Marktsegmentierung: Kategoriale Regression vs. Kontrastgruppenanalyse (Automatic Interaction Detector)

Von Alfred Hamerle und Peter Kemény

Ziel dieser Arbeit ist es, nach einer kurzen Darstellung der Theorie des kategorialen Regressionsansatzes aufzuzeigen, daß sowohl aus formal-statistischer Sicht als auch im Hinblick auf die daraus sich ergebende substantielle Relevanz das Modell der kategorialen Regression ein leistungsfähigeres Instrument zur Identifizierung von Marktsegmenten ist als die bislang dafür überwiegend eingesetzte Kontrastgruppenanalyse (AID). Für eine empirische Fallstudie wird mit Hilfe der kategorialen Regression eine Marktsegmentierung durchgeführt und mit den Ergebnissen einer entsprechenden Kontrastgruppenanalyse verglichen.

1. Einleitung

Ein Hauptziel einer erfolgversprechenden Marktsegmentierung ist, den Gesamtmarkt so in Teilmärkte (Marktsegmente) zu zerlegen, daß dadurch in sich homogene Konsumentengruppen identifiziert werden, die sich in ihrem Präferenz- oder Kaufverhalten möglichst deutlich voneinander unterscheiden. Durch diese segmentspezifische Orientierung soll erreicht werden, daß eine gezielte und damit absatzpolitisch relevante Gestaltung von Marketing-Mix-Aktivitäten erfolgen kann.

Anders ausgedrückt bedeutet dies, daß das Instrument der Marktsegmentierung zur Prognose des Konsumentengruppenverhaltens herangezogen werden soll, indem bei der Segmentbildung diejenigen Merkmale identifiziert werden, die einen bedeutsamen Einfluß auf die Variable „Präferenzverhalten" oder „Kaufverhalten" haben, d.h. deren Variabilität möglichst gut erklären.

Damit ließe sich die im Rahmen des Marktsegmentierungsproblems auftretende Fragestellung – Untersuchung der Wirkung mehrerer unabhängiger Variablen (Einflußfaktoren) auf eine oder mehrere abhängige Variablen – formal-statistisch mit dem klassischen regressions- und varianzanalytischen Methodeninstrumentarium behandeln, falls die zu deren Anwendung notwendigen Voraussetzungen, nämlich metrisches Meßniveau sowie die Gültigkeit einer uni- (bzw. multivariaten) Normalverteilungsannahme, gegeben wären. Häufig liegt jedoch im Bereich der Marktsegmentierung folgende Situation vor:

Für die Segmentierung werden Personenmerkmale, die durch die vier Bereiche

- Demographische Merkmale (Alter, Geschlecht, Familienstand, etc.)
- Geographische Merkmale (Stadtrandwohnlage, Citywohnlage, etc.)
- Sozioökonomische Merkmale (Schulbildung, Berufstätigkeit, Einkommensgruppe, Soziale Schicht, etc.)
- Psychographische Merkmale (Einstellungen, Präferenzen, Preisbewußtsein, Qualitätsbewußtsein, Kaufgewohnheiten, etc.)

typisiert werden können (vgl. dazu etwa *Böcker/Thomas*, 1981), als mögliche Einflußgrößen für das Konsumentenverhalten herangezogen. Diese Variablen sind fast ausschließlich **kategorialer Natur.**

Ferner wird in vielen Fällen das Kauf- bzw. Präferenzverhalten durch eine **dichotome Variable**, etwa Kauf der Marke X vs. Nicht-Kauf, Präferenz für Zeitschrift Z vs. keine Präferenz für Zeitschrift Z, etc. definiert. Damit ist der Fall gegeben, daß sowohl die unabhängigen als auch die abhängige Variable kategorial und somit die **klassischen regressions- und varianzanalytischen Verfahren nicht mehr anwendbar** sind.

Zunächst soll nun diese Form der Datensituation anhand eines Teilaspekts der von der *Infratest Forschung GmbH*, München, durchgeführten **Fallstudie „Informationsgewohnheiten von Frauen"** [1] näher beleuchtet werden. Zugrunde liegt der Datensatz einer Befragung über die Informationsgewohnheiten und Informationsbedürfnisse von Frauen, die von einer Tochtergesellschaft der *Infratest Forschung* in einem europäischen Land durchgeführt wurde. Befragt wurde bei dieser Studie eine repräsentative Auswahl von rund 1000 Frauen im Alter von 18 bis 49 Jahren.

Folgender Teilaspekt war dabei von besonderem Interesse: Wie beeinflussen die kategorialen Personenmerkmale Berufstätigkeit (berufstätig/nicht berufstätig), Alter (18–29 Jahre / 30–39 Jahre / 40–49 Jahre) und Schulbildung (Volksschule ohne Lehre / Volksschule mit Lehre / Realschule / Abitur-Universität) die dichotome Zielvariable Präferenz für Zeitschrift Z (ja/nein) und wie gut erklären sie deren Variabilität? Oder anders ausgedrückt: Gibt es durch spezifische Kategorienkombinationen der unabhängigen Merkmale definierte charakteristische Frauengruppen (Marktsegmente), die sich durch möglichst unterschiedliche Präferenzanteile für Zeitschrift Z auszeichnen?

Fragestellungen dieser Art wurden in Ermangelung geeigneter statistischer Verfahren in der Marktforschung mit Hilfe der von *Morgan/Sonquist* (1963) sowie *Sonquist*

Priv.-Dozent Dr. *Alfred Hamerle*, Lehrstuhl für Statistik, Fakultät für Wirtschaftswissenschaften, Universität Regensburg; Dr. *Peter Kemény*, Leiter des Bereichs Statistische Methoden und Auswertungssoftware, *Bundesverband der Unfallversicherungsträger der Öffentlichen Hand*, München.

et al. (1971) entwickelten **Kontrastgruppenanalyse** (AID, Automatic Interaction Detector) untersucht, wobei die dichotome abhängige Variable in Anteilswerte transformiert wurde; man vergleiche dazu in der angelsächsischen Literatur *Assael* (1970), *Carman* (1970), *Armstrong/Andress* (1970), *Heald* (1972) und *Fielding* (1975). Zur Darstellung der Kontrastgruppenanalyse im deutschsprachigen Raum vergleiche man z.B. *Mayntz* et al. (1974) und *Böhler* (1975).

Die AID-Technik teilt die zu untersuchende Gesamtheit durch **sukzessive binäre Segmentation** so in paarweise disjunkte, durch Kategorienkombinationen der dichotomisierten unabhängigen Variablen definierte Teilgruppen (Splits), daß die dadurch entstehende baumähnliche Struktur (Kontrastgruppenstruktur) einen möglichst hohen Erklärungswert für die Variabilität der abhängigen Variablen leistet. Im Kontext der Clusteranalyse stellt die Kontrastgruppenanalyse ein monothetisches und divisives Verfahren dar, bei dem jeweils pro Segmentation nur eine Variable zur Clusterbildung herangezogen wird. Jedoch diskutierten *Sonquist/Morgan* (1964) ihr Distanzmaß nicht in diesem Zusammenhang.

Mit zunehmender Anwendung der AID-Technik wurden auch die **Grenzen** dieser Methode erkannt und es mehrten sich die kritischen Stimmen, wie etwa *Assael* (1970), *Cramer* (1971), *Einhorn* (1972), *Doyle* (1973) sowie *Doyle/Fenwick* (1975). Die entscheidende Schwäche dieses Verfahrens liegt darin, daß es als rein deskriptive Methode die Stichprobenvariabilität überhaupt nicht berücksichtigt und somit inferentielle Schlußfolgerungen nicht gezogen werden können, d.h. es besteht keine Möglichkeit, die entstandene Baumstruktur und daraus ablesbare Interaktionen auf ihre **statistische Signifikanz** und damit auch **substantielle Relevanz** zu überprüfen. Daß die Anwendung der Kontrastgruppenanalyse reine Zufallsprodukte erzeugen kann, hat *Einhorn* (1972) eindrucksvoll dokumentiert. Er wandte die AID-Technik auf verschiedene Sets von unabhängigen Variablen an, die in überhaupt keiner Beziehung zur abhängigen Variablen standen, und stellte fest, daß jedesmal eine Baumstruktur erzeugt wurde, die in Wirklichkeit gar nicht existierte.

Signifikanzüberlegungen wurden erstmals von *Kass* (1975) sowie *Scott/Knott* (1976) angestellt. Sie untersuchten die Möglichkeit, ob zumindest für jede Segmentationsstufe einzeln die Signifikanz des entstehenden Splits getestet werden kann. Dazu leiteten sie unter der Nullhypothese, daß zwischen den Kategorien des für einen Split ausgewählten Prädiktors keine Unterschiede in den Meßwerten der abhängigen Variablen bestehen, die asymptotische Verteilung der aus dem Splitting-Kriterium der Kontrastgruppenanalyse sich auf natürliche Weise ergebenden Teststatistik her und tabellierten die kritischen Werte dieser Statistik für drei Signifikanzniveaus. *Scott/Knott* (1976) verallgemeinerten das nur für Spezialfälle geltende Resultat von *Kass* (1975) und approximierten unter Zugrundelegung nominalskalierter Prädiktoren die asymptotische Verteilung der Teststatistik durch eine χ^2-Verteilung.

Nach wie vor gibt es aber keinen Test, mit dem es möglich wäre, das Ergebnis einer Kontrastgruppenanalyse insgesamt auf Signifikanz zu überprüfen. Da die Konstruktion eines solchen Gesamttests mit großen theoretischen Schwierigkeiten verbunden sein dürfte, erscheint es sinnvoller, einen gänzlich **anderen Weg** einzuschlagen und die im Rahmen der Marktsegmentierung auftretende asymmetrische Fragestellung für kategoriale Prädiktoren und Zielvariablen im Rahmen der **multivariaten Analyse kategorialer Merkmale** zu behandeln, da hierfür eine Reihe von Verfahren zur Verfügung stehen. Insbesondere wurde für die asymmetrische Fragestellung erstmals von *Grizzle/Starmer/Koch* (1969) ein geschlossener Ansatz im Rahmen eines verallgemeinerten linearen Regressionsmodells vorgestellt. Die theoretischen Eigenschaften der mit Hilfe einer gewichteten Kleinst-Quadrate-Methode gewonnenen Schätzungen bei dieser kategorialen Regression gehen auf *Neyman* (1949) und *Bhapkar* (1961, 1966) zurück. Für verschiedene Anwendungsmöglichkeiten im Bereich der Sozial- und Politikwissenschaften vergleiche man *Grizzle/Williams* (1972), *Johnson/Koch* (1971), *Koch/Reinfurt* (1970, 1971), *Kritzer* (1978, 1979), *Küchler* (1979), *Lehnen/Koch* (1974) sowie *Forthofer/Lehnen* (1981).

Im Bereich der **Marktforschung** wurden in diesem Zusammenhang gewisse Teilaspekte der multivariaten Analyse qualitativer Merkmale untersucht, siehe dazu etwa *Green* et al. (1977), *Flath/Leonard* (1979), *Green* (1978), *DeSarbo/Hildebrandt* (1980), *Dillon* (1979), *Perreault/Young* (1980) sowie *Perreault/Barksdale* (1980), die jedoch lediglich Spezialfälle des allgemeinen kategorialen Regressionsansatzes behandeln. Im deutschsprachigen Raum wurde in dieser Zeitschrift das kategoriale Regressionsmodell (GSK-Ansatz) von *Schwedler* (1982) vorgestellt und auf ein Beispiel von *Green* et al. (1977) angewandt.

Ziel dieser Arbeit ist es, nach einer kurzen Darstellung des kategorialen Regressionsmodells aufzuzeigen, daß aus formalstatistischer Sicht und im Hinblick auf die daraus sich ergebende substantielle Relevanz das Modell der kategorialen Regression ein leistungsfähigeres Instrument zur Identifizierung von Marktsegmenten ist als die bislang dafür überwiegend eingesetzte Kontrastgruppenanalyse. In Abschnitt 2 wird das kategoriale Regressionsmodell vorgestellt und in Abschnitt 3 wird für die Fallstudie „Informationsgewohnheiten von Frauen" ein passendes kategoriales Regressionsmodell konstruiert. Die Modellauswahl erfolgt durch eine schrittweise Suchstrategie auf der Basis statistischer Modelltests. Die sich daraus ergebende Marktsegmentierung wird in Abschnitt 4 behandelt. Sie wird mit den entsprechenden Ergebnissen einer von *Infratest* durchgeführten Kontrastgruppenanalyse verglichen. Schließlich werden im Anhang die wichtigsten Formeln zur Schätzung der unbekannten Regressionsparameter und geeignete Teststatistiken zur Überprüfung der Güte der Anpassung eines Modells sowie zur Prüfung der Signifikanz einzelner Parameter bzw. einer Kombination von Parametern zusammengestellt.

2. Das kategoriale Regressionsmodell

Bei der klassischen metrischen Regression geht man aus vom Modell:

$$\mathbf{y} = \mathbf{X\beta} + \boldsymbol{\varepsilon} \qquad \text{mit } E(\boldsymbol{\varepsilon}) = \mathbf{0}. \tag{1}$$

$\mathbf{y} = (y_1, \ldots, y_n)'$ der Beobachtungsvektor der quantitativen (metrischen) Zielvariablen (Regressand, abhängige Variable),

$\mathbf{X} = (1, \mathbf{x}_1, \ldots, \mathbf{x}_p)$ die Regressorenmatrix; sie enthält die Werte der unabhängigen Variablen (Faktoren, Regressoren),

$\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$ der Parametervektor und

$\boldsymbol{\varepsilon} = (\epsilon_1, \ldots, \epsilon_n)'$ der Vektor der Stör- bzw. Fehlervariablen.

Für die üblichen Schätzverfahren werden Varianzhomogenität und Unkorreliertheit der Fehlervariablen, also

$$E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \sigma^2\mathbf{I}, \tag{2}$$

und für die Konstruktion von Tests und Konfidenzintervallen i.a. eine Normalverteilung der Fehlervariablen vorausgesetzt.

Typische Regressoren im Kontext der Marktsegmentierung sind die in Abschnitt 1 aufgeführten Variablen, also sozioökonomische Merkmale wie Alter, Schulbildung, soziale Schichtzugehörigkeit, Berufstätigkeit etc. oder Persönlichkeitsmerkmale wie Einstellungen, Präferenzen, Kaufgewohnheiten, Preis- und Qualitätsbewußtsein etc.

Die Regressoren sind meist quantitativ, können aber auch qualitativ sein. Die **qualitativen Variablen** sind in $\mathbf{X}$ durch eine geeignete Kodierung („Dummy-Variablen") zu repräsentieren. Sind sämtliche Regressoren qualitativ, so ist die Varianzanalyse die adäquate statistische Auswertungsmethode und $\mathbf{X}$ ist dann nur mit 0, +1, −1 besetzt.

Hier betrachten wir nun den Fall der **kategorialen Regression**, bei dem sowohl die unabhängigen als auch die Zielvariable kategorial sind. Die Zielvariable enthält dabei in der Regel nur wenige Kategorien, oft – auch bei der hier analysierten Fallstudie – ist sie nur binär. Man überzeugt sich leicht, daß dann (2) nicht mehr gilt und daher modifizierte Verfahren anzuwenden sind.

Analog zur Varianzanalyse werden die Kategorien der unabhängigen Merkmale in der „Designmatrix" $\mathbf{X}$ durch geeignete Dummy-Variablen kodiert. Dazu gibt es mehrere Möglichkeiten. Für eine ausführliche Darstellung siehe z.B. *Hamerle/Kemény/Tutz* (1983) bzw. Anhang. Beispielsweise wird für eine dichotome unabhängige Variable – etwa Berufstätigkeit mit den Ausprägungen berufstätig/nicht berufstätig – eine Dummy-Variable x eingeführt, die je nach Vorliegen der beiden Ausprägungen die Werte 1 und 0 annimmt. Sind sonst keine weiteren unabhängigen Variablen im Ansatz enthalten, besitzt das Regressionsmodell die Form:

$$y_i = \beta_0 + x\beta + \epsilon \qquad i = 1, \ldots, n. \tag{3}$$

Insbesondere gilt für eine berufstätige Person j

$$y_j = \beta_0 + \beta + \epsilon \tag{4}$$

und für eine nicht berufstätige Person k

$$y_k = \beta_0 + \epsilon. \tag{5}$$

Der Parameter β bringt also den Einfluß des Faktors Berufstätigkeit auf die abhängige Variable zum Ausdruck. Man nennt ihn den **Haupteffekt** des unabhängigen Merkmals. In analoger Weise können unabhängige Variablen mit mehr als zwei Kategorien oder Interaktionswirkungen mehrerer unabhängiger Merkmale in den Ansatz aufgenommen werden. Für mathematisch-statistische Details vergleiche man die Ausführungen im Anhang.

Zur **Konstruktion der abhängigen Variablen** (Zielvariablen) betrachtet man das allgemeine Layout des Datenmaterials bei der kategorialen Regression. Durch die verschiedenen Kombinationen von Ausprägungen (Faktorstufen) der im Ansatz berücksichtigten unabhängigen Merkmale wird die Population in **I Teilgesamtheiten** zerlegt, wobei I die Anzahl aller möglichen Kombinationen bezeichnet. Das Modell der kategorialen Regression formuliert man nicht wie in (1) in den individuellen Beobachtungswerten, sondern in den relativen Häufigkeiten $\hat{\pi}_{ir}$ der Beobachtungswerte, die jeweils innerhalb einer Subpopulation i in die Kategorie r (r = 1, . . ., R) der Zielvariablen fallen. Sind n_{ir} die entsprechenden absoluten Häufigkeiten, so sieht die typische Anordnung der Daten folgendermaßen aus (vgl. *Tab. 1*).

Teilgesamtheit	Kategorie der Zielvariablen 1	...	R	Total
1	$n_{11}(\hat{\pi}_{11})$	...	$n_{1R}(\hat{\pi}_{1R})$	n_1
.	.		.	.
.	.		.	.
.	.		.	.
I	$n_{I1}(\hat{\pi}_{I1})$	...	$n_{IR}(\hat{\pi}_{IR})$	n_I

Tab. 1: Allgemeine Datenanordnung bei der kategorialen Regression

Die $\hat{\pi}_{ir} = n_{ir}/n_i$ sind konsistente Schätzungen der bedingten Wahrscheinlichkeiten π_{ir}, daß der Wert der Zielvariablen in die r-te Kategorie fällt, wenn eine Merkmalskombination $\mathbf{x}_i$ vorliegt (vgl. Abschnitt 5.1).

Mit

$$\hat{\boldsymbol{\pi}}_i' = (\hat{\pi}_{i1}, \ldots, \hat{\pi}_{i,R-1}) \quad \text{und} \quad \hat{\boldsymbol{\pi}}' = (\hat{\pi}_1', \ldots, \hat{\pi}_I') \tag{6}$$

$$\hat{\boldsymbol{\pi}} = \mathbf{X\beta} + \boldsymbol{\varepsilon} \tag{7}$$

bzw. in allgemeineren Ansätzen:

$$\mathbf{f}(\hat{\boldsymbol{\pi}}) = \mathbf{X\beta} + \boldsymbol{\varepsilon} \tag{8}$$

mit $\mathbf{f}(\hat{\boldsymbol{\pi}}) = (f_1(\hat{\boldsymbol{\pi}}), \ldots, f_N(\hat{\boldsymbol{\pi}}))'$, $N \leqslant I(R-1)$.

Das Modell der kategorialen Regression geht in wesentlichen Teilen auf *Grizzle/Starmer/Koch* (1969) zurück. Für eine ausführliche Darstellung vergleiche man beispielsweise *Hamerle/Kemény/Tutz* (1983).

Das Modell (7) besitzt den Vorteil der leichteren Interpretierbarkeit als (8), da die Regressionskoeffizienten wegen $\boldsymbol{\pi} = \mathbf{X\beta}$ direkt als prozentuale Anteile gedeutet werden

können, mit denen die verschiedenen Faktorstufen zu den bedingten Wahrscheinlichkeiten beitragen. Ein Nachteil besteht darin, daß die für eine weitere Beobachtung geschätzten Wahrscheinlichkeiten nicht notwendig zwischen Null und Eins liegen müssen. Nach unserer Erfahrung ist dies allerdings bei richtig spezifizierter Designmatrix praktisch nie der Fall. Der Vorteil von (8) besteht u.a. darin, daß dies durch geeignete Wahl von f auf jeden Fall vermieden werden kann. Besondere Bedeutung in Theorie und Praxis haben sog. **Logit-Ansätze**. Dabei wird z.B. für binäre Zielvariablen (R = 2)

$$f_i(\hat{\boldsymbol{\pi}}) = \log \frac{\hat{\pi}_{i1}}{1 - \hat{\pi}_{i1}}, \quad i = 1, \ldots, I \tag{9}$$

gesetzt.

Die unbekannten Parameter $\beta_0, \beta_1, \ldots, \beta_p$ des Modells werden aus den beobachteten Daten mit Hilfe einer gewichteten Kleinste-Quadrate-Methode geschätzt. Die inferenzstatistischen Ausführungen zum kategorialen Regressionsmodell sind im Anhang zusammengestellt.

3. Anwendung der kategorialen Regression

3.1. Das Datenmaterial

Im folgenden wird der bereits im ersten Abschnitt beschriebene Datensatz zu einer von der *Infratest Forschung GmbH* durchgeführten Fallstudie „Informationsgewohnheiten von Frauen" im Rahmen des kategorialen Regressionsmodells analysiert. Bei einem Teilaspekt, der die Präferenz für Zeitschrift Z zum Gegenstand hat, wurden als mögliche Einflußgrößen die Merkmale „Berufstätigkeit", „Alter" und „Schulbildung" gewählt, und zwar in der folgenden Kategorisierung:

Faktoren

(1) Berufstätigkeit (B)
- berufstätig (B1)
- nicht berufstätig (B2)

(2) Alter (A)
- 18–29 Jahre (A1)
- 30–39 Jahre (A2)
- 40–49 Jahre (A3)

(3) Schulbildung (S)
- Volksschule ohne Lehre (S1)
- Volksschule mit Lehre (S2)
- Realschule (S3)
- Abitur/Universität (S4)

Zielvariable

Präferenz für Zeitschrift Z
- ja
- nein

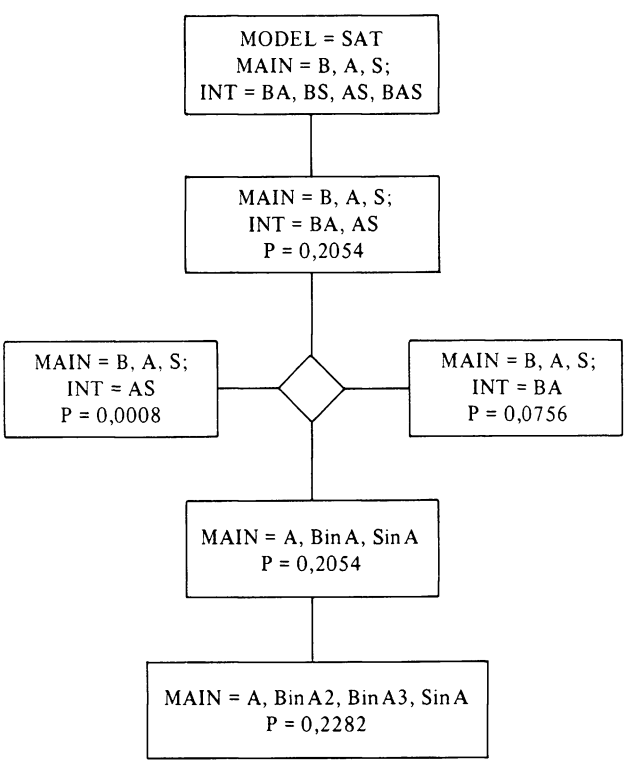
Bei einer Stichprobe von N = 941 befragten Frauen ergab sich entsprechend *Tab. 1* die folgende *Tab. 2.*

Berufstätigkeit	Alter	Schulbildung	Zeitschrift (Z) Z	nicht Z
berufstätig	18–29	Volksschule ohne Lehre	1	14
		Volksschule mit Lehre	32	49
		Realschule	20	34
		Abitur/Universität	8	3
	30–39	Volksschule ohne Lehre	9	23
		Volksschule mit Lehre	31	57
		Realschule	11	26
		Abitur/Universität	5	7
	40–49	Volksschule ohne Lehre	1	33
		Volksschule mit Lehre	12	50
		Realschule	5	11
		Abitur/Universität	1	7
nicht berufstätig	18–29	Volksschule ohne Lehre	3	24
		Volksschule mit Lehre	12	41
		Realschule	19	20
		Abitur/Universität	14	13
	30–39	Volksschule ohne Lehre	1	37
		Volksschule mit Lehre	12	68
		Realschule	14	43
		Abitur/Universität	4	7
	40–49	Volksschule ohne Lehre	11	54
		Volksschule mit Lehre	14	53
		Realschule	8	15
		Abitur/Universität	1	3

Tab. 2: Datenanordnung für die Fallstudie „Informationsgewohnheiten von Frauen"

3.2. Modellbildung und Interpretation

Ausgangspunkt der Analyse ist der Modellansatz (7) $\hat{\boldsymbol{\pi}} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, d.h. es soll untersucht werden, wie die drei genannten Faktoren die Präferenzrate (Leserate) für Zeitschrift Z beeinflussen. Es wurde versucht, ausgehend vom saturierten Modell durch Weglassen von nicht signifikan-



(MAIN bedeuten Haupteffekte, INT bedeuten Interaktionen)

ten Effekten ein möglichst einfaches, den Daten aber noch angepaßtes Modell zu finden. Sämtliche numerischen Berechnungen wurden mit dem **Programm NONMET II** von *Herbert M. Kritzer* durchgeführt. Die **Modellsuche** ist im folgenden **Flußdiagramm** wiedergegeben, wobei mit P die Überschreitungswahrscheinlichkeit des jeweiligen Werts der Teststatistik $Q_c(\beta)$ des Modelltests bezeichnet wird (vgl. Formel (A. 11) im Anhang).

Bei den letzten beiden Modellen handelt es sich um **konditionale Modelle** (vgl. Abschnitt 5.3). Insbesondere zeigt das **Modell <A,BinA2,BinA3,SinA>** mit einem Wert der Goodness-of-fit-Teststatistik von 12,92 bei zehn Freiheitsgraden (Überschreitungswahrscheinlichkeit P = 0,2282) eine gute Anpassung an den vorliegenden Datensatz. Es ergab sich:

Parameterschätzung	(geschätzte) Varianz	$Q_c(\hat{\beta}_i)$	Überschreitungswahrscheinlichkeit P
$\hat{\beta}_0$ = 0,271	0,000315	233,18	0,00
$\hat{\beta}_{A1}$ = 0,082	0,00057	11,81	0,00
$\hat{\beta}_{A2}$ = −0,009	0,00062	0,13	0,71
$\hat{\beta}_{BinA2}$ = 0,089	0,00048	16,28	0,00
$\hat{\beta}_{BinA3}$ = −0,045	0,00042	4,78	0,03
$\hat{\beta}_{S1inA1}$ = −0,263	0,00173	39,84	0,00
$\hat{\beta}_{S1inA2}$ = −0,140	0,00129	15,09	0,00
$\hat{\beta}_{S1inA3}$ = −0,110	0,00146	8,29	0,00
$\hat{\beta}_{S2inA1}$ = −0,038	0,00154	0,92	0,34
$\hat{\beta}_{S2inA2}$ = −0,014	0,00137	0,14	0,71
$\hat{\beta}_{S2inA3}$ = 0,003	0,00178	0,00	0,95
$\hat{\beta}_{S3inA1}$ = 0,064	0,00205	2,01	0,16
$\hat{\beta}_{S3inA2}$ = 0,026	0,00191	0,37	0,55
$\hat{\beta}_{S3inA3}$ = 0,128	0,00403	4,05	0,04

Insgesamt läßt sich das erhaltene Modell wie folgt **interpretieren**: Bei 27,1% aller Frauen ist eine Präferenz für Zeitschrift Z festzustellen ($\hat{\beta}_0$ = 0,271). Dabei spielt das Alter der befragten Frauen für die Lesegewohnheit die wichtigste Rolle.

In der Gruppe der 18–29jährigen Frauen lag das Interesse an Zeitschrift Z mit 35,3% ($\hat{\beta}_{A1}$ = 0,082) signifikant über der durchschnittlichen Leserate. Dieser positive Effekt in der jüngsten Alterskategorie wird zusätzlich verstärkt bei denjenigen Frauen, welche den höchsten Schulbildungsstand (Abitur/Universität) aufweisen ($\hat{\beta}_{S4\text{ in }A1}$ = 0,236). In dieser Gruppe beträgt die Leserate 58,9%.

Demgegenüber ist bei den 18–29jährigen Frauen mit dem niedrigsten Schulbildungsnachweis (Volksschule ohne Lehre) ein äußerst schwach ausgeprägtes Interesse an Zeitschrift Z festzustellen ($\hat{\beta}_{S1\text{ in }A1}$ = −0,263). Die Leserate dieser Frauengruppe sinkt auf 9,0%.

Ferner ist in der Gruppe der 40–49jährigen Frauen ein unterdurchschnittliches Leseinteresse an Zeitschrift Z festzustellen ($\hat{\beta}_{A3}$ = −0,073). Dieser negative Effekt in der dritten Alterskategorie wird allerdings bei denjenigen Frauen, welche Realschulbildung aufweisen, wieder aufgehoben ($\hat{\beta}_{S2\text{ in }A3}$ = 0,128). In dieser Gruppe beträgt der Anteil der Leserinnen 32,6%.

Eine ähnliche Feststellung gilt für die nicht berufstätigen 40–49jährigen Frauen. 24,3% dieser Gruppe zeigen eine Präferenz für Zeitschrift Z und liegen damit über dem Durchschnitt dieser Altersgruppe ($\hat{\beta}_{B2\text{ in }A3}$ = 0,045). Demgegenüber sinkt der Anteil der berufstätigen Leserinnen dieser Altersgruppe auf 15,3% ($\hat{\beta}_{B1\text{ in }A3}$ = −0,045).

Bei den 30–39jährigen Frauen insgesamt ist im Gegensatz zu den beiden anderen Altersgruppen keine signifikante Abweichung von der durchschnittlichen Präferenzrate festzustellen. Allerdings spielt die Unterscheidung nach „berufstätigen" und „nicht berufstätigen" Frauen eine diesbezügliche Rolle.

Während bei den erstgenannten die Lesequote um 8,9% ($\hat{\beta}_{B1\text{ in }A2}$ = 0,089) über dem Durchschnitt liegt, ist sie bei den nicht berufstätigen Frauen dieser Altersgruppe unterdurchschnittlich ($\hat{\beta}_{B2\text{ in }A2}$ = −0,089).

Völlig ohne Bedeutung für das Lesebedürfnis der befragten Frauen war der Schulbildungsstand „Volksschule mit Lehre".

4. Marktsegmentierung

4.1. Marktsegmentierung mit Hilfe von AID

Die im Rahmen des Marktsegmentierungsproblems auftretenden statistischen Fragestellungen wurden im Falle kategorialer Segmentierungs- und Kriteriumsmerkmale überwiegend mit Hilfe der Kontrastgruppenanalyse (AID) untersucht. Durch diesen Segmentierungsvorgang sollen bei der Segmentbildung diejenigen Merkmale identifiziert werden, die einen bedeutsamen Einfluß auf die Kriteriumsvariable, wie etwa hier „Präferenz für Zeitschrift Z", haben. Bei der von *Infratest* durchgeführten AID-Studie ergab sich der in *Abb. 1* dargestellte Segmentationsbaum.

4.2. Marktsegmentierung durch kategoriale Regression

Das in Abschnitt 2 vorgestellte kategoriale Regressionsmodell bietet gegenüber der AID den Vorteil, sich durch geeignete statistische Tests dagegen abzusichern, daß die gefundenen Marktsegmente reine Zufallsprodukte sind. Es werden nämlich zur Segmentbildung nur die aufgrund der Teststatistik $Q_c(\hat{\beta}_i)$ (vgl. (A. 14)) als signifikant von Null verschieden erkannten Regressionskoeffizienten herangezogen. Diese Regressionskoeffizienten sind aufgrund der Effektkodierung analog zur Varianzanalyse als Mittelwerte (d.h. hier durchschnittliche Prozentsätze) bzw. als sukzessive Abweichungen von Mittelwerten zu interpretieren. Beispielsweise ist $\hat{\beta}_0$ = 0,271 (= 27,1%) die geschätzte durchschnittliche Leserate in der Gesamtpopulation. *Abb. 2* enthält die aus dem vorliegenden Datenmaterial nach diesem Modell resultierenden absatzpolitisch relevanten Marktsegmente, definiert als Frauengruppen mit signifikanter Abweichung von der durchschnittlichen Leserate von 27,1%.

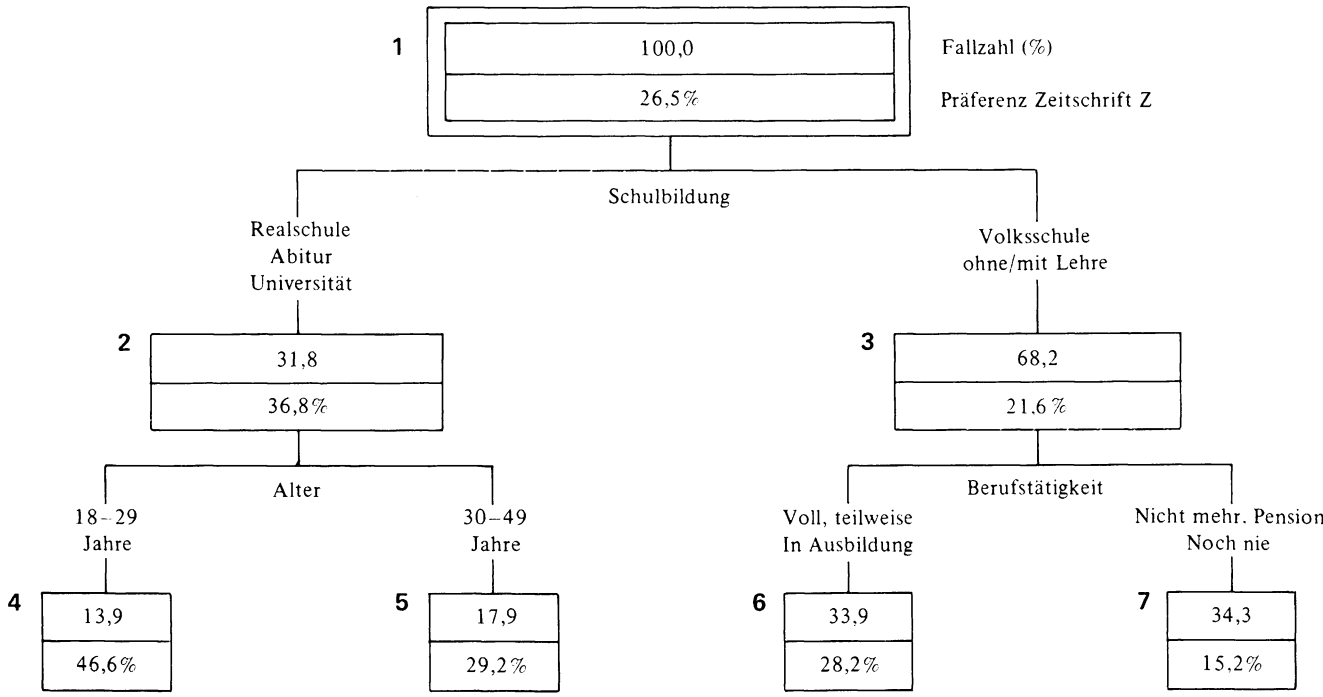
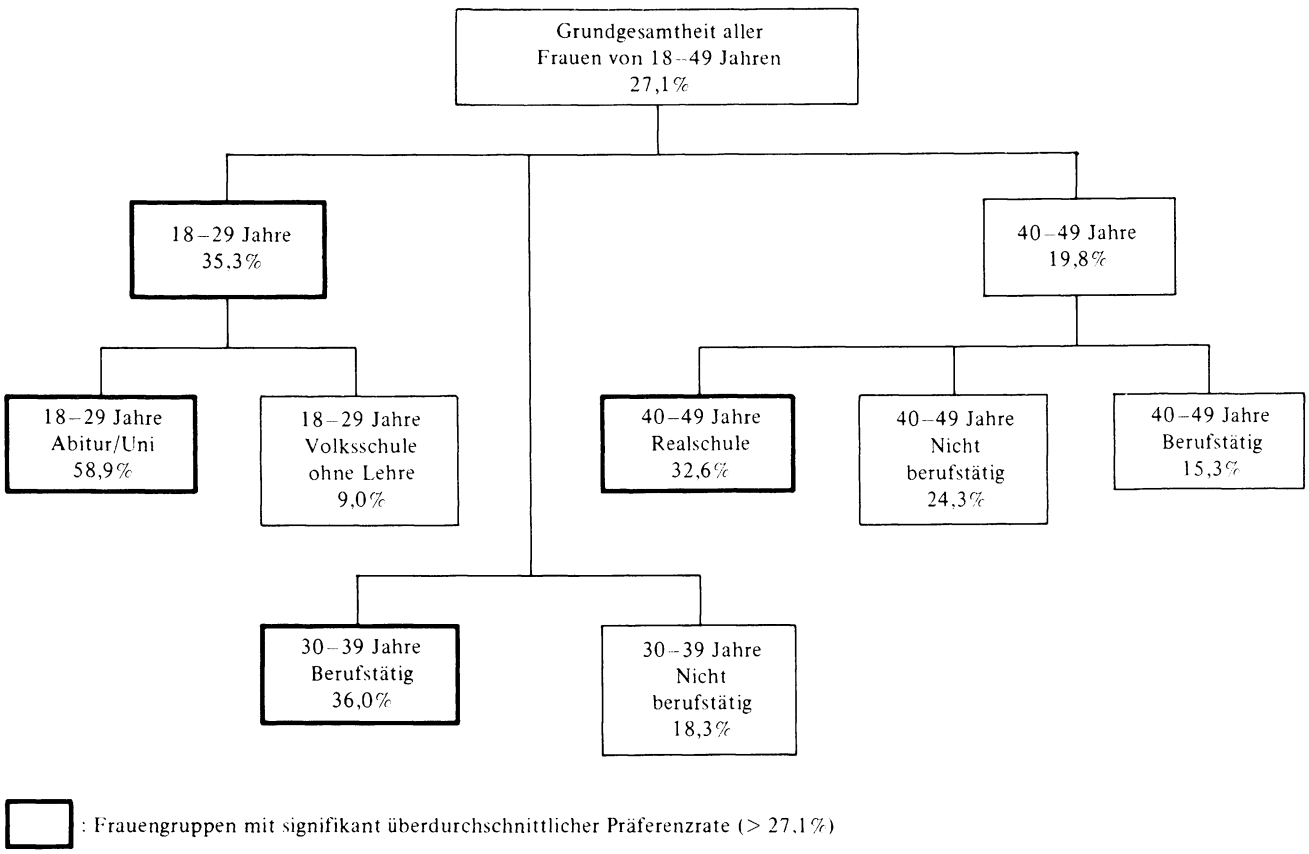


Abb. 1: Marktsegmentierung des Datensatzes durch AID



: Frauengruppen mit signifikant überdurchschnittlicher Präferenzrate (> 27,1%)

Abb. 2: Graphische Darstellung der Frauengruppen mit signifikanter Abweichung von der durchschnittlichen Präferenzrate (27,1%)

4.3. Vergleich

Der Vergleich der beiden Tabellen zeigt einige **substantielle Unterschiede** im Hinblick auf die resultierenden Marktsegmente. Da die AID auf binäre Splits beschränkt ist, erhält man bei der Segmentierung gelegentlich undifferenzierte, durch zusammengelegte Kategorien der unabhängigen Variablen definierte Marktsegmente. Beispielsweise ergibt sich für das Marktsegment 4 in *Abb. 1* eine geschätzte Leserate von 46,6%, während im entsprechenden Marktsegment in *Abb. 2*, das durch die signifikanten Regressionskoeffizienten $\beta_{A1} = 0{,}082$ und $\beta_{S4 \text{ in } A1} = 0{,}236$ festgelegt ist, eine geschätzte Leserate von 58,9% resultiert. Ferner können durch die erforderliche Zusammenlegung von Kategorien und durch die in der Segmentierungsvorschrift enthaltene Varianzmaximierung zwischen den Gruppen signifikante Interaktionen übersehen und andererseits nicht signifikante Interaktionen auf einer bestimmten Segmentationsstufe künstlich erzeugt werden. Man vergleiche etwa die durch AID erhaltenen Segmente 6 und 7 in *Abb. 1*, die auf eine Interaktion zwischen Berufstätigkeit und Schulbildung hindeuten, die jedoch im kategorialen Regressionsmodell nicht signifikant ist.

Abschließend sei bemerkt, daß die in Abschnitt 2 vorgestellte Methode der kategorialen Regression die von *Green* et al. (1977), *Flath/Leonard* (1979) u.a. behandelten Modelle zur Marktsegmentierung als Spezialfälle enthält. Darüber hinaus ist das Modell der kategorialen Regression ein geeignetes Instrumentarium zur Marktsegmentierung, das auch in theoretischer Hinsicht der Kontrastgruppenanalyse überlegen ist.

Anhang

1. Kodierung der unabhängigen Merkmale

Besitzt ein unabhängiges Merkmal A I Kategorien (Faktorstufen), so lassen sich diese durch I – 1 Dummy-Variablen erfassen, z.B. in einer **„Dummykodierung"** der Form:

$$x_i^A = \begin{cases} 1 & \text{falls Kategorie i der Variablen A vorliegt,} \\ 0 & \text{sonst} \end{cases} \qquad i = 1, \ldots, I-1. \tag{A.1}$$

Die i-te Variable x_i^A (i=1, . . .,I–1) kodiert dabei nur das Vorliegen bzw. Nichtvorliegen der i-ten Ausprägung des Merkmals A. Das Vorliegen der I-ten (Referenz-)Kategorie ist implizit erfaßt durch die Kodierung $x_i^A = 0$ für i=1, . . .,I–1. Die zu den Variablen x_i^A gehörenden Koeffizienten β_i werden wie in der Varianzanalyse Haupteffekte genannt.

Eine unmittelbar an die Varianzanalyse angelehnte Darstellung ergibt sich durch die **„Effektkodierung"**. Die Kodierung erfolgt hier mit den I–1 Variablen

$$x_i^A = \begin{cases} 1 & \text{falls Kategorie i der Variablen A vorliegt,} \\ -1 & \text{falls Kategorie I der Variablen A vorliegt,} \\ 0 & \text{sonst} \end{cases} \qquad i = 1, \ldots, I-1. \tag{A.2}$$

Die Effektkodierung (A.2) ist eine unmittelbare Konsequenz der in der Varianzanalyse üblichen Restriktionen für die Parameter. Dort wird die Summe der Haupteffekte einer Variablen A a priori gleich Null gesetzt. Daraus ergibt sich für den Parameter der Referenzkategorie

$$\beta_I = -\sum_{i=1}^{I-1} \beta_i .$$

Der Parameter β_I wird nicht in den Regressionsansatz einbezogen, sondern durch die restlichen ausgedrückt, wobei für eine Beobachtung aus der Referenzkategorie in der **X**-Matrix bei den zu $\beta_1, \ldots, \beta_{I-1}$ gehörenden Stellen jeweils eine –1 zu setzen ist.

Neben den Haupteffekten können wie in der Varianzanalyse auch **Interaktionseffekte** in den Ansatz aufgenommen werden. Sie messen den gemeinsamen Einfluß einer bestimmten Kombination von Kategorien von zwei und mehr unabhängigen Variablen. Formal werden sie durch Produkte der Dummy-Variablen für die Haupteffekte miteinbezogen. Der Datenvektor **x** wird dann erweitert um Zwei-Faktor-Interaktionen, wie z.B. $x_i^A x_j^B$, bzw. Drei-Faktor-Interaktionen, wie z.B. $x_i^A x_j^B x_k^C$, etc.

2. Kategoriales Regressionsmodell, Parameterschätzung und Teststatistiken

Das kategoriale Regressionsmodell ist gegeben durch (vgl. (7) und (8))

$$\hat{\boldsymbol{\pi}} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{A.3}$$

mit $\hat{\boldsymbol{\pi}}_i = (\hat{\pi}_{i1}, \ldots, \hat{\pi}_{i,R-1})$ und $\hat{\boldsymbol{\pi}}' = (\hat{\boldsymbol{\pi}}_1', \ldots, \hat{\boldsymbol{\pi}}_I')$
bzw. im allgemeineren Ansatz

$$\mathbf{f}(\hat{\boldsymbol{\pi}}) = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{A.4}$$

mit $\mathbf{f}(\hat{\boldsymbol{\pi}}) = (f_1(\hat{\boldsymbol{\pi}}), \ldots, f_N(\hat{\boldsymbol{\pi}}))'$, $N \leqslant I(R-1)$.

Zur Schätzung der unbekannten Parameter $\beta_0, \beta_1, \ldots, \beta_p$ benötigt man die Kovarianzmatrix der Fehlervariablen. Beim Modell (A.3) ergibt sich die Blockdiagonalmatrix

$$\operatorname{cov}(\boldsymbol{\varepsilon}) = \operatorname{cov}(\hat{\boldsymbol{\pi}}) = \begin{pmatrix} \operatorname{cov}(\hat{\boldsymbol{\pi}}_1) & & 0 \\ & \ddots & \\ 0 & & \operatorname{cov}(\hat{\boldsymbol{\pi}}_I) \end{pmatrix} \tag{A.5}$$

mit

$$\operatorname{cov}(\hat{\boldsymbol{\pi}}_i) = \frac{1}{n_i} \begin{pmatrix} \pi_{i1}(1-\pi_{i1}) & \cdots & -\pi_{i1}\pi_{i,R-1} \\ & \ddots & \\ & \cdots & \pi_{i,R-1}(1-\pi_{i,R-1}) \end{pmatrix} \tag{A.6}$$

Im allgemeinen Ansatz (A.4) ist

$$\operatorname{cov}(\boldsymbol{\varepsilon}) = \mathbf{V} = \mathbf{H} \operatorname{cov}(\hat{\boldsymbol{\pi}}) \mathbf{H}' \tag{A.7}$$

mit

$$\mathbf{H} = \left(\frac{\partial f_n(\hat{\pi})}{\partial \pi_{ir}} \right) \tag{A.8}$$

$n = 1, \ldots, N;\ i = 1, \ldots, I;\ r = 1, \ldots, R-1$

die asymptotische Kovarianzmatrix von $\mathbf{f}(\hat{\boldsymbol{\pi}})$ bzw. $\boldsymbol{\varepsilon}$. Schätzwerte $\hat{\boldsymbol{\beta}}$ für $\boldsymbol{\beta}$ erhält man aus einem **verallgemeinerten Kleinst-Quadrat-Prinzip**

$$Q_e(\boldsymbol{\beta}) = (\mathbf{f}(\hat{\boldsymbol{\pi}}) - \mathbf{X}\boldsymbol{\beta})' \hat{\mathbf{V}}^{-1} (\mathbf{f}(\hat{\boldsymbol{\pi}}) - \mathbf{X}\boldsymbol{\beta}) \rightarrow \min_{\boldsymbol{\beta}}! \tag{A.9}$$

Dabei werden in $\hat{\mathbf{V}}$ die unbekannten Parameter π_{ir} durch die konsistenten Schätzungen $\hat{\pi}_{ir}$ ersetzt. Gilt $\operatorname{rg}(\mathbf{X}) = p + 1 \leqslant I(R-1)$, erhält man

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}' \hat{\mathbf{V}}^{-1} \mathbf{X})^{-1} \mathbf{X}' \hat{\mathbf{V}}^{-1} \mathbf{f}(\hat{\boldsymbol{\pi}}) . \tag{A.10}$$

Ein grundlegendes Problem besteht in der **Spezifikation der Designmatrix X**, d.h. man sucht ein möglichst einfaches Modell für das gemeinsame Einwirken der unabhängigen Merkmale (Faktoren), welches das vorliegende Datenmaterial angemessen beschreibt. Dazu läßt sich mit Hilfe der gewichteten Quadratsumme der Residuen

$$Q_e(\hat{\boldsymbol{\beta}}) = (\mathbf{f}(\hat{\boldsymbol{\pi}}) - \mathbf{X}\hat{\boldsymbol{\beta}})' \hat{\mathbf{V}}^{-1} (\mathbf{f}(\hat{\boldsymbol{\pi}}) - \mathbf{X}\hat{\boldsymbol{\beta}}) \tag{A.11}$$

ein Anpassungstest konstruieren. Bei Gültigkeit des Modells $\mathbf{f}(\boldsymbol{\pi}) = \mathbf{X}\boldsymbol{\beta}$ ist $Q_e(\hat{\beta})$ asymptotisch χ^2-verteilt mit $N - p - 1$ Freiheitsgraden. Für eine Ableitung der asymptotischen Verteilung von $Q_e(\boldsymbol{\beta})$ siehe *Hamerle/Kemény/Tutz* (1983).

Im Ansatz der kategorialen Regression ist es möglich, durch Einbeziehung sämtlicher Interaktionswirkungen ein Modell zu erhalten, bei dem die Designmatrix quadratisch und nichtsingulär ist. Man nennt solche Modelle **saturiert.**

Seien beispielsweise zwei dichotome Faktoren und eine dichotome Zielvariable gegeben. Bezieht man neben den Haupteffekten der beiden Faktoren auch die Interaktionswirkung AB ein, erhält man das saturierte Modell (in Effektkodierung)

$$\begin{pmatrix}\hat{\pi}_1\\ \hat{\pi}_2\\ \hat{\pi}_3\\ \hat{\pi}_4\end{pmatrix} = \begin{pmatrix}1 & 1 & 1 & 1\\ 1 & 1 & -1 & -1\\ 1 & -1 & 1 & -1\\ 1 & -1 & -1 & 1\end{pmatrix}\begin{pmatrix}\beta_0\\ \beta_A\\ \beta_B\\ \beta_{AB}\end{pmatrix} + \begin{pmatrix}\epsilon_1\\ \epsilon_2\\ \epsilon_3\\ \epsilon_4\end{pmatrix} \qquad (A.12)$$

Das Ziel der kategorialen Regression besteht jedoch darin, den empirischen Sachverhalt möglichst einfach zu beschreiben, d.h. ein Modell mit möglichst wenigen Parametern zu finden. Demnach sind unsaturierte Modelle von zentraler Bedeutung, wobei allerdings darauf zu achten ist, daß sie den empirischen Sachverhalt noch angemessen beschreiben. Zur Überprüfung dient die Teststatistik (A.11).

Hat man ein passendes Modell gefunden, kann man verschiedene Hypothesen bezüglich der im Modell enthaltenen Parameter statistisch überprüfen. Diese Tests betreffen in der Regel Teile des Modells, insbesondere den Beitrag einzelner Regressionskoeffizienten. Die entsprechende Nullhypothese läßt sich stets schreiben als allgemeine lineare Hypothese

$$\mathbf{C}\boldsymbol{\beta} = \mathbf{0} \qquad (A.13)$$

mit einer geeigneten Matrix C, $rg(\mathbf{C}) = k$.
Eine geeignete Teststatistik zur Überprüfung von (A.13) ist

$$Q_C(\boldsymbol{\beta}) = (\mathbf{C}\boldsymbol{\beta})' \left[\mathbf{C}(\mathbf{X}' \hat{\mathbf{V}}^{-1} \mathbf{X})^{-1} \mathbf{C}'\right]^{-1} (\mathbf{C}\boldsymbol{\beta}) . \qquad (A.14)$$

$Q_C(\boldsymbol{\beta})$ ist bei Gültigkeit der Nullhypothese asymptotisch χ^2-verteilt mit k Freiheitsgraden. Zum Beweis vergleiche man wieder *Hamerle/Kemény/Tutz* (1983).

3. Konditionale Modelle

Eine spezielle Möglichkeit, Interaktionswirkungen in leichter interpretierbarer Weise zu behandeln, bieten sog. **„konditionale" Modelle.** Für eine ausführliche Beschreibung von konditionalen Modellen vergleiche man *Küchler* (1979), *Forthofer/Lehnen* (1981) oder *Hamerle/Kemény/Tutz* (1983). Bei der Konstruktion von konditionalen Modellen werden **geschachtelte Effekte** ermittelt. Man wählt z.B. eine unabhängige Variable als Bezugsfaktor aus und berechnet dann Haupt- und Interaktionseffekte der übrigen unabhängigen Merkmale getrennt für jede Kategorie der ausgewählten Variablen.

Seien beispielsweise die dichotomen Faktoren A, B und eine dichotome Zielvariable gegeben. Wählt man nun Faktor A, dessen Haupteffekt unverändert berechnet wird, als Bezugsfaktor aus, so lassen sich die konditionalen Effekte

β_{BinA1} und β_{BinA2}

definieren und das saturierte konditionale Modell ist gegeben durch (Effektkodierung; Modell (A.3))

$$\begin{pmatrix}\hat{\pi}_1\\ \hat{\pi}_2\\ \hat{\pi}_3\\ \hat{\pi}_4\end{pmatrix} = \begin{pmatrix}1 & 1 & 1 & 0\\ 1 & 1 & -1 & 0\\ 1 & -1 & 0 & 1\\ 1 & -1 & 0 & -1\end{pmatrix}\begin{pmatrix}\beta_0\\ \beta_A\\ \beta_{BinA1}\\ \beta_{BinA2}\end{pmatrix} + \epsilon \qquad (A.15)$$

Es werden also für jede Teilgesamtheit, die durch die Stufen des ausgewählten Bezugsfaktors festgelegt sind, getrennt die Effekte des anderen Faktors berechnet.

Vom statistischen Standpunkt aus ist ein konditionales Modell zum entsprechenden nicht konditionalen Modell äquivalent, d.h. die Teststatistiken $Q_e(\boldsymbol{\beta})$ des Modelltests sind identisch. Konditionale Modelle liefern lediglich eine zusätzliche Möglichkeit, vorhandene Interaktionswirkungen zu analysieren und zu interpretieren.

Anmerkung

[1] Wir danken Herrn Dr. *Stadler* von der *Infratest Forschung GmbH*, München, für die Überlassung des Datenmaterials.

Literaturverzeichnis

Armstrong, J.S.; Andress, J.G. (1970): Exploratory analysis of marketing data: trees vs. regression, in: Journal of Marketing Research, Vol. 7 (1970), S. 487–492.

Assael, H. (1970): Segmenting markets by group purchasing behaviour: an application of the AID technique, in: Journal of Marketing Research, Vol. 7 (1970), S. 153–158.

Bhapkar, V.P. (1961): Some tests for categorical data, in: Annals of Mathematical Statistics, Vol. 32 (1961), S. 72–81.

Bhapkar, V.P. (1966): A note on the equivalence of two test criteria for hypotheses in categorical data, in: Journal of the American Statistical Association, Vol. 61 (1966), S. 228–235.

Böcker, F.; Thomas, L. (1981): Marketing, Stuttgart 1981.

Böhler, H. (1977): Methoden und Modelle der Marktsegmentierung, Stuttgart 1977.

Carman, J.M. (1970): Correlates of brand loyalty: some positive results, in: Journal of Marketing Research, Vol. 7 (1970), S. 67–76.

Cramer, E. (1971): Book review, in: Psychometrika, Vol. 36 (1971), S. 440–442.

DeSarbo, W.S.; Hildebrandt, D.K. (1980): A marketer's guide to log-linear models for qualitative data analysis, in: Journal of Marketing, Vol. 44 (1980), S. 40–51.

Dillon, W.R. (1979): The performance of the linear discriminant function in nonoptimal situations and the estimation of classification error rates: a review of recent finding, in: Journal of Marketing Research, Vol. 16 (1979), S. 370 ff.

Doyle, P. (1973): The use of Automatic Interaction Detector and similar search procedures, in: Operations Research Quarterly 24, S. 465–467.

Doyle, P.; Fenwick, I. (1975): The pitfalls of AID analysis, in: Journal of Marketing Research, Vol. 12 (1975), S. 408–413.

Einhorn, H. (1972): Alchemy in the behavioral sciences, in: Public Opinion Quarterly, Vol. 36 (1972), S. 367–378.

Fahrmeir, L.; Hamerle, A. (1981): Kategoriale Regression in der betrieblichen Planung, in: Zeitschrift für Operations Research, Vol. 25B, S. 63–78.

Fielding, A. (1975): Binary segmentation: The Automatic Interaction Detector and related techniques for exploring data structure, in: *O'Muircheartaigh, C.A.; Payne, C.* (Eds.), Exploring data structures, New York 1975.

Flath, D.; Leonard, E.W. (1979): A comparison of two logit models in the analysis of qualitative marketing data, in: Journal of Marketing Research, Vol. 16 (1979), S. 533–538.

Forthofer, R.N.; Lehnen, R.G. (1981): Public program analysis, a new categorical data approach, Belmont, Cal. 1981.

Green, P.E.; Carmone, F.J.; Wachspress, D.P. (1977): On the analysis of qualitative data in marketing research, in: Journal of Marketing Research, Vol. 14 (1977), S. 52–59.

Green, P.E. (1978): An AID/Logit procedure for analyzing large multiway contingency tables, in: Journal of Marketing Research, Vol. 15 (1978), S. 132–136.

Grizzle, J.E.; Starmer, C.F., Koch, G.G. (1969): Analysis of categorical data by linear models, in: Biometrics, Vol. 25 (1969), S. 489–504.

Grizzle, J.E.; Williams, O.D. (1972): Log-linear models and tests of independence for contingency tables, in: Biometrics, Vol. 28 (1972), S. 137–156.

Hamerle, A.; Kemény, P.; Tutz, G. (1983): Kategoriale Regression, in: *Fahrmeir, L.; Hamerle, A.* (Hrsg.): Multivariate statistische Verfahren, Berlin 1983, Kap. 6.

Heald, G.I. (1972): The application of the AID program and multiple regression techniques to the assessment of store performance and site selection, in: Operations Research Quarterly, Vol. 23 (1972), S. 445–457.

Johnson, W.D.; Koch, G.G. (1971): A note on the weighted least squares analysis of the Ries-Smith contingency table data, in: Technometrics, Vol. 13 (1971), S. 438–447.

Kass, G.V. (1975): Significance testing in Automatic Interaction Detection (AID), in: Applied Statistics, Vol. 24 (1975), S. 178–189.

Kritzer, H.M. (1978): An introduction to multivariate contingency table analysis, in: American Journal of Political Science, Vol. 22 (1978), S. 187–226.

Kritzer, H.M. (1979): Analyzing contingency tables by weighted least squares: an alternative to the Goodman approach, in: Political Methodology, Vol. 6 (1979), S. 277–326.

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

Mayntz, R.; Holm, K.; Hübner, P. (1974): Einführung in die Methoden der empirischen Soziologie, Opladen 1974.

Morgan, J.N.; Sonquist, J.A. (1963): Problems in the analysis of survey data and a proposal, in: Journal of the American Statistical Association, Vol. 58 (1983), S. 415–434.

Neyman, J. (1949): Contributions to the theory of the chi-square-test. Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability, Berkeley 1949.

Perreault, W.D.; Barksdale, H.C. (1980): A model-free approach for analysis of complex contingency data in survey research, in: Journal of Marketing Research, Vol. 17 (1980), S. 503–515.

Perreault, W.D.; Young, F.W. (1980): Alternating Least Squares Optimal Scaling: Analysis of nonmetric data in marketing research, in: Journal of Marketing Research, Vol. 17 (1980), S. 1–13.

Schwedler, E. (1982): Neuere statistische Verfahren zur Analyse qualitativer Daten und ihre Anwendungsmöglichkeit in der Marketingforschung, in: Marketing ZFP, 4. Jg. (1982), S. 45–52.

Scott, A.J.; Knott, M. (1976): An approximate test for use with AID, in: Applied Statistics, Vol. 25 (1976), S. 103–106.

Sonquist, J.A.; Morgan, J.A. (1964): The detection of interaction effects. Monograph No. 35, Survey Research Centre, Institute for Social Research, University of Michigan 1964.

Sonquist, J.A.; Baker, E.L.; Morgan, J.N. (1971): Searching for structure, Michigan 1971.

Zusammenfassung

Im vorliegenden Beitrag wird gezeigt, daß sowohl aus formalstatistischer Sicht als auch im Hinblick auf die daraus sich ergebende substantielle Relevanz das Modell der kategorialen Regression ein leistungsfähigeres Instrument zur Identifizierung von Marktsegmenten ist als die bislang dafür überwiegend eingesetzte Kontrastgruppenanalyse (AID). Für eine empirische Fallstudie werden ein passendes kategoriales Regressionsmodell durch eine schrittweise Suchstrategie konstruiert und die daraus sich ergebende Marktsegmentierung mit den entsprechenden Ergebnissen einer früher durchgeführten Kontrastgruppenanalyse verglichen. Dabei erhält man einige substantielle Unterschiede, insbesondere im Hinblick auf die Identifikation absatzpolitisch relevanter Marktsegmente.

Summary

The main purpose of the paper is to show that both from a statistical point of view as well as regarding substantial conclusions the model of categorical regression is a more powerful method for identifying relevant market segments than AID (Automatic Interaction Detector) which is still often used in this context. An empirical illustration is shown by fitting a categorical regression model in a stepwise manner to the data collected in a case study in 'Female reading habits'. The resulting market segments are compared with those obtained by the application of the AID technique. Substantial differences are found to be primarily due to the lack of a possibility of testing significance in the AID framework.