Modell als ein System von Abhängigkeitsbeziehungen interpretieren läßt. Die Daten sind Längsschnittbeobachtungen für die Bundesrepublik Deutschland. Für fünf Variable (siehe Tab. 4), die Beschäftigten, die Bruttoanlageinvestitionen, das Bruttoeinkommen aus Unternehmertätigkeit und Vermögen, den privaten Verbrauch und die Exporte wur-

Tabelle 4: Volkswirtschaftliche Variablen (von der Lippe 1977)

Variablennummer	Variablennamen
1	Beschäftigtenzahl
2	Bruttoanlageinvestitionen
3	Bruttoeinkommen aus Unternehmertätigkeit und Vermögen (ohne den Sektor Staat)
4	Privater Verbrauch
5	Exporte

Beobachtungszahl: 24.
Gut passendes multiplikatives Modell: 145/245/345 (LQ – χ^2 = 0,44 bei 3 Freiheitsgraden).

den Wachstumsraten für 24 Jahre ermittelt mit der Absicht, den Einfluß eines gemeinsamen Trends weitgehend auszuschalten (von der Lippe 1977). Der Autor stellte die Daten zusammen, um beurteilen zu können, ob in den Nachkriegsjahren die Beschäftigungslage mehr von der Investitionshöhe abhängig war oder mehr von der Nachfrage, wobei die Nachfrage den privaten Verbrauch und die Exporte umfaßt.

Ein multiplikatives Modell, das die beobachtete Korrelationsmatrix gut beschreibt, ist das Modell 145/245/345. Das beinhaltet, daß für ein vorgegebenes Nachfrageniveau (Variable 4 und 5) Beschäftigungslage (1), Investitionen (2) und Unternehmereinkommen (3) unabhängig voneinander sind. Mit anderen Worten: bei vorgegebenem Nachfrageniveau war aus Kenntnis der Höhe von Investitionen und Unternehmereinkommen kein verbesserter Rückschluß auf die Beschäftigungslage möglich. Wie gering die Abweichungen sind, die sich zwischen der beobachteten Korrelationsmatrix und der unter diesen Modellannahmen geschätzten Matrix ergeben, wird in Tabelle 5 gezeigt und spiegelt sich auch deutlich in der Likelihood-Quotienten-Prüfgröße wieder.

Tabelle 5: Korrelationen der volkswirtschaftlichen Variablen

Variablennummer	1	2	3	4	5
1	1,00	0,59 (0,55)	0,47 (0,52)	0,67	0,44
2	0,12	1,00	0,49 (0,45)	0,80	0,07
3	0,12	0,09	1,00	0,55	0,39
4	0,49	0,57	0,36	1,00	0,04
5	0,55	0,04	0,43	0,40	1,00

Obere Hälfte: r_{ij}, einfache beobachtete Korrelationen.
In Klammern: r^*_{ij}, einfache, durch Modell 145/245/345 implizierte Korrelationen.
Untere Hälfte: $r_{ij.klr}$, partielle Korrelationen.

Die gute Anpassung dieses Modells bedeutet, daß ein bestimmtes System von linearen Abhängigkeitsbeziehungen ebensogut mit den Beobachtungen zu vereinbaren ist. Man geht von dem folgenden vollständigen, rekursiven Gleichungssystem aus,

$$
\begin{aligned}
X_1 &= a_{12}X_2 + a_{13}X_3 + a_{14}X_4 + a_{15}X_5 + U_1 \\
X_2 &= a_{23}X_3 + a_{24}X_4 + a_{25}X_5 + U_2 \qquad (1) \\
X_3 &= a_{34}X_4 + a_{35}X_5 + U_3 \; .
\end{aligned}
$$

für das angenommen wird, daß die Fehler (U_i) normalverteilt und unkorreliert sind (vergl. Goldberger 1964).

Modell 145/245/345 impliziert nun, daß genau drei Regressionskoeffizienten gleich Null sind, die Koeffizienten a_{12}, a_{13} und a_{23} (Wermuth 1979). Man erhält damit ein unvollständiges rekursives System, in dem jede der Variablen 1, 2 und 3 unmittelbar von den Variablen 4 und 5 abhängig ist. Die Beziehungen zwischen den Variablenpaaren (1,2), (1, 3) und (2,3) werden dagegen nur noch als mittelbare Abhängigkeiten angesehen, als erklärbar durch die Beziehungen, die zwischen den übrigen Variablen bestehen. Inhaltlich bedeutet die gute Anpassung dieses unvollständigen rekursiven Systems insbesondere, daß die Beschäftigungslage in den Jahren 1950–74 als nur mittelbar von Investitionen und Unternehmereinkommen abhängig angesehen werden kann. Die hohen partiellen Korrelationskoeffizienten zwischen den Paaren (1,4) und (1,5) zeigen andererseits (siehe Tabelle 5), daß die Annahme einer nur mittelbaren Abhängigkeit der Beschäftigtenzahl (Variable 1) von der Nachfrageseite her (Variablen 4 und 5) nicht mit den Beobachtungen zu vereinbaren ist. In diesem Beispiel scheint uns damit die gute Beschreibung der beobachteten Korrelationen durch ein multiplikatives Modell zu einer inhaltlichen Interpretation zu führen, die diskussionsanregend ist.

Tabelle 6: Beobachtete Häufigkeiten der medizinischen Variablen

Untergewicht Mutter	Wohnort Großstadt	Zigaretten pro Tag	Alter der Mutter	Schwangerschaftsdauer in Tagen					
				197–260 p. Mort.		261–270 p. Mort.		271 u. mehr p. Mort	
				ja	nein	ja	nein	ja	nein
ja	nein	keine	< 25	1	10	1	11	1	132
			25–29	3	19	1	31	1	193
			⩾ 30	1	19	0	29	0	138
		1–5	< 25	0	0	0	6	0	55
			25–29	1	2	0	10	0	66
			⩾ 30	0	3	0	7	0	27
		⩾ 6	< 25	1	2	0	4	1	12
			25–29	1	2	0	1	1	26
			⩾ 30	1	2	0	2	0	14
	ja	keine	< 25	1	9	0	15	1	112
			25–29	2	23	0	33	0	166
			⩾ 30	3	5	2	21	2	96
		1–5	< 25	1	5	0	10	0	36
			25–29	0	4	0	2	0	47

			⩾ 30	0	2	0	5	0	21
			< 25	0	5	0	4	0	20
		⩾ 6	25–29	3	3	0	5	0	26
			⩾ 30	1	1	0	1	0	14
			< 25	10	38	1	54	3	457
		keine	25–29	10	49	2	87	2	568
			⩾ 30	15	50	0	78	4	440
			< 25	2	15	0	21	0	189
	nein	1–5	25–29	5	15	0	21	1	167
			⩾ 30	0	10	0	15	1	76
			< 25	0	8	0	7	0	55
		⩾ 6	25–29	2	5	0	6	1	55
			⩾ 30	1	2	0	4	0	24
nein			< 25	6	24	3	35	1	282
		keine	25–29	3	39	2	56	2	422
			⩾ 30	10	25	0	46	4	262
			< 25	2	11	0	18	0	110
	ja	1–5	25–29	0	19	0	15	1	108
			⩾ 30	1	8	0	6	0	55
			< 25	1	7	0	9	0	84
		⩾ 6	25–29	0	5	1	17	1	67
			⩾ 30	1	3	1	5	0	46

III. Definitionen und Eigenschaften multiplikativer Modelle

Definitionen und Eigenschaften von multiplikativen Modellen lassen sich angeben, ohne daß man sich auf eine bestimmte Wahrscheinlichkeitsverteilung bezieht. Aber Schätzwerte bei den jeweiligen Modellannahmen und Prüfgrößen für die Güte der Anpassung eines multiplikativen Modells an Beobachtungswerten sind bisher lediglich für zwei spezielle Verteilungen abgeleitet worden, für die Multinomialverteilung und für die multivariate Normalverteilung. Für beide dieser Verteilungen sind die Berechnungsformeln für Schätzwerte und Prüfgrößen so einfach, daß man – im Prinzip zumindest – nur auf einen Taschenrechner angewiesen ist. Bevor wir auf diese Verteilungen eingehen (Abschnitt B), geben wir zunächst allgemeine Definitionen und formulieren die damit ableitbaren Eigenschaften multiplikativer Modelle als Thesen.

A. Definitionen und Thesen

Wir beschreiben, inwiefern es duale Charakterisierungen für ein multiplikatives Modell gibt, einerseits anhand einer Liste von Variablenpaaren mit partiellen Nullassoziationen

und andererseits anhand einer Liste von zusammengehörigen Variablengruppen. Mit den Thesen wird gezeigt, wie man jeweils von einer dieser Charakterisierungen zur anderen gelangt und wie man für eine beliebige Liste mit partiellen Nullassoziationen oder Indexgruppen entscheiden kann, ob sie ein multiplikatives Modell kennzeichnen.

Gegeben sei ein p-dimensionaler Zufallsvektor, der eine Verteilungsfunktion besitze. Die Variablen dieses Vektors werden mit Indices, für $1 \leqslant j \leqslant p$, bezeichnet.

Definition 1

Ein Variablenpaar, das bedingt unabhängig ist gegeben alle übrigen $(p - 2)$ Variablen, wird ein Paar mit partieller Nullassoziation genannt.

Eine Indexliste $I \subseteq \tilde{I} = \{(i, j) \mid 1 \leqslant i < j \leqslant p\}$ bezeichne alle jene Variablenpaare, die in einer gegebenen p-dimensionalen Verteilung die Bedingung der partiellen Nullassoziationen erfüllen.

Definition 2

Eine Indexliste I heißt reduzibel falls für jedes in I enthaltene Paar (i, j) gilt, daß für alle $h = 1, \ldots, i - 1$ die Paare (h, i) oder (h, j) oder beide in I enthaltenen sind.

These 1

Eine Indexliste I kennzeichnet dann und nur dann ein multiplikatives Modell, wenn die Variablen so numeriert werden können, daß I reduzibel ist (Wermuth 1979).

Es sei C eine Teilmenge von $\{1, \ldots, p\}$, $\underset{\sim}{X}_C$ bezeichne einen Teilvektor von $(X_1, \ldots, X_p)$, der alle X_k mit $k \in C$ enthält, $f_C(\underset{\sim}{X}_C)$ sei die Verteilungsfunktion von $\underset{\sim}{X}_C$ und $f_\emptyset = 1$. Weiterhin seien für ein reduzibles I Indexmengen wie folgt für $i = 1, \ldots, p$ definiert:

$$A_i = \{j \mid j > i \text{ und } (i, j) \notin I\}$$

$$B_i = \{j \mid j > i \text{ und } (i, j) \in I\}$$

$$C_i = \{i\} \cup A_i$$

These 2

Die zu einem reduziblen I gehörige Dichte oder Wahrscheinlichkeitsfunktion ist

$$f(X_1, \ldots, X_p) = \left(\prod_{i=1}^{p-1} f_{C_i}(X_{C_i}) / f_{A_i}(X_{A_i}) \right) f_p(X_p).$$

Sie vereinfacht sich in der Regel durch Kürzen.

These 3

Eine zu einem reduziblen I gehörige Modellinterpretation ist für alle i:

X_i ist unabhängig von $\underset{\sim}{X}_{B_i}$ für $A_i = \emptyset$

X_i ist unabhängig von $\underset{\sim}{X}_{B_i}$ gegeben $\underset{\sim}{X}_{A_i}$ für $A_i \neq \emptyset$.

Es sei $\{N_t\} = \{N_1, \ldots, N_T\}$ eine Sammlung von Teilmengen von $\{1, \ldots, p\}$ derart, daß kein Element von $\{N_t\}$ in einem anderen Element enthalten ist und $\{N_t\}$ weise jene Variablengruppen aus, die die gemeinsame Verteilung generieren.

These 4

Eine Indexmenge $\{N_t\}$ kennzeichnet dann und nur dann ein multiplikatives Modell, wenn die Variablen so numeriert werden können, daß jedes N_t aus Elementen besteht, die in keinem N_i mit $i > t$ vorkommt und solchen Elementen, die Teilmenge eines N_j mit $j > t$ sind (Sundberg 1975).

Definition 3

Unter der Modellbezeichnung eines multiplikativen Modells versteht man die durch Schrägstriche getrennten Indexgruppen N_t.

These 5

Aus der Modellbezeichnung eines multiplikativen Modells erhält man die zugehörige Liste der partiellen Nullassoziationen als all jene Paare, die in keiner Indexgruppe gemeinsam vorkommen ($(i, j) \in I \Leftrightarrow \{i, j\} \not\subseteq N_t$).

These 6

Aus einem reduziblen I erhält man die zugehörige Modellbezeichnung nach Streichen aller echten Teilmengen aus der Menge der C_i (d.h. $\{N_t\} \subseteq \{C_1, \ldots, C_p\}$).

These 7

Eine Variablennumerierung mit reduziblem I erfüllt Sundbergs Kriterium für ein multiplikatives Modell und umgekehrt (Wermuth 1979).

Bei einer graphentheoretischen Betrachtung der multiplikativen Modelle (Darroch. Lauritzen, Speed 1979) zeigt sich eine partielle Nullassoziation in einem Graphen als ein Paar nicht unmittelbar verbundener Eckpunkte. Eine sogenannte Clique entspricht einer Indexgruppe N_t in der Modellbezeichnung.

B. Spezielle verteilungstheoretische Annahmen

Nur für zwei Verteilungen können wir bisher Maximum-Likelihood Schätzwerte und Likelihood-Quotienten-Prüfgrößen für die Güte der Anpassung eines multiplikativen Modells an die Beobachtungswerte angeben. Im Falle der multivariaten Normalverteilung wurden diese abgeleitet (Wermuth 1976a, 1979) aus der Theorie der Kovarianzselektion (Dempster 1972) und im Fall der Multinomialverteilung ergeben sie sich (Goodman 1970, Bishop 1971) aus der Theorie der logarithmisch-linearen Modelle

(Birch 1963, Bishop, Fienberg und Holland 1975). Für vier Variable wurden alle möglichen multiplikativen Modelle, ihre Interpretation sowie Prüfgrößen und Schätzwerte unter den beiden genannten verteilungstheoretischen Annahmen von Wermuth (1976a) beschrieben.

1. Partielle Nullassoziation

Eine partielle Nullassoziation des Paares (i, j) bedeutet bei einer multivariaten Normalverteilung, daß die Konzentration (Dempster 1969) dieses Paares gleich Null ist, beziehungsweise daß der partielle Korrelationskoeffizient unter Ausschaltung aller übrigen (p − 2) Variablen für dieses Paar gleich Null ist. Nehmen wir als Beispiel p = 5 Variable. Es bezeichne σ_{ij} und σ^{ij} das Element (i, j) in der Kovarianzmatrix Σ bzw. in der Inversen Σ^{-1}. Dann sind σ_{ij} und σ^{ij} Kovarianz und Konzentration des Paares (i, j), und der partielle Korrelationskoeffizient $\rho_{ij.klr}$ ist gleich dem standardisierten Element (i, j) in der Inversen Σ^{-1}:

$$(2) \qquad \rho_{ij.klr} = \sigma^{ij}(\sigma^{ii}\sigma^{jj})^{-1/2}$$

Bei einer Multinomialverteilung zeigt sich die partielle Nullassoziation des Paares (i, j) auf andere Weise: alle Wechselwirkungen, die dieses Paar betreffen, sind in der logarithmisch-linearen Parametrisierung gleich Null gesetzt, d.h. die zweifache Wechselwirkung und alle Wechselwirkungen höherer Ordnung mit dem Paar (i, j) verschwinden. Das Fehlen dieser Wechselwirkungen wirkt sich auf die erwarteten Häufigkeiten aus: Bei p = 5 Variablen bezeichne m_{ijklr} die erwartete Häufigkeit in jeder der Zellen einer IxJxKxLxR Kontingenztafel, und es sei z.B. $m_{i.klr} = \sum_{j=1}^{J} m_{ijklr}$. Dann bedeutet partielle Nullassoziation für Paar (1, 2), daß sich die erwartete Häufigkeit in jeder der KxLxR Teiltafeln multiplikativ aus Randhäufigkeiten wie folgt zusammensetzt:

$$(3) \qquad m_{ijklr} = m_{i.klr}\, m_{.jklr}/m_{..klr}$$

2. Maximum Likelihood-Schätzwerte

Für beide genannten Verteilungen gibt es programmierte iterative Verfahren, mit denen Maximum-Likelihood (ML)-Schätzer für beliebige log-lineare Modelle bzw. Kovarianzselektionsmodelle berechnet werden können. In der Untergruppe der multiplikativen Modelle lassen sich die Schätzer jedoch jeweils in geschlossener Form darstellen.

Bei der Multinomialverteilung für eine $J_1xJ_2x \ldots xJ_p$ Tafel läßt sich der ML-Schätzer für die erwarteten Zellenhäufigkeiten mit Randtafelhäufigkeiten analog zur Faktorisierungseigenschaft der Verteilung (These 2) schreiben. Bezeichnen wir bei p = 5 Varia-

blen mit n_{ijklr} und $\hat{m}_{ijklr}$ beobachtete und geschätzte Zellenhäufigkeiten, so ist z.B. für Modell 145/245/345

$$\hat{m}_{ijklr} = n_{i...lr}\, n_{.j.lr}\, n_{..klr}/n_{...lr}\, n_{...lr}\,.$$

Die Anzahl der Freiheitsgrade ist gleich der Anzahl der zu Null gesetzten Wechselwirkungsparameter. Es seien X_i, X_{A_i}, X_{B_i} wie für These 3 definiert, dann ergibt sich diese Anzahl für jedes X_i mit $B_i \neq 0$ als $(J_i - 1)\left(\prod_{k \in B_i} J_k - 1\right)$ für $A_i = \emptyset$ oder als

$(J_i - 1)\left(\prod_{k \in B_i} J_k - 1\right) \prod_{l \in A_i} J_l$ für $A_i \neq \emptyset$.

Auf diese Weise errechnen sich z.B. die Freiheitsgrade für Modell 145/245/345 als $(I - 1)(JK - 1)LR + (J - 1)(K - 1)LR$.

Im Falle der multivariaten Normalverteilung läßt sich für ein gegebenes multiplikatives Modell der ML-Schätzer $(\hat{P})$ der Korrelationsmatrix aus Regressionsschätzungen nach der Methode der kleinsten Quadrate bestimmen.

Gegeben sei die reduzible Liste der partiellen Nullassoziationen, I, eines multiplikativen Modells und k sei die erste der Variablen, für die kein Paar (k, l), mit $2 \leqslant k < l \leqslant p$, in I enthalten ist, dann schreibt man zunächst ein vollständiges, rekursives Gleichungssystem (siehe (1)) mit den Variablen $i = 1, \ldots, k - 1$ als abhängigen Variablen und setzt all jene Regressionskoeffizienten zu Null, die in I enthaltene Variablenpaare betreffen. Es bezeichne weiterhin r_{ij} den beobachteten Korrelationskoeffizienten des Paares (i, j) und α_{ij} den standardisierten Regressionskoeffizienten der Variablen j in der i'ten Regressionsgleichung, dann erhält man zunächst die Schätzer (für festes i) nach der Methode der kleinsten Quadrate, $\hat{\alpha}_{ij}$, als Lösung der sogenannten Normalgleichungen:

$$(4) \qquad r_{ij} = \Sigma_l \hat{\alpha}_{il} r_{lj}$$

für $i < j \leqslant l$ und (i, j) (i, l), nicht in I.

Bezeichnen wir mit $\hat{\rho}_{ij}$ Element (i, j) des ML-Schätzers $\hat{P}$, so erhält man nunmehr $\hat{P}$ aufbauend in der Reihenfolge $i = p - 1, p - 2, \ldots, 1$ aus:

$$(5) \qquad \hat{\rho}_{ij} = \begin{cases} r_{ij} & \text{für (i, j) nicht in I} \\ \Sigma_l \hat{\alpha}_{il} \hat{\rho}_{lj} & \text{für (i, j) in I und} \\ & \text{(i, l) nicht in I} \end{cases}$$

Gleichung (5) zeigt, daß die durch die Modellannahmen implizierte Korrelation $\hat{\rho}_{ij}$ nur dann von der beobachteten Korrelation r_{ij} abweichen kann, wenn Paar (i, j) eine partielle Nullassoziation hat, also in I enthalten ist. Es sei nochmals betont, daß die Voraussetzung für die Gültigkeit der Gleichung (5) eine reduzible Liste I, partieller Null-

assoziationen ist: nur in diesem Fall kann das Nullmuster in den Regressionskoeffizienten eines unvollständigen rekursiven Gleichungssystems dem Nullmuster in den partiellen Assoziationen der Kovarianzmatrix entsprechen.

3. Likelihood-Quotienten Prüfgrößen

Für ein gegebenes multiplikatives Modell mit r partiellen Nullassoziationen kann man die Güte der Anpassung dieses Modells an Beobachtungswerte anhand einer Likelihood-Quotienten-Prüfgröße (LQ-χ^2) beurteilen, die für große Stichproben annähernd chi-Quadrat-verteilt ist.

Bezeichnet bei multivariat-normalverteilten Größen $\underset{\sim}{R}$ die beobachtete Korrelationsmatrix und $\underset{\sim}{\hat{P}}$ den ML-Schätzer (5) so ist

$$(6) \qquad LQ - \chi^2 = 2n(\ln \det \underset{\sim}{\hat{P}} - \ln \det \underset{\sim}{R})$$

bei r Freiheitsgraden.

Bezeichnen für multinomial-verteilte Größen in einer $J_1 x J_2 x \ldots J_p$ Kontingenztafel $\hat{m}_{j_1 \ldots j_p}$ die ML-Schätzer, so ist

$$LQ - \chi^2 = -2(n_{j_1 \ldots j_p} \ln \hat{m}_{j_1 \ldots j_p}) - (n_{j_1 \ldots j_p} \ln n_{j_1 \ldots j_p})$$

mit Freiheitsgraden, die gleich der Anzahl der für $\hat{m}$ zu Null gesetzten Wechselwirkungsparameter ist (vgl. III B 2).

Für alle multiplikativen Modelle ist die Berechnung dieser Prüfgrößen möglich, ohne daß zuvor die ML-Schätzer explizit bestimmt werden müssen. Wesentlich ist nur, die zum Modell gehörige Faktorisierungsregel zu kennen. Sie legt fest, wie det $\hat{P}$, bzw. $\hat{m}$ in beobachtete Größen zerlegt werden kann.

Nehmen wir als Beispiel wieder Modell 145/245/345. Es bezeichne R_{145} die 3x3 Teilmatrix von R mit den Variablen 1, 4 und 5, weiterhin seien R_{245}, R_{345}, R_{45} entsprechend definiert. Dann errechnet sich det $\hat{P}$ wie folgt:

$$\det \hat{P} = \det R_{145} \det R_{245} \det R_{345} / \det R_{45} \det R_{45} .$$

Ähnlich ergibt sich für den Ausdruck $\Sigma n \ln \hat{m}$:

$$\Sigma n_{ijklr} \ln \hat{m}_{ijklr} = (\Sigma n_{i\ldots lr} \ln n_{i..lr})$$
$$(\Sigma n_{.j.lr} \ln n_{.j.lr}) + (\Sigma n_{..klr} \ln n_{..klr}) - 2 (\Sigma n_{...lr} \ln n_{..lr}) .$$

In Situationen, in denen nur die Anpassungsprüfung, nicht aber die ML-Schätzer an sich interessieren, führt diese Eigenschaft multiplikativer Modelle zu einer beachtlichen Rechenzeitersparnis.

IV. Diskussion

Wir versuchten, die Anwendung multiplikativer Modelle bei der Datenanalyse anhand von unterschiedlichen Datenbeispielen zu motivieren. Wir stellten solche Eigenschaften multiplikativer Modelle zusammen, die uns als wesentlich für das Verständnis dieser Modelle erscheinen. Schließlich zeigten wir, wie einfach für zwei spezielle Verteilungen die Berechnung von Schätzwerten und Prüfgrößen ist: im Fall der Multinomialverteilung ist es lediglich erforderlich, Randtabellenhäufigkeiten zu multiplizieren oder zu logarithmieren; im Fall der multivariaten Normalverteilung werden für die Prüfgrößen Determinanten von Teilen der beobachteten Korrelationsmatrix berechnet, und die Schätzwerte werden im wesentlichen mit Hilfe der Methode der kleinsten Quadrate bestimmt.

Abgesehen von der Einfachheit in den Berechnungen sind multiplikative Modelle vor allem aus folgenden Gründen attraktiv. Einerseits ist jedes Modell eindeutig definiert, so daß die Anpassung an Daten als statistische Hypothese geprüft werden kann, andererseits umfaßt die Gruppe der multiplikativen Modelle so viele unterschiedliche Strukturen, daß sich Modellsuchverfahren sinnvoll formulieren lassen. Das bedeutet, daß die Modelle auch in einem der Hypothesenprüfung vorgelagerten Forschungsstadium eingesetzt werden können. In diesem Fall führen sie zu Hypothesen über Zusammenhangsstrukturen, die an neuen Daten überprüft werden können oder aber sie führen zu zusammenfassenden, vereinfachenden Beschreibungen komplexer Zusammenhänge, deren inhaltliche Bedeutung zur Diskussion gestellt wird.

Anhand der Prüfgrößen und der Abweichungen zwischen beobachteten und den unter Modellannahmen implizierten Werten ist jederzeit leicht nachvollziehbar, ob die strukturvereinfachenden Annahmen mit den Beobachtungen gut zu vereinbaren sind oder nicht. Nicht überprüft werden kann damit, ob die verteilungstheoretischen Annahmen zutreffen.

Da multiplikative Modelle ohne Bezug auf spezielle Verteilungen definiert sind, ist es wünschenswert für andere mehrdimensionale Verteilungen Einzelheiten auszuarbeiten, die Schätzwerte und Prüfgrößen betreffen. Insbesondere jene Verteilungen der mehrdimensionalen Exponentialfamilie, die gleichzeitig diskrete und stetige Variable einschließen (Dempster 1971) scheinen uns hierbei für Anwendungen wichtig zu sein.

ZUSAMMENFASSUNG

Der Nutzen multiplikativer Modelle für die Datenanalyse wird anhand je eines medizinischen, soziologischen und volkswirtschaftlichen Datenbeispiels beschrieben. Wichtige

Eigenschaften der multiplikativen Modelle werden als Thesen zusammengestellt. Ferner werden die Berechnungsformeln für Maximum-Likelihood-Schätzwerte und Likelihood-Quotienten Prüfgrößen für zwei spezielle Verteilungen angegeben: für die Multinomialverteilung und für die multivariate Normalverteilung.

*

SUMMARY

We describe how to use multiplicative models in the analysis of data with a medical set of data, a sociological set of data, and an economic set of data. We present important properties of multiplicative models as propositions. Furthermore, we give computing formulas for maximum-likelihood estimates and likelihood-ratio statistics in the case of two special distributions: for the multinomial distribution and for the multivariate normal distribution.

*

Literaturverzeichnis

Andersen A. H. (1974): „Multidimensional Contingency Tables", Scandinavian Journal of Statistics, 1, 115–127.

Birch, M. W. (1963): „Maximum-likelihood in Three-Way Contingency Tables", Journal of the Royal Statistical Society, Series B, 25, 220–233.

Bishop. Y. M. M. (1971): „Effects of Collapsing Multidimensional Contingency Tables", Biometrics, 25, 383–400.

ders., Fienberg, S. E., Holland, P. W. (1975): Discrete Multivariate Analysis-Theory and Practice, M. I. T. Press Cambridge, Mass.

Darroch, J. N., Lauritzen, S., Speed, T. P. (1979): „Markov Fields and Linear Interaction Models for Contingency Tables", (erscheint in: The Annals of Statistics).

Dempster, A. P. (1971): „An Overview of Multivariate Data Analysis", Multivariate Analysis, 1, 316–347.

ders. (1972): „Covariance selection", Biometrics, 28, 157–175.

DFG-Forschungsbericht (1978): Schwangerschaftsverlauf und Kindesentwicklung, Harald Boldt Verlag, Boppard.

Goldberg, A. S. (1971): „Discerning a Causal Pattern among Data on Voting Behavior" in Blalock, H. M. Jr. (ed): Causal Models in the Social Sciences, Aldine-Atherton, Chicago, 33–48.

Goldberger, A. S. (1964): Economic Theory, John Wiley & Sons, New York.

Goodman, L. A. (1970): „The Multivariate Analysis of Qualitative Data: Interaction among Multiple Classifications", Journal of the American Statistical Association, 65, 226–256.

Haberman, S. J. (1974): The Analysis of Frequency Data, The University of Chicago Press, Chicago.

Kellerer, H. G. (1964): Verteilungsfunktionen mit gegebenen Marginalverteilungen", Zeitschrift für Wahrscheinlichkeitstheorie, 3, 247–270.

von der Lippe, P. (1977): „Beschäftigungswirkung durch Umverteilung?", WSI-Mitteilungen, 8, 505–512.

Speed, T. P. (1978): „Graphical Methods in the Analysis of Data", (unpublished lecture notes).

Sundberg, R. (1975): „Some Results about Decomposable (or Markov-type) Models for Multidimensional Contingency Tables: Distribution of Marginals and Partitioning of Tests", Scandinavian Journal of Statistics, 2, 71–79.

Wermuth, N. (1976a): „Analogies between Multiplicative Models in Contingency Tables and Covariance Selection", Biometrics, 32, 95–108.

ders. (1976b): „Model Search Among Multiplicative Models", Biometrics, 32, 253–263.

ders. (1976c), „Explorative Analyses of Multidimensional Contingency Tables", in: Proceedings of the 9'th International Biometric Conference, Boston, August 22–27, 1976, Vol. I, p. 279–295, The Biometric Society, Raleigh.

ders. (1978) Zusammenhangsanalysen Medizinischer Daten, Lecture Notes in Medizinischer Informatik und Statistik, Band 5, Springer Verlag, Berlin.

ders. (1979) „Linear Recursive Equations, Covariance Selection, and Path Analysis, (erscheint im: Journal of the American Statistical Association).

Wright, S. (1934) „The Method of Path Coefficients", The Annals of Mathematical Statistics, 5, 161–215.