Gerhard Arminger

Methode, Statistik und Modell in den Sozialwissenschaften

> Was sich überhaupt sagen läßt,
> läßt sich klar sagen; und wovon
> man nicht reden kann, darüber
> muß man schweigen.
>
> (Wittgenstein: Vorrede zum
> Tractatus logico - philosophicus)

Abstract: The relationship between methods, statistics and models in the social sciences is discussed. New models generalizing commonly used linear models to deal with qualitative and ordinal data are introduced; their basic similarity to linear models is pointed out. Rate models and stochastic linear differential equations to model social processes in continuous time are mentioned. The implications of weak substantial theory and the correct use of statistical significance tests for any kind of model are demonstrated.

1. Problemstellung und Vorrede

In diesem Aufsatz werden folgende Problembereiche aus der durchaus subjektiven Sicht des Autors diskutiert:

- Der Stand der den Sozialwissenschaften zur Verfügung stehenden und verwendeten mathematischen und stochastisch formulierten Modelle. Es werden nicht einzelne Modelle, sondern Modellklassen angesprochen, die häufig mißverständlich als statistische Methoden oder Instrumente bezeichnet werden. Der Schwerpunkt liegt auf Modellen der Struktur und der Dynamik und nicht auf Meß- und Klassifikationsmodellen.
- Das Verhältnis zwischen Methode, Statistik und Modell.
- Die Verwendung von Daten und Modellen zum Erzeugen oder Testen von Hypothesen (exploratorische versus konfirmatorische Statistik).

Analyse & Kritik 5 (1983), S. 3–36

Die genannten Probleme werden an Hand von verwendeten Verfahren angesprochen. Neuere Entwicklungen werden in nicht technischer Weise eingeführt, bei Problemen des Schätzens und Testens sowie bei Algorithmen wird auf die einschlägige Literatur verwiesen.

Die durch die statistischen Probleme im engeren Sinne häufig verwischten und manchmal nicht unmittelbar transparenten fundamentalen Ähnlichkeiten zwischen Modellen werden hervorgehoben. Darauf aufbauend werden ihre Implikationen aufgezeigt. Es wird durchwegs versucht, möglichst einfach und durchsichtig zu formulieren, damit die Aussagen leicht kritisiert werden können. Das Verhältnis von Methode und Modell sowie der Mißbrauch statistischer Testtheorie wird von Kriz (1981) ausführlich an Beispielen aus der Literatur behandelt, so daß ich mich - der Kürze eines Aufsatzes angemessen - auf einer vom einzelnen Beispiel losgelösten Ebene bewegen kann.

Die subjektive Sicht und Kenntnis des Autors, die den einzelnen Überlegungen zu Grunde liegt, läßt sich am besten mit einigen Stichworten beschreiben: ich bin der Auffassung, daß sozialwissenschaftliche Theorie ab irgendeinem Punkt, der der Einigung zwischen Wissenschaftlern bedarf, an der Realität überprüfbar sein muß. Jede empirische Forschung ist - wenn auch manchmal kaum bewußt - von Theorien oder zumindest Vorstellungen und Primärerfahrungen geleitet. Sozialwissenschaftliche Theorie und Empirie sollen in stetem Wechselspiel stehen. Der Formulierung von Begriffen und Hypothesen soll eine genaue Analyse der sozialen Situationen, auf die sich der Forscher bezieht, vorangehen. Jede Art und Weise, Kenntnisse zu sammeln und damit Hypothesen zu erzeugen, sei es Introspektion, Befragung, Gruppeninterview oder Verwendung von Literatur, Archiv- oder statistisches Material, ist zulässig. Bei der Analyse sozialer Situationen wird das Verstehen von Symbolen und Handlungen aus dem Kontext heraus von besonderer Bedeutung sein. Trotzdem sollte versucht werden, Resultate von Beobachtung und Befragung zu kategorisieren und damit der Behandlung als Merkmal mit eindeutiger Zuordnung jedes Untersuchungselements zu einer Merkmalskategorie zugänglich zu ma-

chen. Das Meßniveau der Daten soll nicht künstlich verändert werden, da sonst implizite Annahmen über eine Transformation (z.B. ordinal auf quantitativ) oder Informationsverlust (quantitativ auf qualitativ) die unangenehmen Folgen sind. Schließlich soll man einige Anstrengung auf sich nehmen, zu überprüfen, wie gut beobachtete Phänomene durch Variable und Modelle erklärt werden. Bekanntlich führt dies zu einer bescheideneren Einschätzung der Erklärungskraft eigener Theorien und Hypothesen, die größerer Neugier und Offenheit in der wissenschaftlichen Arbeit nur nützlich sein kann.

2. Strukturmodelle

Da eine Reihe von Implikationen und Problemen bereits bei einfachen Modellen aufgezeigt werden kann, behandeln wir zunächst Modelle, bei denen keine wechselseitige Veränderung der Variablen über die Zeit unterstellt wird. Solche Modelle können durch Erhebung von Querschnittsdaten überprüft werden. Sie bilden den weitaus größten Teil der Modelle in den Sozialwissenschaften. Entsprechend werden am häufigsten Querschnittsdaten erhoben. Da sie nicht auf die Erfassung von Veränderungen und Prozessen abzielen, werden sie als Strukturmodelle bezeichnet.

2.1 Einfache lineare Modelle

Eines der einfachsten Modelle in den Sozialwissenschaften ist die Darstellung von Messungen als Summe von Mittelwert und Fehler:

$$y_i = \mu + e_i , \quad i=1,\ldots,n \qquad (1)$$

$$E(y_i) = \mu \qquad \text{(E ist Erwartungswertoperator)}$$

Die Beobachtungen von y_i sind quantitativ und statistisch unabhängig. Dieses Modell eines gleichen Mittelwerts für alle Beobachtungen ist wenig befriedigend, da wir in aller Regel Wirkungen von anderen Variablen auf y untersuchen wollen,

also Unterschiede zwischen Elementen auf erklärende Faktoren zurückführen wollen. Ein Modell, das auf die Heterogenität der untersuchten Population abstellt, ist z.B.:

$$y_i = (\mu_i+\delta_i) + e_i \qquad (2)$$

$$\mu_i = \beta_i x_{i1} + \ldots + \beta_p x_{ip}$$

$$E(y_i) = \mu_i + \delta_i$$

Der Erwartungswert wird aufgeteilt auf eine systematische Komponente, μ_i, die durch eine Linearkombination von exogenen Variablen x_{ij} und Parametern β_j (Regressionskoeffizienten) beschrieben wird, und eine spezifische Komponente δ_i. Die durch die beobachteten Werte x_{ij} erzeugten μ_i werden als beobachtete Heterogenität, die δ_i als unbeobachtete Heterogenität bezeichnet. In linearen Modellen ist sie nicht von der Fehlerkomponente e_i zu trennen; daher werden die üblichen Regressionsmodelle wie folgt beschrieben:

$$y_i = \mu_i + (\delta_i+e_i) \qquad (3)$$

$$\mu_i = \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$$

$$E(y_i) = \mu_i$$

Diese Schreibweise hat zur Folge, daß die unbeobachtete Heterogenität, die z.B. durch uns unbekannte exogene Faktoren entsteht, nicht hinreichend beachtet wird. Auf die Folgen dieses sogenannten Spezifikationsfehlers werden wir noch zu sprechen kommen. Vor allem aber ist zu beachten, daß hier ein lineares Modell in den Parametern $\underline{\beta}$ spezifiziert wird. Selbst durch die einfache Gleichung (3) wird also bereits ein theoretisches Modell für einen Zusammenhang zwischen y und den erklärenden Variablen $\underline{x}$ beschrieben, das sowohl die Anzahl und die Auswahl der erklärenden Variablen als auch die Art des Zusammenhangs genau festlegt.

Da aus n Beobachtungen y_i nicht mehr als n Parameter des gesamten Modells, also inklusive Verteilung des Fehlers, ge-

schätzt werden können, werden häufig folgende Annahmen getroffen:

$$y_i = \mu_i + e_i \qquad (4)$$

$$y_i \sim N(\mu_i, \sigma^2)$$

$$\mu_i = \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$$

$$E(e_i e_j) = 0, \quad i \neq j$$

Es wird unterstellt, daß die Fehler e_i für alle Elemente gleiche Varianz aufweisen (Homoskedastizität). Auf Grund der Unabhängigkeit sind sie unkorreliert. Für die Konstruktion von Konfidenzintervallen und Tests wird Normalverteilung angenommen. Schätzt man die Werte von β_j mit dem Maximum Likelihood (ML) Verfahren oder der Methode der kleinsten Quadrate, lassen sich folgende Kennwerte berechnen und unter Verwendung der Matrixschreibweise kompakt darstellen:

$\underline{y} = (y_i)$ $\quad i=1,\ldots, n$

$\underline{X} = (x_{ij})$ $\quad i=1,\ldots, n;\ j=1,\ldots, p$

$\underline{\beta} = (\beta_j)$ $\quad j=1,\ldots, p$

$\underline{b} = (b_j)$ $\quad j=1,\ldots, p$ sind die Schätzwerte für $\underline{\beta}$

$$\underline{b} = (\underline{X}'\underline{X})^{-1}\,\underline{X}'\underline{y} \qquad (5)$$

$$\hat{\underline{y}} = \underline{X}\underline{b} \qquad (6)$$

$$s^2 = \frac{1}{(n-p)} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \qquad (7)$$

$$R^2 = 1 - \left(\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \Big/ \sum_{i=1}^{n} (y_i - \bar{y})^2 \right) \qquad (8)$$

$$\underline{V} = s^2 (\underline{X}'\underline{X})^{-1} \qquad (9)$$

s^2 = geschätzte Fehlervarianz

$\underline{V}$ = geschätzte Kovarianzmatrix der Regressionskoeffizienten ($\underline{b}$)

R^2 = multiples Bestimmtheitsmaß

Die Diagonalelemente von $\underline{V}$ sind die geschätzten Varianzen von b_j, mit ihrer Hilfe lassen sich die üblichen t-Tests und Konfidenzintervalle sowie die F-Tests berechnen.

2.2 Die Folgen schlechter Theorie

Untersucht man die in Gleichung (4) unterstellten Annahmen und die Verwendung der Regressionsanalyse in den Sozialwissenschaften näher, kann man sich des Eindrucks nicht erwehren, daß viele Anwender einerseits die unterstellten Annahmen und ihre Implikationen überhaupt nicht kennen und berücksichtigen, andererseits die in der Design Matrix $\underline{X}$ enthaltenen Möglichkeiten, Hypothesen zu formulieren, nicht wahrnehmen.

Die schwerwiegendsten Fehler werden zweifellos bei der Spezifikation gemacht. Zunächst ist die Auswahl der Variablen zu überlegen. Fehlen wichtige Variable, so wird der Spezifikationsfehler δ_i groß; die Folgen sind in der Regel ein geringes R^2, das ein Maß für die Erklärungskraft der exogenen Variablen darstellt,und eine hohe Instabilität der geschätzten Regressionskoeffizienten. In diesem Fall enthalten die Parameter b_j einen starken indirekten Effekt, der aus der Korrelation der Variablen x_j mit dem Spezifikationsfehler entsteht. Besonders bei hoch aggregierten Daten kommt es leicht zu völliger Veränderung von b_j, wenn zusätzliche exogene Variable, die im Spezifikationsfehler enthalten sind, in die Regression eingeführt werden. Dies kann sofort an folgenden Beziehungen abgelesen werden.

Für die abhängigen Variablen y und z seien lineare Modelle spezifiziert. Die Erwartungswerte werden als bedingte Erwartungswerte in Abhängigkeiten von exogenen Variablen $\underline{x}$ geschrieben, so daß gilt:

$$E(y|\underline{x}) = \underline{x}\,\underline{\beta} \tag{10}$$

$$E(z\ \underline{x},y) = \underline{x}\,\underline{\gamma} + \alpha y \tag{11}$$

Lassen wir für z die erklärende Variable y weg - sie wird zum Spezifikationsfehler - erhalten wir den Erwartungswert von z in Abhängigkeit von $\underline{x}$ allein.

$$E(z|\underline{x}) = \int E(z|\underline{x},y)\, f(y|\underline{x})\, d(y|\underline{x}) \tag{12}$$

$$= \underline{x}(\underline{\gamma} + \alpha\underline{\beta})$$

$f(y|\underline{x})$ ist die durch $\underline{x}$ bedingte Dichte von y, $\underline{\gamma}$ ist der direkte Effekt, $\alpha\underline{\beta}$ der indirekte Effekt von $\underline{x}$, der durch den Wegfall von y erzeugt wird.

In gleicher Weise wirkt es sich aus, wenn der Zusammenhang nicht in $\underline{x}$, sondern nur in einer Funktion $\underline{g}(\underline{x})$ linear ist. Aus der Biologie sind zahlreiche Fälle bekannt, für die das zutrifft. Auch dieser Fall läßt sich sofort an Gleichung (13) ablesen.

$E(y|x) = \underline{g}(\underline{x})\underline{\beta}$ sei der "wahre" Zusammenhang

Wird $\underline{g}(\underline{x})$ durch $\underline{x}$ ersetzt, erhalten wir: (13)

$$E(y|\underline{x}) = \underline{x}\,\underline{\beta} + (\underline{g}(\underline{x}) - \underline{x})\underline{\beta}$$

Der Ausdruck $(g(\underline{x})-\underline{x})\underline{\beta}$ tritt dann als Spezifikationsfehler mit den oben geschilderten Konsequenzen auf.

Denkt man daran, daß in der empirischen Sozialforschung nur selten Werte von $R^2 > 0{,}5$ erzielt werden, sind die Resultate, die aus den Regressionskoeffizienten abgeleitet werden, nur mit äußerster Vorsicht zu genießen, wenn man an den oben beschriebenen Spezifikationsfehler denkt. Andererseits ist dies nach meiner Auffassung nur zum geringen Teil den empirischen Sozialforschern vorzuwerfen, sondern zum größten Teil dem Theoriekonzept und der mangelnden Theoriebildung in den Sozialwissenschaften. Zu oft erschöpft sich sogenannte Theorie in nebulosen Begriffsexplikationen und der Verschleierung einfa-

cher Tatbestände durch einen Wust von wissenschaftlich klingenden Wortneuschöpfungen lateinischer oder altgriechischer Herkunft. Selten findet der empirische Forscher Hilfestellung bei der Auswahl von abhängigen und unabhängigen Variablen und bei der Formulierung der Art des Zusammenhangs. So vertraut er sich - nach mehr oder weniger geglückter Operationalisierung seiner Variablen - blind einem Modell an, das wie oben gezeigt, eine Reihe von Annahmen mit ausgeprägten Folgewirkungen hat. Es erscheint in diesem Zusammenhang merkwürdig, daß statistische Modelle wie die Regressionsrechnung nach Auffassung vieler Sozialwissenschaftler "nur" Methoden- oder Instrumentalcharakter haben, obwohl, wie oben gezeigt wurde, dadurch in Wirklichkeit die Vorstellungen über Wirkungszusammenhänge bereits weitgehend festgelegt werden. Selbst die einfache Festlegung, den Mittelwert zu berechnen, impliziert bereits ein Modell. In anderen Wissenschaften, etwa der Biologie, ist es durchaus üblich, an Stelle von Mittelwerten bestimmte Quantile zu setzen, etwa zur Beschreibung des Anfangs oder Endes von Epidemien. Dies dürfte auch für die Analyse von Diffusionen (z.B. Innovationen, Moden) in den Sozialwissenschaften interessant sein. Wir halten also fest, daß jede sogenannte statistische Methode in den Sozialwissenschaften ein Modell eines Wirkungs- oder Meßzusammenhangs darstellt. Wie Kriz (1981) ausführt, ist die Verwendung von Methoden nur das Aufgreifen verschiedener Wege, um zum gleichen Resultat zu gelangen. Wie bereits an trivialen Beispielen erkennbar ist, führen unterschiedliche Modelle zu sehr verschiedenen Ergebnissen. Zusätzlich können wir in der Regel feststellen, daß die einfachen "Methoden" auch besonders restriktiv sind und häufig nur untaugliche Modelle abgeben.

Neben den Spezifikationsfehlern spielen mögliche Fehler durch Annahme der Homoskedastizität der Fehler und der Normalverteilung eine geringere Rolle. Bei quantitativen abhängigen Variablen tritt eine je nach Beobachtung verschiedene Fehlervarianz häufig auf. Sie müssen bei der Schätzung berücksichtigt werden. Ein Beispiel ist die größere Streuung der Sparquoten bei zunehmendem Einkommen von Haushalten. Heteroskedastizität kann leicht durch Residualanalyse entdeckt und durch

gewichtete Regression in einem zweistufigen Verfahren abgefangen werden (Dhrymes 1974). Die Annahme der Normalverteilung ist primär für die Konstruktion von Konfidenzintervallen und Tests notwendig, sie spielt für einfache Modelle der Form in Gleichung (4) eine untergeordnete Rolle. Sie wird allerdings bei komplexeren Modellen, z.B. LISREL als Grundlage des Schätzverfahrens kritisch. Darauf werden wir noch zurückkommen.

Auf der anderen Seite wird übersehen, daß sich viele qualitative Aussagen leicht in Aussagen über spezielle Formen der Designmatrix $\underline{X}$ und des Parametervektors $\underline{\beta}$ übersetzen lassen. Dies gilt in besonderem Ausmaß, wenn die unabhängigen Variablen nominal skaliert sind und daher in $\underline{X}$ als Dummy Variable auftreten, die bekanntlich wie folgt definiert sind:

Eine qualitative Variable A mit Kategorien $(A_0, A_1, \ldots A_m)$ wird aufgelöst in m Dummy Variable x_j, j=1,..., m, für die gilt:

$$x_j = \begin{cases} 1 & \text{wenn die Beobachtung in } A_j \text{ fällt, } j=1,\ldots,m \\ 0 & \text{sonst} \end{cases} \qquad (14)$$

A_o, das beliebig wählbar ist, ist durch die Kombination $x_1 = x_2 = \ldots = x_m = 0$ festgelegt. Diese Festlegung von Dummy Variablen wird als "cornered effect" Reparametrisierung bezeichnet. Eine häufig verwendete Alternative ist die Reparametrisierung durch "centered effects":

$$x_j = \begin{cases} 1 & \text{wenn die Beobachtung in } A_j \text{ fällt, } j=1,\ldots,m \\ -1 & \text{sonst} \end{cases} \qquad (15)$$

Die Einbeziehung von Dummy Variablen in die Regressionsanalyse führt bekanntlich zu Varianz- und Kovarianzanalyse. Die Verwendung von centered effects entspricht der Reparametrisierung $\beta_o + \beta_1 + \ldots + \beta_m = 0$ in der Varianzanalyse für das lineare Modell:

$$y_{ij} = \mu + \beta_j + e_{ij} \qquad (16)$$

$$y_{ij} \sim N(\mu+\beta_j, \sigma^2)$$

Durch Verwendung von Dummy Variablen und geeignete Spezifikation von Spalten der Designmatrix können folgende Typen von Vorstellungen über Wirkungszusammenhänge leicht in das Regressionsmodell übersetzt werden:

- Interaktionen. Sie treten auf, wenn Variable nur in einer bestimmten Kombination einen Effekt auf die abhängige Variable haben. Sie werden durch Multiplikation von Spalten von $\underline{X}$ erzeugt, die die einzelnen Variablen beschreiben.

- Gruppenspezifische Regression. Dies ist ein Spezialfall der Interaktion. In einer Kovarianzanalyse wird angenommen, daß eine quantitative Variable in unterschiedlicher Weise, die von anderen exogenen Variablen abhängt, auf y wirkt.

- Konditionale Effekte. Sie können auch als Interaktionen ohne vorgelagerte Haupteffekte angesehen werden. Sie treten dann auf, wenn eine exogene Variable nur innerhalb einer Ausprägung einer vorhergehenden Variablen möglich ist. Ein Beispiel ist der Einfluß von Beruf und Bildung auf das Einkommen. Bestimmte Berufe, z.B. Arzt, Rechtsanwalt, sind nur innerhalb einer bestimmten Bildungsstufe möglich.

- Restriktionen. Viele Aussagen lassen sich über Restriktionen der Parameter formulieren. Besonders wichtig ist die Restriktion der Gleichheit von Koeffizienten. Sie läßt sich über die Addition von Spalten der Designmatrix erzielen. Restriktionen allgemeiner Art, z.B. daß ein Regressionskoeffizient größer als ein anderer sein muß, wie sie zur Erreichung von Ordinalität dienen, können ebenfalls formuliert werden, die Schätzverfahren müssen dann allerdings modifiziert werden (Judge et al. 1981). Für die Restriktionen werden wir in einem der nächsten Abschnitte ein Beispiel formulieren. Insgesamt läßt sich festhalten: Wenn die sozialwissenschaftliche Theorie bezüglich der Auswahl der

exogenen Variablen, des Variablenzusammenhangs und der Parameter genauere Angaben liefern könnte, könnten die in der empirischen Forschung verwendeten Regressionsmodelle wesentlich besser spezifiziert und eingesetzt werden. Die hier formulierten Aussagen lassen sich auf alle später vorgestellten Modelle übertragen.

2.3 Explorative und konfirmatorische Statistik

In den sozialwissenschaftlichen Zeitschriften hat sich in den letzten Jahren die Kennzeichnung von Ergebnissen und Koeffizienten durch Sternchen eingebürgert. Durch ein, zwei oder gar drei Sternchen wird angegeben, ob ein Koeffizient auf einem bestimmten Testniveau (z.B. $\alpha=0,05$ oder $\alpha=0,01$) signifikant von 0 verschieden ist. Man führt also Tests durch und verwendet die Ergebnisse, um Hypothesen zu überprüfen. Diese Hypothesenprüfung durch Konfidenzintervalle und Tests ist das eigentliche Anliegen der schließenden Statistik, die deshalb auch mit dem Beinamen konfirmatorisch belegt wird. In aller Regel wird dabei übersehen, daß ein gewähltes Testniveau nur für jeweils einen Test an einer Stichprobe gilt. Will man an einer Stichprobe k Tests durchführen, wie dies etwa bei einer Regressionsanalyse, in der jeder Koeffizient getestet wird, der Fall ist, so muß das Testniveau entsprechend kleiner gewählt werden. Eine - allerdings sehr konservative - Faustregel ist das Verfahren von Bonferroni. Soll bei k Tests das Testniveau insgesamt gleich α sein, so ist für den einzelnen Test das Niveau $\alpha^+ = \alpha|k$ zu wählen. Dies führt natürlich bei einer großen Zahl von Tests sofort zu dem Ergebnis, daß kein Koeffizient mehr signifikant ist. Ein weiteres Resultat ist, daß die herkömmlichen Regeln zur Festlegung der Größe von Stichproben in Zweifel gezogen werden müssen. Auch hier wird ja unterstellt, daß nur ein Test oder ein Konfidenzintervall durchgeführt bzw. berechnet wird. Das hier skizzierte Verfahren der Festlegung von Signifikanzniveaus läßt sich etwas abschwächen (S. Holm 1979), trotzdem eignet sich das Sternchenverfahren nicht für die Überprüfung von Hypothesen. Diese Einsicht hindert allerdings den Sozialforscher nicht, dieses Verfahren im

Sinne der explorativen Statistik zur Erzeugung von Hypothesen zu nutzen. Wie in einigen Beiträgen zu dem Sammelband von Victor et al. (1980) ausgeführt wird, läßt sich jedes Verfahren der konfirmatorischen Statistik explorativ anwenden. Von Hypothesenprüfung im strengen Sinn kann allerdings dann nicht mehr gesprochen werden.

2.4 Komplexe lineare Modelle

Der in Gleichung (4) eingeführte Regressionsansatz läßt sich auf multivariate abhängige Variable erweitern. An Stelle der abhängigen Variablen y_i tritt der Vektor von abhängigen Variablen $\underline{y}_i$, der Vektor von Regressionskoeffizienten $\underline{\beta}$ wird durch eine Matrix $\underline{B}$ ersetzt.

$$\underline{y}_i = \underline{\mu}_i + \underline{e}_i \tag{17}$$

$$\underline{y}_i \sim N(\underline{\mu}_i, \underline{\Sigma})$$

$$\underline{\mu}_i = \underline{B}\,\underline{x}_i$$

$$E(\underline{e}_i\underline{e}_j) = \underline{0}$$

Modell und Rechentechnik der multivariaten Regressions-, Varianz- und Kovarianzanalyse werden ausführlich bei Bock (1975) und K. Holm (1979) behandelt. Die Parameter $\underline{B}$ in Gleichung (17) können wie im univariaten Fall durch die Methode der kleinsten Quadrate oder durch Maximum Likelihood geschätzt werden. An den zuvor angeführten Implikationen des Modells im Fall von Spezifikationsfehlern ändert sich nichts.

Wesentlich erweitert wurden die in (17) angeführten linearen Modelle durch die Einführung von Strukturgleichungen aus der Ökonometrie und von latenten Variablen aus der faktorenanalytischen Tradition der Psychometrie. Mit dieser Vermählung von Ökono- und Psychometrie sind vor allem Jöreskog (1982) und das Programmsystem LISREL auf der einen Seite und Wold (1982) mit dem Programmsystem PLS verbunden. Da diese Modelle in diesem Heft an anderer Stelle ausführlich behandelt werden, stellen

wir sie nur in dem uns interessierenden Zusammenhang dar.

LISREL Modelle können durch folgende Gleichungen für einzelne Beobachtungen gekennzeichnet werden:

$$\underline{B}\underline{\eta} = \underline{\Gamma}\underline{\xi} + \underline{\zeta} \qquad \text{Strukturmodell} \qquad (18)$$

$$\underline{\eta} = \underline{B}^{-1}\underline{\Gamma} + \underline{\zeta} \qquad \text{Reduzierte Form} \qquad (19)$$

$$\underline{y} = \underline{\Lambda}_y\underline{\eta} + \underline{\varepsilon} \qquad \text{Meßmodell für } \underline{y} \qquad (20)$$

$$\underline{x} = \underline{\Lambda}_x\underline{\xi} + \underline{\delta} \qquad \text{Meßmodell für } \underline{x} \qquad (21)$$

$$E(\underline{\eta}) = E(\underline{\xi}) = E(\underline{\zeta}) = E(\underline{\varepsilon}) = E(\underline{\delta}) = \underline{0} \qquad (22)$$

$$E(\underline{\xi}\underline{\xi}') = \underline{\Phi},\ E(\underline{\zeta}\underline{\zeta}') = \underline{\psi}$$

$$E(\underline{\varepsilon}\underline{\varepsilon}') = \underline{\theta}_\varepsilon,\ E(\underline{\delta}\underline{\delta}') = \underline{\theta}_\delta$$

Beobachtet werden nur $\underline{y}$ und $\underline{x}$, geschätzt werden sollen die Regressionsparameter der latenten endogenen Variablen, $\underline{B}$, der exogenen Variablen, $\underline{\Gamma}$, und die Kovarianzmatrix $\underline{\Phi}$ der exogenen Variablen sowie der Störungen $\underline{\psi}$. Das Strukturmodell allein entspricht den klassischen ökonometrischen Modellen bei unabhängigen Beobachtungen. Schreibt man das Strukturmodell in reduzierter Form (Gleichung (19)) an, die allein geschätzt werden kann, erhalten wir das multivariate Modell von Gleichung (17).

Die eindeutige Trennung von $\underline{B}$ und $\underline{\Gamma}$, wenn nur der Ausdruck $\underline{B}^{-1}\underline{\Gamma}$ berechnet werden kann, ist das Identifikationsproblem der Ökonometrie. Man beachte, daß das Strukturgleichungsmodell auch für abhängige Beobachtungen verwendet werden kann. Diese Eigenschaft wird später benutzt.

Zusätzlich zum Strukturmodell werden lineare Meßmodelle der klassischen konfirmatorischen Faktorenanalyse (Arminger 1979) eingeführt. Die Restriktionen der Unkorreliertheit von Fehlern können weggelassen werden, was diese Modelle besonders brauchbar zur Analyse von Paneldaten (Jöreskog und Sörbom 1977) macht.

Die Schätzung der Ladungsmatrizen $\underline{\Lambda}_y$, $\underline{\Lambda}_x$ und der Kovarianzmatrizen $\underline{\theta}_\varepsilon$ und $\underline{\theta}_\delta$ wirft in der Regel zusätzliche Identifikationsprobleme auf.

Die Schätzung der Koeffizienten und Kovarianzmatrizen erfolgt nach dem Maximum-Likelihood Prinzip, das auf der Annahme der multivariaten Normalverteilung des beobachteten Vektors ($\underline{x}'$, $\underline{y}'$) beruht und numerisch auf der geschätzten Kovarianzmatrix von ($\underline{x}'$,$\underline{y}'$) aufbaut. Die Schätzung auf Grund der Kovarianz bzw. Korrelationsmatrix bietet den Vorteil, daß ordinale und gemischt ordinale und quantitative Daten auch behandelt werden können. Hier ist allerdings die Annahme der multivariaten Normalverteilung von ausschlaggebender Bedeutung. Sind die ordinalen Variablen durch latente normalverteilte Variable erzeugt, so lassen sich die Korrelationskoeffizienten der latenten Variablen aus den Häufigkeitstabellen der ordinalen Variablen schätzen und werden als polychorische bzw. polyseriale Korrelationskoeffizienten bezeichnet. Sie wurden auch in die neue Version LISREL V nach der Berechnungsmethode von Olsson (1979) und Olsson et al. (1982) eingebaut.

Im Gegensatz zu LISREL basiert PLS (Partial Least Squares) nicht auf der ML Schätzung, sondern stellt sowohl für das Struktur- als auch für das Meßmodell ein Verfahren iterativer Kleinste Quadrate Schätzungen dar. Die statistischen Eigenschaften sind nicht so günstig wie im Fall von LISREL, das Verfahren ist auch nur dann konsistent, wenn sowohl die Zahl der Beobachtungen als auch die Zahl der Indikatoren für die latenten Variablen groß werden (Hui and Wold 1982).

Verlangt bereits die einfache Regressionsanalyse genaue Überlegungen bei der Auswahl der Variablen und der Art des Zusammenhangs, um Fehler bei der Spezifikation zu vermeiden, so erhöht sich die Komplexität bei LISREL und PLS Modellen erheblich. Zusätzlich zum Strukturmodell müssen zwei Meßmodelle spezifiziert werden. Zur Instabilität der Regressionskoeffizienten $\underline{B}$ und $\underline{\Gamma}$ tritt die Instabilität der Ladungsmatrizen $\underline{\Lambda}_y$ und $\underline{\Lambda}_x$, wenn ein neues Modell geschätzt wird. Dies bedeutet, daß das Hinzufügen oder Wegnehmen einer Variablen jeweils

die Hypothese über den Zusammenhang von manifesten und latenten Variablen verändert. Nach meiner Auffassung ist es anzuraten, das Meßmodell nur einmal zu schätzen und für alle weiteren Modelle mit Hilfe von Restriktionen gleich zu halten.

Außerdem ist zu beachten, daß sich der Modelltyp - lineares Modell mit Normalverteilung - nicht geändert hat, so daß alle Implikationen der einfachen linearen Modelle weiter gelten. Trotz dieser Probleme ist die Verwendung dieses Modells den auf den ersten Blick einfacheren Modellen der Bildung von Summen und (un)gewichteten Indizes vorzuziehen, da sie den Sozialforscher zwingt, sein Modell explizit zu machen. Es sollte nicht vergessen werden, daß jede Indexbildung ebenfalls ein - wenn auch besonders einfaches - Meßmodell darstellt, das eben auf Grund seiner restriktiven Annahmen meist falsch ist.

2.5 Verallgemeinerte lineare Modelle: Anwendung auf qualitative Variable

Bis jetzt haben wir nur lineare Modelle mit normalverteilten Fehlern betrachtet. Damit lassen sich qualitative abhängige Variable wie Ausbildung, Beruf, Parteipräferenz oder soziale Mobilität nicht befriedigend behandeln. Für diesen Zweck wurden die sogenannten loglinearen Modelle entwickelt, die eine Verallgemeinerung des herkömmlichen χ^2 Tests auf Anpassung oder Unabhängigkeit nominal skalierter Variablen darstellen.

Loglineare Modelle für Kontingenztabellen sind wie die oben diskutierten linearen Modelle Spezialfälle der von Nelder und Wedderburn (1972) entdeckten Klasse der verallgemeinerten linearen Modelle, die auch als GLM (Generalized Linear Models) Ansatz bezeichnet wird. Wir führen daher zunächst den GLM Ansatz ein und gehen dann im einzelnen auf Probleme der Analyse von qualitativen und ordinalen Daten ein.

Wir gehen wieder von n unabhängigen Beobachtungen y_i aus. Gleichung (4) wird in zweierlei Hinsicht verallgemeinert. An Stelle der Normalverteilung tritt die exponentielle Familie von Verteilungen und zwischen Erwartungswert und Linearkombination

wird eine Verbindungsfunktion (link) geschoben, so daß Gleichung (4) zu Gleichungen (24) - (29) erweitert wird:

$$y_i = \mu_i + e_i; \quad E(y_i) = \mu_i; \quad E(e_i e_j) = 0 \qquad i \neq j \tag{24}$$

$$f(y_i|\theta_i) = \exp\{[y_i\theta_i - b(\theta_i)]/a_i(\phi) + c(y_i,\phi)\} \tag{25}$$

$f(\cdot)$ ist die Dichte der exponentiellen Familie

θ_i heißt kanonischer Parameter

ϕ heißt Dispersionsparameter

$a(.)$, $b(.)$, $c(.)$ sind geeignet gewählte Funktionen

$$\mu_i = \frac{db(\theta_i)}{d\theta_i} = b'(\theta_i) \tag{26}$$

$$V(y_i) = b''(\theta)a_i(\phi) \tag{27}$$

$$\eta_i = g(\mu_i) \qquad \text{ist die link Funktion} \tag{28}$$

$$\eta_i = \sum_{j=1}^{p} x_{ij}\beta_j \qquad \text{ist der lineare Prädiktor} \tag{29}$$

An die Funktionen sind gewisse Regularitätsbedingungen zu stellen, die sichern, daß jedem kanonischen Parameter genau ein Erwartungswert zugeordnet wird. Sind kanonischer Parameter und linearer Prädiktor gleich, ergeben sich spezielle Eigenschaften, die aus dem Blickwinkel der mathematischen Statistik wünschenswert sind.

Die exponentielle Familie umfaßt sowohl diskrete Verteilungen, die zur Analyse von Häufigkeiten (qualitative Variable) verwendet werden, z.B. Poisson-, Binomial- und negative Binomialverteilung als auch stetige Verteilungen zur Analyse spezieller Fehlerverteilungen, z.B. die Normal-, die Gamma-, die Pareto- und die inverse Gaußverteilung. Ausführliche Beispiele sind in Andersen (1980) und Arminger (1983a) angegeben.

Es läßt sich zeigen, daß die Multinomialverteilung, die die Verteilung von Häufigkeiten in Kontingenztabellen beschreibt, als das Produkt unabhängiger Poissonverteilungen geschrieben werden kann. Daher lassen sich die drei in den Sozialwissenschaften am häufigsten verwendeten Modelle in den nächsten Gleichungen darstellen:

$$y_i \sim N(\mu_i, \sigma^2) \tag{30}$$

$$g(\mu_i) = \mu_i$$

$$\theta_i = \eta_i = \mu_i = \sum_{j=1}^{p} x_{ij}\beta_j$$

$$\phi = \sigma^2$$

Gleichung (30) beschreibt das bekannte lineare Modell mit Normalverteilung.

Im nächsten Modell ist die abhängige Variable die Häufigkeit der ersten Ausprägung einer dichotomen abhängigen Variablen. Diese Häufigkeit ist für jede Kombination der unabhängigen Variablen binomial verteilt mit Wahrscheinlichkeit π_i und Stichprobengröße m_i:

$$y_i \sim B(\pi_i, m_i) \quad \text{mit} \quad \mu_i = m_i\pi_i \tag{31}$$

$$\eta_i = g(\mu_i) = \ln(\pi_i/(1-\pi_i)) \quad \text{logit link}$$

$$\eta_i = \Phi^{-1}(\pi_i) \quad \text{probit link}$$

$$\eta_i = \ln(-\ln(1-\pi_i)) \quad \text{Komplementäres log log link}$$

$$\eta_i = \sum_{j=1}^{p} x_{ij}\beta_j$$

$$\phi = 1$$

Nur für den Fall des logit link sind linearer Prädiktor und kanonischer Parameter identisch. Jede der angegebenen link Funktion stellt ein Modell dar, für das detaillierte Vorstellungen über die zu Grunde liegenden Prozesse entwickelt wurden,

nämlich Choice Models der Psychometrie und Ökonometrie (Manski 1981) im Fall von logit und probit, Infektionsmodelle der Epidemiologie (Arminger 1982) im dritten Fall. Die Häufigkeiten y_i können auch die Werte $\{0,1\}$ annehmen, so daß nicht nur Häufigkeiten, sondern auch individuelle Daten behandelt werden können. Dies tritt immer dann auf, wenn die exogenen Variablen x_j quantitativ sind und damit individuell verschiedene Werte annehmen können, z.B. Einkommen. Man beachte, daß alle link Funktionen die Eigenschaft besitzen, daß die aus der inversen Transformation geschätzten Wahrscheinlichkeiten immer im Intervall $[0,1]$ liegen.

Dies wäre in einem linearen Modell nicht der Fall, da bei beliebig großen x_j ab einem bestimmten Punkt der Wert 0 unterschritten bzw. überschritten wird, wenn der geschätzte Regressionskoeffizient ungleich 0 ist. Eine weitere Implikation ist, daß es vom jeweiligen Standort von η abhängt, wie groß der Effekt von x_j auf die Wahrscheinlichkeit ist. Befindet man sich bereits in der Nähe von 1 bei π, so bedarf es größerer Zuwächse in x_j, um π noch zu erhöhen als bei $\pi = 0,5$, sofern der geschätzte Regressionskoeffizient größer 0 ist. Das Modell ist eben nicht mehr im Erwartungswert linear, sondern in einer Transformation des Erwartungswerts. Da $\pi \varepsilon [0,1]$ sein muß, ist diese Eigenschaft durchaus wünschenswert und entspricht unseren Beobachtungen der Realität.

Im letzten Modell ist die abhängige Variable y eine Häufigkeit, die einer Poissonverteilung mit Erwartungswert μ folgt.

$$y_i \sim P\mu_i \quad \text{mit} \quad E(y_i) = \mu_i \tag{32}$$

$$\theta_i = \eta_i = \ln \mu_i \qquad \text{loglineares link}$$

$$\eta_i = \sum_{j=1}^{p} x_{ij}\beta_j$$

$$\phi = 1$$

Sind die x_{ij} ausschließlich Dummy Variable, ist (32) äquivalent zur Multinomialverteilung, die eine Kontingenztabelle

beschreibt (Haberman 1974, Arminger 1976). Den Zusammenhang mit dem herkömmlichen χ^2 Test auf Unabhängigkeit kann man sofort an Hand eines einfachen Beispiels herstellen. Sei y_{ij} $i = 1,2$, $j = 1,2$ die Häufigkeit in einer 2x2 Kontingenztabelle. Verwenden wir centered effects für die x_{ij}, die beschreiben, in welcher Zelle der Tabelle wir uns befinden, so erhalten wir für $Ey_{ij} = \mu_{ij}$ folgende Darstellung in Matrixschreibweise:

	B_1	B_2
A_1	y_{11}	y_{12}
A_2	y_{21}	y_{22}

2 x 2 Tabelle

$$\begin{bmatrix} \eta_{11} \\ \eta_{12} \\ \eta_{21} \\ \eta_{22} \end{bmatrix} = \begin{bmatrix} \ln \mu_{11} \\ \ln \mu_{12} \\ \ln \mu_{21} \\ \ln \mu_{22} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix} \cdot \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix}$$

Die letzte Spalte (Interaktion) entsteht aus Multiplikation der zweiten und dritten Spalte von $\underline{X}$, die die Haupteffekte beschreiben. Einfaches Nachrechnen zeigt, daß $\beta_4 = 0$ äquivalent zur statistischen Unabhängigkeit von A und B ist. Sind zusätzlich β_2 und $\beta_3 = 0$, folgt die Tabelle [AB] einer Gleichverteilung. Mit Hilfe des linearen Modells im Logarithmus von μ_i, lassen sich also Modelle über die Wahrscheinlichkeiten einer Tabelle spezifizieren. Die gängigsten Modelle sind bedingte Gleichverteilung und Unabhängigkeit in mehrdimensionalen Kontingenztabellen.

Wie leicht mit Modell (32) Hypothesen spezifiziert und überprüft werden können, läßt sich am relationalen Ansatz zur Analyse sozialer Strukturen (Marsden 1981) zeigen. Der Einfachheit

halber nehmen wir drei soziale Schichten I, II und III, geordnet nach fallendem Prestige an. In jeder Schicht werden Personen befragt, welcher Schicht ihre drei besten Freunde angehören. Dann läßt sich als empirische Regelmäßigkeit feststellen, daß die Diagonalzellen der entsprechenden Kontingenztabelle weitaus am stärksten besetzt sind und daß die Zellen unterhalb der Diagonale stärker besetzt sind als die oberhalb der Diagonale. Wir illustrieren dies durch eine Kontingenztabelle, in der die Anzahl von + die Stärke der Besetzung symbolisiert.

		Schicht der Freunde I	II	III
Schicht	I	+++	0	0
des	II	+	+++	0
Befragten	III	0	+	+++

Ein einfaches Modell wäre dann das sogenannte "differential inbreeding", das für die Tabelle statistische Unabhängigkeit annimmt, aber für jede Diagonalzelle einen - jeweils für jede Schicht verschieden starken - positiven Effekt, die Freunde aus der eigenen Schicht zu wählen postuliert. Mit herkömmlichen Methoden, eine Kontingenztabelle zu analysieren, läßt sich diese Hypothese nur schwer formulieren und überprüfen. Mit Hilfe geeigneter Spezifikationen der Designmatrix $\underline{X}$ ist dies leicht.

Die Koeffizienten β, mit denen die Spalten von $\underline{X}$ multipliziert werden, stehen über den zu ihnen gehörigen Spalten. Es wurde die Reparametrisierung über "cornered effects" gewählt, die Bezugskategorie ist Schicht I, dementsprechend sind β_1^A und β_1^B gleich 0. β_1 ist die Regressionskonstante, also die Häufigkeit der Schicht I, β_2^A und β_3^A geben die Differenzen in den Randhäufigkeiten von II und III im Vergleich zu I an; analog sind β_2^B und β_3^B definiert. Sind nur β_1, β_2^A, β_3^A, β_2^B, β_3^B ungleich 0, ist

$$
\begin{bmatrix} \ln \mu_{11} \\ \ln \mu_{12} \\ \ln \mu_{13} \\ \ln \mu_{21} \\ \ln \mu_{22} \\ \ln \mu_{23} \\ \ln \mu_{31} \\ \ln \mu_{32} \\ \ln \mu_{33} \end{bmatrix} = \begin{bmatrix} \beta_1 & \beta_2^A & \beta_3^A & \beta_2^B & \beta_3^B & \beta_2^D & \beta_3^D \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{bmatrix}
$$

die Hypothese der Unabhängigkeit formuliert. Wir lassen zusätzlich Diagonaleffekte β_2^D, β_3^D für die Diagonalzellen zu. Die entsprechenden Spalten weisen eine 1 für μ_{22} bzw. μ_{33} auf und eine 0 sonst.

Ein Spezialfall ist das "constant inbreeding", in dem ein gleich starker Effekt für alle Diagonalzellen angenommen wird. Die Parameter β_2^D und β_3^D werden dann ersetzt durch einen Koeffizienten β^D, die beiden letzten Spalten von $\underline{X}$ werden gestrichen, an ihre Stelle tritt der Vektor (100010001) transponiert. Alle anderen Hypothesen, z.B. Symmetrie oder Quasi-Symmetrie lassen sich in ähnlicher Weise über $\underline{X}$ spezifizieren.

Bevor wir weiter auf die Analyse qualitativer abhängiger Variablen eingehen, geben wir noch die Berechnungsmethode für den GLM Ansatz an, der einige interessante Schlüsse zuläßt. Es wird nach dem Maximum Likelihood Prinzip geschätzt; die ML Schätzer für β_j werden mit Hilfe iterativer gewichteter Regression berechnet. Detaillierte Ableitungen sind in Arminger (1982) enthalten. Wir benutzen wieder Matrixschreibweise:

q = 0,1... sei der Laufindex der Iteration

$$\underline{b}^{q+1} = (\underline{X}'\underline{W}^q\underline{X})^{-1}\,\underline{X}'\underline{W}^q(\underline{\eta}^q+\underline{r}^q) \tag{33}$$

$$\underline{\eta}^q = \underline{x}\underline{\beta}^q \tag{34}$$

$$\underline{r}^q = (r_i^q) \quad i = 1,\ldots,n \quad \text{mit} \quad r_i^q = (y_i-\mu_i^q)(d\eta_i^q/d\mu_i^q) \tag{35}$$

$$\mu_i^q = g^{-1}(\eta_i^q) \quad i = 1,\ldots, n$$

$$\underline{W}^q = \text{diag}\,\{w_i^q\}\ i = 1,\ldots,n \tag{36}$$
ist die Diagonalmatrix der Gewichte

$$w_i^q = (V^q(y_i)(d\eta_i^q/d\mu_i^q)^2)^{-1}$$

$V^q(y_i)$ ist die geschätzte Varianz von y_i.

Vergleicht man (33) - (36) mit der üblichen Berechnung, erkennt man, daß die link Funktion über die Ableitung $(d\eta_i/d\mu_i)$ in die Berechnung eingeht und daß mit der Varianz gewichtet wird. Ist die Berechnung beendet - in der Regel sind es 4 oder 5 Iterationen - erhält man analog zur Regressionsanalyse eine geschätzte Kovarianzmatrix der Schätzer, die unter Regularitätsbedingungen normal verteilt sind (Nordberg 1980, Küsters 1983).

$$\hat{\underline{V}} = (\underline{X}'\underline{W}'\underline{X})^{-1} \tag{37}$$

$\hat{\underline{V}}$ ist die Kovarianzmatrix von $\underline{b}$.

Auch hier lassen sich wieder Tests und Konfidenzintervalle konstruieren. Das Analogon zu R^2 ist im allgemeinen Fall die Devianz, die im Spezialfall der loglinearen Modelle definiert ist als:

$$G^2 = 2\sum_{i=1}^{n} y_i \ln(y_i/\mu_i) \tag{38}$$

Gleichungen (33) - (37) gelten allgemein für den GLM Ansatz, daher lassen sie sich auch auf Häufigkeiten als Spezialfall übertragen. Die in (32) angegebenen loglinearen Modelle sind

sowohl für die Analyse von Kontingenztabellen allgemein als auch von abhängigen qualitativen Variablen zu verwenden. Im ersten Fall werden bestimmte Randverteilungen als variabel angenommen, im zweiten Fall ist die Verteilung für jede Kombination der unabhängigen Variablen durch die Stichprobe fest vorgegeben. Es kann gezeigt werden, daß die von Grizzle, Starmer und Koch (1969) entwickelte gewichtete Regression zur linearen oder logistischen Analyse von relativen Häufigkeiten, die im Programmsystem NONMET verwendet und im deutschen Sprachraum von Küchler (1979) propagiert wird, der erste Schritt des Iterationsverfahrens der Gleichungen (33) - (36) ist (Arminger 1983a).

Gegen die bisher von Sozialwissenschaftlern verwendeten Verfahren, die als spezielle (log)lineare Modelle zu charakterisieren sind (z.B. Programmsystem ECTA oder NONMET) lassen sich folgende Einwände formulieren:

- Quantitative unabhängige Variable sind nicht zugelassen.
- Treten fehlende Zellen auf, so können die Parameter nicht geschätzt werden.
- Die "Varianz" der qualitativen abhängigen Variablen ist nicht definiert. Ein "saturiertes" Modell mit allen Interaktionen der exogenen Variablen erklärt immer 100 % der Devianz. Es kommt dadurch zum paradoxen Ergebnis, daß die Einführung zusätzlicher exogener Variablen immer weniger erklärt, sofern man nur die Haupteffekte betrachtet.
- Wenn die Zellen der Kontingenztabelle nur gering besetzt sind, sind die asymptotischen Eigenschaften der Schätzer ungeklärt. Insbesondere ist keine Analyse auf individueller Ebene, sondern nur auf aggregierter Ebene möglich.

Wissenschaftsgeschichtlich ist es durchaus von Interesse, daß diese Probleme innerhalb der Ökonometrie bereits Anfang der 70er Jahre befriedigend gelöst wurden (McFadden 1973), während sich innerhalb der Sozialwissenschaften und selbst in der Statistik ausschließlich Goodman's Schule (Goodman 1978) durchsetzte. Aus Gleichung (32) ist nun sofort ersichtlich, daß in

der iterierten Regression (33) nicht nur Dummy Variable, die qualitative unabhängige Variable beschreiben, zulässig sind. Das Problem der fehlenden Zellen ist in der iterierten Regression identisch mit der Invertierbarkeit der Kreuzproduktmatrix ($\underline{X}'\underline{W}'\underline{X}$) in (33), die ausschließlich vom Rang von $\underline{X}$ abhängt. Ist die Anzahl p der exogenen Variablen größer als der Rang von $\underline{X}$, müssen so lange Spalten von $\underline{X}$ und die korrespondierenden Parameter β_j gestrichen werden, bis $\underline{X}$ wieder vollen Spaltenrang hat. Damit sind die Parameter wieder schätzbar. Beide Probleme sind im Programmsystem GLIM 3, das einen großen Teil der GLM Modelle enthält, gelöst (Arminger 1983a).

Die beiden verbleibenden Probleme lassen sich ebenfalls in GLIM mit Hilfe des Tricks lösen, daß jede Person s-fach vorkommt, mit Häufigkeit 1 in der für sie zutreffenden Kategorie und mit Häufigkeit 0 in den anderen (s-1) Kategorien der abhängigen Variablen (Arminger 1983b). Daraus läßt sich auch eine Formel für die Devianz einer qualitativen abhängigen Variablen ableiten, die nicht von der Anzahl und Auswahl der exogenen Variablen abhängt. Damit ist analog zu R^2 die Berechnung des Anteils an erklärter Devianz möglich. Sei n die Stichprobengröße und p_k der beobachtete Anteil in Kategorie k, k=1,...,s der abhängigen Variablen. Die Devianz D ist dann wie folgt definiert:

$$D = 2n \sum_{k=1}^{s} p_k \ln(1/p_k) \qquad (39)$$

Dies entspricht der Quadratsumme

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

im linearen Modell.

Insgesamt lassen sich damit alle Aussagen über die Regression aus Gleichung (4) auf verallgemeinerte lineare Modelle, insbesondere auf Modelle zur Analyse qualitativer Variablen, voll übertragen, nur die tatsächliche Berechnung ist etwas komplizierter. Daraus folgt, daß alle Anmerkungen zu Spezifikations-

fehlern und konfirmatorischer vs. explorativer Statistik auch für diese Modelle gelten. Da $\underline{X}$ keinen Einschränkungen unterliegt, sind weiter alle Möglichkeiten vorhanden, die Matrix $\underline{X}$ so zu spezifizieren, daß Interaktionen, konditionale Effekte und Restriktionen aufgespürt oder überprüft werden können. Auch hier sind die "Theoretiker" aufgerufen, soziologische Phantasie in der Auswahl von Variablen und der Art des Zusammenhangs walten zu lassen.

Am Schluß dieses Abschnittes sei noch erwähnt, daß es mit Hilfe von zusammengesetzten link Funktionen (Thompson und Baker 1981) möglich ist, die von McCullagh (1980) entwickelten Modelle für abhängige ordinale Variable in den GLM Ansatz einzubetten. Dies gilt auch für die Latent Class Analysis (Goodman 1978), die als faktorenanalytische Reduktion von manifesten auf latente qualitative Variable angesehen werden kann (Arminger 1983b). Damit ist ein wichtiger Schritt zur Ausdehnung der verallgemeinerten linearen Modelle auf Strukturgleichungen mit latenten Variablen analog zum LISREL Ansatz für lineare Modelle getan. Das Strukturmodell von Gleichung (18) läßt sich für endogene qualitative Variable im Logitmodell direkt übertragen (Schmidt und Strauss 1975); sind die endogenen Variablen sowohl nominal als auch intervall skaliert, sind multivariate Probitmodelle vorzuziehen, die allerdings erhebliche numerische Probleme aufwerfen (Heckman 1978).

Insgesamt läßt sich festhalten, daß in den letzten 15 Jahren Statistik, Ökonometrie, Psychometrie und die "Methodiker" der Soziologie den Sozialwissenschaften Modelle zur Verfügung gestellt haben, die eine einheitliche Betrachtung der Verknüpffung beliebiger Meßniveaus erlauben. Es liegt nun an den Sozialwissenschaftlern, Variablen und mögliche Beziehungen zu spezifizieren.

3. Dynamische Modelle

Wir sind bis jetzt davon ausgegangen, daß die Beobachtungen für jedes Element der Stichprobe voneinander unabhängig sind, wie dies für Querschnittsdaten der Fall ist. Für die Analyse

sozialer Prozesse jedoch kann dies nicht mehr angenommen werden. Wir beobachten dann ein Element zu mehreren Zeitpunkten oder erheben retrospektiv Stationen seiner individuellen Geschichte, wie dies etwa bei der Erforschung von Berufskarrieren, lebensgeschichtlichen Zyklen oder Krankheitsverläufen der Fall ist. Wir beschränken uns dabei auf Modelle für eine relativ große Stichprobe mit wenigen Beobachtungen pro Element in der Zeit, z.B. Panelbefragungen. Dieser Fall ist typisch für die Sozialwissenschaften. Der konträr gelagerte Fall von kleinen Stichproben mit vielen Beobachtungen in der Zeit tritt eher in der Ökonomie auf und wird dort mit den Modellen der Zeitreihenanalyse (Box und Jenkins 1976) oder der Spektralanalyse (Anderson 1973) behandelt. Charakteristisch für die Prozeßbetrachtung ist, daß der Meßwert der abhängigen Variablen y zum Zeitpunkt t von früheren Meßwerten von y und von exogenen Variablen x, die sich ebenfalls in der Zeit ändern können, abhängt. Diese Abhängigkeit kompliziert sowohl die Modellbildung, da diese Abhängigkeit auch spezifiziert werden muß, als auch die numerischen Methoden, die zur Schätzung herangezogen werden müssen. Es ist daher verständlich, daß an dieser Stelle nur ein kleiner Ausschnitt der vorhandenen Möglichkeiten angerissen werden kann.

Beeinflussen sich mehrere Variablen in der Zeit gegenseitig, sprechen wir von dynamischen Systemen, deren Modellierung nach meiner Auffassung eine der wichtigsten Aufgaben der Sozialwissenschaft ist. Von Bedeutung ist noch die Unterscheidung, ob zeitliche Veränderungen nur zu bestimmten Zeitpunkten (Prozesse mit diskretem Parameterraum) oder zu beliebigen Zeitpunkten (Prozesse mit stetigem Parameterraum) auftreten können. Hier beschränken wir uns auf stetige Prozesse, die wohl den Regelfall in den Sozialwissenschaften darstellen.

3.1 Diskreter Zustandsraum: Rate Modelle

In der Soziologie wurde bereits vor ca. 20 Jahren begonnen (Coleman 1964), Modelle für den Übergang von einer Merkmalsausprägung j in eine andere k des selben Merkmals zu formulieren. Eine treibende Kraft war die Erforschung sozialer Mobili-

tät. Zunächst wurde als theoretisches Modell eine einfache Markoffkette verwendet. Sie weist unter anderen folgende Eigenschaften auf:

Sei $\underline{P} = (p_{jk})$ $j,k = 1,\ldots s$ die zeitunabhängige Matrix der Übergangswahrscheinlichkeiten vom Zustand j zum Zeitpunkt t-1 in den Zustand k zum Zeitpunkt t. Unter bestimmten Regularitätsbedingungen gilt dann:

$$\underline{P}(t-k,t) = \underline{P}^k \tag{40}$$

Bei Anwendung dieses einfachen Modells auf Mobilitätstabellen wurde festgestellt, daß regelmäßig die Wahrscheinlichkeiten in der gleichen Klasse wie der Vater bzw. der Großvater zu bleiben, unterschätzt wurden, d.h. die Mobilität wurde überschätzt.

Das einfache Modell wurde daher in dreierlei Hinsicht abgeschwächt (Tuma et al. 1979 , Coleman 1981, Hannan und Tuma 1983):

- an Stelle der diskreten Übergangswahrscheinlichkeiten werden Übergangsraten in stetigen Prozessen untersucht. Die Übergangsrate $r_{jk}(t)$ ist definiert als:
$$r_{jk}(t) = \lim_{\Delta t \to 0} \frac{P_{jk}(t,t+\Delta t)}{\Delta t} \quad , \; r_{jk} \geq 0 \tag{41}$$
- die Annahme, daß sich alle Personen durch gleiche Übergangswahrscheinlichkeiten charakterisieren lassen (Homogenität der Population) wurde zu Gunsten der Annahme, daß sich Individuen je nach sie kennzeichnenden Eigenschaften in den Übergangsraten unterscheiden, fallen gelassen (Heterogenität der Population).
- die Annahme, daß sich die Übergangsraten in der Zeit nicht ändern (Stationarität des Prozesses) wurde aufgegeben.

Kann man die individuelle Übergangsrate $r_{ijk}(t)$ schätzen, so läßt sich z.B. daraus berechnen, wie lange ein Individuum im Durchschnitt in einem Zustand bleibt und wie wahrscheinlich es ist, daß es in einen bestimmten anderen Zustand wechselt. Diese Modelle sind daher besonders geeignet zur Analyse sogenann-

ter "event histories" und werden zur Zeit vor allem in Karriereverläufen und in der Organisationsforschung angewandt. Die Spezifikation der oben genannten Erweiterungen läßt sich wieder analog zum einfachen Regressionsmodell durchführen. Gebräuchliche Modelle für r_i sind (die Indizes r_{jk} werden der einfacheren Notation halber weggelassen):

$$\ln r_i(t) = \sum_{j=1}^{p} x_{ij}\beta_j \tag{42}$$

$$\ln r_i(t) = \sum_{j=1}^{p} x_{ij} + \varepsilon_i \tag{43}$$

$$\ln r_i(t) = \beta_0(t) + \sum_{j=1}^{p} x_{ij}\beta_j \tag{44}$$

$$\ln r_i(t) = \sum_{j=1}^{p} x_{ij}\beta_j + \left(\sum_{k=1}^{q} x_{ik}\gamma_k\right) \exp\left(-\left(\sum_{1=1}^{r} x_{i1}\delta_1\right)t\right) \tag{45}$$

Da die Bedingung r größer gleich 0 erfüllt sein muß, werden an Stelle der r die logarithmierten Werte eingesetzt. Gleichung (42) ist das einfachste Modell, das nur beobachtete Heterogenität der Population einschließt. Im Gegensatz dazu beinhaltet (43) mit ε_i auch Heterogenität, die nicht beobachtet werden kann. Gleichung (44) ist ein Modell für Zeitabhängigkeit, die allerdings für alle Personen gleich ist, und beobachtete Heterogenität. Das letzte Modell schließlich ist ein parametrisches Modell - das verallgemeinerte Makeham Gompertz Gesetz - für die Einschließung von Heterogenität und individueller Zeitabhängigkeit. Die exogenen Variablen x_{ij} sind bekannt, die Koeffizienten β_j, γ_k, δ_1 sind zu schätzen. Diese Schätzung erfolgt nach dem Maximum Likelihood Prinzip, wenn die Zeiten, wann jemand seinen Zustand wechselt, bekannt sind oder mit nicht parametrischen Schätzverfahren, wie dem Partial Likelihood Ansatz (Cox 1975). Die tatsächliche Berechnung kann mit Hilfe des Computerprogramms RATE (Tuma 1980) erfolgen, einige Modelle lassen sich auch auf Poissonmodelle zurückführen und können daher mit GLIM 3 berechnet werden (Arminger 1983c).

Auch für die in Gleichungen (42) und (45) angegebenen Modelle gelten, da sie in der Struktur Gleichung (4) entsprechen, alle zum einfachen Regressionsmodell gemachten Aussagen über Spezifikation, Konfidenzintervalle und Tests.

3.2 Stetiger Zustandsraum: Dynamische Systeme

Im letzten Abschnitt wurde jeweils nur die Veränderung einer qualitativen Variablen untersucht, nicht jedoch ein System von sich gegenseitig in der Zeit verändernden Variablen. Für quantitative Variable lassen sich derartige Systeme mit Hilfe stochastischer Differentialgleichungen entwickeln. Dieser Ansatz zur Analyse sozialer Prozesse wurde ebenfalls von Coleman (1968) formuliert und in der Organisationssoziologie von Doreian und Hummon (1976) und Doreian (1981) aufgegriffen und weiter entwickelt. Hier wird nur der Fall linearer stochastischer Differentialgleichungen angegeben, der allerdings den weitaus größten Anteil bisher erschienener Literatur zu diesem Thema ausmacht.

Sei $\underline{y}(t)$ der Vektor der endogenen, sich gegenseitig in der Zeit beeinflussenden, Variablen; $\underline{x}$ sei ein Vektor von exogenen Variablen und $\underline{u}(t)$ ein Fehler, der selbst einem stochastischen Prozeß folgt. In Matrixschreibweise kann das System in Form der nächsten Gleichung geschrieben werden:

$$d\underline{y}(t)/dt = \underline{A}\underline{y}(t) + \underline{B}\underline{x} + \underline{u}(t) \qquad (46)$$

$d\underline{y}(t)/dt$ ist die Änderungsrate in den einzelnen Variablen. Sind $\underline{A}$ und $\underline{B}$ sowie ein Startwert $\underline{y}(0)$ bekannt, lassen sich die Erwartungswerte von $\underline{y}$ für beliebige Zeitpunkte größer 0 berechnen. Für das Verhalten des Systems ist vor allem $\underline{A}$ von Bedeutung.

Die Diagonalzellen von $\underline{A}$ geben an, inwieweit jede Variable sich selbst verstärkt oder abschwächt. Die Werte außerhalb der Diagonale geben an, wie jede Variable die anderen im System beeinflußt; z.B. läßt sich damit die Frage nach Reziprozität sofort beantworten. An den Eigenwerten von $\underline{A}$ läßt sich ablesen, ob das System zu einem Gleichgewichtspunkt tendiert (alle

Eigenwerte haben einen Realteil kleiner 0) oder nicht, und ob Oszillationen (komplexe Eigenwerte) auftreten oder nicht. $\underline{A}$ und $\underline{B}$ gemeinsam legen den Gleichgewichtspunkt fest, - wenn Gleichgewicht existiert - und sind damit der Ausgangspunkt für strukturelle Analysen. Die Größe der Eigenwerte von $\underline{A}$ legt auch fest, wie rasch sich das System ändert. Obwohl die Bedeutung dieses Denkmodells für die Sozialwissenschaften bereits 1969 von Blalock erkannt und dargestellt wurde, wurde Gleichung (46) auf Grund numerischer Schwierigkeiten bei der Schätzung von $\underline{A}$ und $\underline{B}$ nur wenig als Modell verwendet. Auch dieses Modell läßt sich aber unter Annahme der Normalverteilung des integrierten Fehlers nach dem ML Prinzip unter Hinzunahme von latenten Variablen aus Paneldaten (Arminger 1983d) schätzen. Wie bei allen zuvor genannten Verfahren läßt sich auch hier die Kovarianzmatrix der Schätzer angeben. Dies ermöglicht wieder die Konstruktion von Konfidenzintervallen und Tests.

Da auch Gleichung (46) im wesentlichen ein lineares Modell - wenn auch nicht für $\underline{y}(t)$, sondern für die Veränderungsraten - spezifiziert, bleiben alle Aussagen über Fehlspezifikation und die Verwendung statistischer Tests erhalten.

4. Zusammenfassung von Diskussion

Ausgehend vom einfachsten statistischen Modell zur Abbildung sozialer Realität, dem Mittelwert, wurde versucht zu zeigen, daß dem Sozialwissenschaftler in den letzten 15 Jahren eine Fülle von Modellen bereitgestellt wurde, um für beliebige Meßniveaus sowohl strukturelle als auch dynamische Modelle zu formulieren. Ferner wurde gezeigt, daß allen besprochenen Modellen die gleiche Struktur zu Grunde liegt, wenn sie auch durch neue Fehlerverteilungen und link Funktionen erweitert wird. Dies stellt eine große Erleichterung dar, da es genügt, das Meßniveau aller Variablen und die Designmatrix $\underline{X}$ anzugeben. Diese Arbeit, die ja aus den verwendeten Variablen, ihren Operationalisierungen und den inhaltlichen Hypothesen folgt, kann und soll niemand dem Sozialwissenschaftler abnehmen.

Jeder Sozialwissenschaftler, der eine bestimmte statistische Methode anwendet, sollte sich darüber im klaren sein, daß er in Wirklichkeit ein Modell über soziale Realität anlegt. Verwendet er ein besonders einfaches Modell, ist es auf Grund der Verknüpfung der Variablen meist zu restriktiv und daher falsch. Wenn er ein falsches Modell verwendet hat bzw. erhebliche Fehlspezifikationen getroffen hat, sollte dies aber nicht der Statistik angelastet werden, sondern dem eigenen Unvermögen, wichtige Variablen herauszufinden und spezielle Effekte durch Interaktionen oder Bedingungen zu formulieren. Schließlich sollte die Tatsache, daß man gezwungen ist, Variable klar zu formulieren und Zusammenhänge genau zu definieren, und die Möglichkeit, die eigenen Hypothesen zu überprüfen, nicht negativ, sondern positiv bewertet werden.

Bibliographie

Andersen, E.B. (1980), Discrete statistical models with social science applications, Amsterdam

Arminger, G. (1976), Loglineare Modelle zur Analyse nominal skalierter Variablen, Wien

- (1979), Faktorenanalyse, Stuttgart
- (1982), Klassische Anwendungen verallgemeinerter linearer Modelle in der empirischen Sozialforschung, in: ZUMA Arbeitsbericht No. 1982/03, 1-124, Mannheim
- (1983a), Multivariate Analyse von qualitativen abhängigen Variablen mit verallgemeinerten linearen Modellen, Zeitschrift für Soziologie 12, 49-64
- (1983b), Analysis of qualitative individual data and of latent class models with generalized linear models, in: Measuring the unmeasurable: Proceedings of the advanced research workshop on qualitative spatial data, P. Nijkamp (ed.), The Hague, in Druck
- (1983c), Analysis of event histories with generalized linear models, in: Progress in stochastic modeling of social processes, A. Diekmann/P. Mitter (Eds.), New York, in Druck
- (1983d), Estimation of parameters of linear stochastic differential equations and their covariances from panel data, paper presented at the Annual Meeting of the American Sociological Association in Detroit, 1983

Blalock, H.M. (1969), Theory construction, Englewood Cliffs

Bock, R.D. (1975), Multivariate statistical methods in behavioral research, New York

Coleman, J.S. (1964), Introduction to mathematical sociology, New York

- (1968), The mathematical study of change, in: Methodology in social research, H.M. and A. Blalock (Eds.), New York

- (1981), Longitudinal data analysis, New York

Cox, D.R. (1975), Partial likelihood, in: Biometrika 62, 269-276

Dhrymes, Ph.Y. (1974), Econometrics, New York

Doreian, P./ N.P. Hummon (1976), Modelling social processes, New York

Doreian, P. (1981), Models or organizational change, in: Mathematische Analyse von Organisationsstrukturen und Prozessen, W. Sodeur (Ed.), Duisburg

Goodman, L.A. (1978), Analysing qualitative/categorial data, London

Grizzle, J.E./C.F. Starmer/G.G. Koch (1969), Analysis of categorial data by linear models, in: Biometrics 25, 489-504

Haberman, S.J. (1974), The analysis of frequency data, Chicago

Heckman, J.J. (1978), Dummy endogenous variables in a simultaneous equation system, in: Econometrica 46, 931-959

Hannan, M./ N. Tuma (1983), Dynamic analysis of qualitative variables: applications to organizational demography, in: Measuring the unmeasurable: Proceeding of the advanced research workshop on qualitative spatial data, P. Nijkamp (Ed.), The Hague, in Druck

Holm, K. (1979), Die Befragung, Bd. 6, München

Holm, S. (1979), A simple sequentially rejective multiple test procedure, in: Scandinavian Journal of Statistics 6, 65-70

Hui, B.S./H. Wold (1982), Consistency and consistency at large of partial least squares estimates, in: Systems under indirect observation, K.G. Jöreskog/H. Wold (Eds.), Amsterdam

Jöreskog, K.G./D. Sörbom (1977), Statistical models and methods for analysis of longitudinal data, in: Latent variables in socio - economic models, D.J. Aigner/A.S. Goldberger (Eds.), Amsterdam

Jöreskog, K.G. (1982), The LISREL approach to causal model building in the social sciences, in: Systems under indirect observation, K.G. Jöreskog/H. Wold (Eds.), Amsterdam

Judge, G.G./W.E. Griffiths/R.C. Hill/T.C. Lee (1980), The theory and practice of econometrics, New York

Kriz, J. (1981), Methodenkritik empirischer Sozialforschung, Stuttgart

Küchler, M. (1979), Multivariate Analyseverfahren, Stuttgart

Küsters, U.L. (1983), Likelihood Theorie für Folgen von stochastisch unabhängigen nicht identisch verteilten Zufallsvariablen aus regulären Exponentialfamilien, Diplomarbeit in Ökonometrie am Fachbereich Wirtschaftswissenschaften, Universität Wuppertal

Manski, C.F. (1981), Structural models for discrete data: the analysis of discrete choice, in: Sociological Methodology 1981, S. Leinhardt (Ed.), San Francisco

Marsden, P.V. (1981), Models and methods for characterizing the structural parameters of groups, in: Social Networks 3, 1-27

McCullagh, P. (1980), Regression methods for ordinal data, in: Journal of the Royal Statistical Society B 42, 109-142

McFadden, D. (1973), Conditional logit analysis of qualitative choice behavior, in: Frontiers of Econometrics, P. Zarembka (Ed.), New York

Nelder, J.A./R.W.M. Wedderburn (1972), Generalized Linear Models, Journal of the Royal Statistical Society A 135, 370-383

Nordberg, L. (1980), Asymptotic normality of maximum likelihood estimators based on independent, unequally distributed observation in exponential family models, in: Scandinavian Journal of Statistics 7, 27-32

Olsson, U. (1979), Maximum likelihood estimation of the polychoric correlation coefficient, in: Psychometrika 44, 443-460

Olsson, U./F. Drasgow/N.J. Dorans (1982), The polyserial correlation coefficient, in: Psychometrika 47, 337-347

Schmidt, P./R.P. Strauss, Estimation of models with jointly dependent qualitative variables: a simultaneous logit approach, in: Econometrica, 745-755

Thompson, R./R.J. Baker , Composite link functions in generalized linear models, in: Applied Statistics 30, 125-131

Tuma, N./ M. Hannan/L. Groeneveld (1979), Dynamic analysis of event histories, in: American Journal of Sociology 84, 820-854

Tuma, N. (1980), Invoking rate, ZUMA, Mannheim

Victor, N./ W. Lehmacher/W. van Einerem (Eds.) (1980), Explorative Datenanalyse, Berlin

Wilson, T.P. (1980), Exponential family regression models, unpublished manuscript, University of California, Santa Barbara

Wittgenstein, L. (1980), Tractatus logico - philosophicus, 16. Aufl., Frankfurt am Main

Wold, H. (1982), Soft modeling: the basic design and some extensions, in: Systems under indirect observation, K.G. Jöreskog/H. Wold (Eds.), Amsterdam