Sonderdruck aus:

Mitteilungen
aus der
Arbeitsmarkt- und
Berufsforschung

Franz Egle

Regressionsschätzung mit qualitativen Variablen

8. Jg./1975

1

Mitteilungen aus der Arbeitsmarkt- und Berufsforschung (MittAB)

Die MittAB verstehen sich als Forum der Arbeitsmarkt- und Berufsforschung. Es werden Arbeiten aus all den Wissenschaftsdisziplinen veröffentlicht, die sich mit den Themen Arbeit, Arbeitsmarkt, Beruf und Qualifikation befassen. Die Veröffentlichungen in dieser Zeitschrift sollen methodisch, theoretisch und insbesondere auch empirisch zum Erkenntnisgewinn sowie zur Beratung von Öffentlichkeit und Politik beitragen. Etwa einmal jährlich erscheint ein „Schwerpunktheft", bei dem Herausgeber und Redaktion zu einem ausgewählten Themenbereich gezielt Beiträge akquirieren.

Hinweise für Autorinnen und Autoren

Das Manuskript ist in dreifacher Ausfertigung an die federführende Herausgeberin
Frau Prof. Jutta Allmendinger, Ph. D.
Institut für Arbeitsmarkt- und Berufsforschung
90478 Nürnberg, Regensburger Straße 104
zu senden.

Die Manuskripte können in deutscher oder englischer Sprache eingereicht werden, sie werden durch mindestens zwei Referees begutachtet und dürfen nicht bereits an anderer Stelle veröffentlicht oder zur Veröffentlichung vorgesehen sein.

Autorenhinweise und Angaben zur formalen Gestaltung der Manuskripte können im Internet abgerufen werden unter http://doku.iab.de/mittab/hinweise_mittab.pdf. Im IAB kann ein entsprechendes Merkblatt angefordert werden (Tel.: 09 11/1 79 30 23, Fax: 09 11/1 79 59 99; E-Mail: ursula.wagner@iab.de).

Herstellung

Satz und Druck: Tümmels Buchdruckerei und Verlag GmbH, Gundelfinger Straße 20, 90451 Nürnberg

Verlag

W. Kohlhammer GmbH, Postanschrift: 70549 Stuttgart: Lieferanschrift: Heßbrühlstraße 69, 70565 Stuttgart: Telefon 07 11/78 63-0;
Telefax 07 11/78 63-84 30: E-Mail: waltraud.metzger@kohlhammer.de, Postscheckkonto Stuttgart 163 30.
Girokonto Städtische Girokasse Stuttgart 2 022 309.
ISSN 0340-3254

Bezugsbedingungen

Die „Mitteilungen aus der Arbeitsmarkt- und Berufsforschung" erscheinen viermal jährlich. Bezugspreis: Jahresabonnement 52,- €
inklusive Versandkosten: Einzelheft 14,- € zuzüglich Versandkosten. Für Studenten, Wehr- und Ersatzdienstleistende wird der Preis
um 20 % ermäßigt. Bestellungen durch den Buchhandel oder direkt beim Verlag. Abbestellungen sind nur bis 3 Monate vor Jahresende möglich.

Zitierweise:

MittAB = „Mitteilungen aus der Arbeitsmarkt- und Berufsforschung" (ab 1970)
Mitt(IAB) = „Mitteilungen" (1968 und 1969)
In den Jahren 1968 und 1969 erschienen die „Mitteilungen aus der Arbeitsmarkt- und Berufsforschung" unter dem Titel
„Mitteilungen", herausgegeben vom Institut für Arbeitsmarkt- und Berufsforschung der Bundesanstalt für Arbeit.

Internet: http://www.iab.de

Regressionsschätzung mit qualitativen Variablen

(Darstellung methodischer Probleme und Lösungsansätze am Beispiel einer Untersuchung zur Berufswahlsituation von Jugendlichen)

*Franz Egle**

Während zur Analyse quantitativer, metrisch skalierter Merkmale vergleichsweise viele und weitentwickelte statistische Instrumente zur Verfügung stehen, ist man bei der Untersuchung qualitativer, nominal skalierter Merkmale auf wenige Analysenmethoden beschränkt. Untersuchungen dieser Art bilden aber den Schwerpunkt in wirtschafts- und sozialwissenschaftlichen Einrichtungen wie etwa dem IAB.

Mit dem folgenden Beitrag wird deshalb versucht, die üblicherweise nur bei der Analyse quantitativer Merkmale verwendete multiple Regressionsanalyse auf qualitative Merkmale anzuwenden. Es werden die dabei auftretenden mathematisch-statistischen Probleme und die sich daraus ergebenden Besonderheiten bei der Ergebnis-Interpretation erörtert.

Die verschiedenen Regressionsansätze werden am Beispiel einer Untersuchung der Berufswahlsituation von Jugendlichen diskutiert. Dabei wird eine der Fragen zur Art der gewünschten beruflichen Tätigkeit (und zwar die Problemlösungsbereitschaft) herangezogen und ihre Abhängigkeit von den qualitativen, nominal skalierten Merkmalen „Geschlecht“ und „Schulbildung“ analysiert.

In der Regel setzt die Regressionsanalyse metrisch skalierte Merkmale voraus. Mit Hilfe von „Dummy-Variablen“ gelingt jedoch auch eine Anwendung auf nominal skalierte Merkmale. Dabei treten im wesentlichen drei Probleme auf:

1. Lineare Abhängigkeit der Spalten der Beobachtungsmatrix (Multikollinearität)
2. Ungleiche Varianz der Störvariablen (Heteroskedastizität)
3. Linearität der Funktionsform

Zur Lösung des Multikollinearitätsproblems werden zwei Methoden erörtert: zum einen die schätzbaren Funktionen — hierbei werden anstatt der ursprünglichen Regressionskoeffizienten gewichtete Parameter geschätzt — und zum anderen die a-priori-Restriktionen von Parameterwerten, bei denen im einfachsten Fall so viele Spalten der Beobachtungsmatrix gestrichen werden, bis deren Rang gleich der Spaltenzahl ist.

Während die zweite Lösung eine gewisse Willkür enthält und man je nach der Auswahl der a priori festzulegenden Parameter unterschiedliche Schätzwerte erhalten kann, bekommt man mit der ersten Methode eindeutige Schätzwerte, falls es solche überhaupt gibt.

Die Heteroskedastizität der Varianz der Störvariablen läßt sich mittels der von Aitken entwickelten verallgemeinerten Methode der kleinsten Quadrate in die Analyse einbeziehen. Als Alternative zur linearen bewährt sich eine logistische Funktionsform.

Zugunsten der Schnelligkeit und Einfachheit der Berechnungen kann für hinreichend genaue Ergebnisse die Heteroskedastizität außer acht gelassen, auf logistische Funktionsformen verzichtet und das Multikollinearitätsproblem durch a-priori-Restriktionen von Parameterwerten gelöst werden.

Die Untersuchung wurde im IAB durchgeführt.

Gliederung

* Die notwendigen Computerprogramme wurden von G. Apfelthaler geschrieben.

1. Zielsetzung und Möglichkeiten

Die herkömmliche Aufbereitung statistischer Daten in Tabellenform stößt bei der Analyse sehr schnell an Grenzen: ein- und zwei- oder auch dreidimensionale Tabellen lassen oft Zusammenhänge nicht erkennen, höherdimensionale Tabellen sind meistens unübersichtlich. Eine Kausalanalyse über die Art der linearen Abhängigkeiten zwischen Variablen ist nicht möglich. Dazu ist ein multivariates Verfahren nötig, d. h. ein Verfahren, das es gestattet, mehrere Variablen in der statistischen Analyse gleichzeitig zu betrachten und gemeinsam auszuwerten. Die multiple Regressionsanalyse ist ein solches Verfahren. Jedoch werden hierbei in der Regel meßbare Größen unterstellt. Der vorliegende Aufsatz hat das Ziel, zu prüfen, unter welchen Bedingungen eine Anwendung der Regressionsanalyse auch auf Daten möglich ist, die ausschließlich auf qualitativen Merkmalen basieren. Diese Prüfung erfolgte beispielhaft an einem durch eine Jugendzeitschrift im Jahre 1971 erhobenen Datenmaterial zu Fragen der Berufswahl.

Für diese Untersuchung wurde eine der Fragen zur Art der gewünschten beruflichen Tätigkeit mit der Merkmalsausprägung „Ja, ich möchte Probleme lösen und mir häufig über eine Sache den Kopf zerbrechen“[1]) herangezogen und ihre Abhängigkeit von den Merkmalen „Geschlecht“ und „Schulbildung“ analysiert.

Es konnte gezeigt werden, daß das Regressionsmodell trotz der in den folgenden Abschnitten herausgearbeiteten Schwierigkeiten elastisch genug ist, um auch auf diesen Problemkreis anwendbar zu sein. Dies zu wissen ist u.a. deshalb von Bedeutung, weil die Regressionsanalyse mehr zu leisten vermag (falls die allerdings z. T. restriktiven Voraussetzungen des Regressionsmodells erfüllt sind) als die herkömmlichen, d. h. auf keinem bzw. keinem explizit formuliertem Modell basierenden Analysemethoden, insbesondere die tabellarische Auswertung von Häufigkeitsverteilungen.

So kann z. B. mit der Regressionsanalyse

1) für jede Merkmalsausprägung der unabhängigen Variablen über die geschätzten Regressionskoeffizienten ihr spezifischer Einfluß auf die abhängige Variable isoliert werden. Die Interpretation dieser Koeffizienten unterscheidet sich jedoch bei qualitativen Variablen von der bei quantitativen Variablen, wie im empirischen Teil dieser Arbeit noch näher erläutert wird;
2) der Einfluß eines bestimmten Merkmals auf die abhängige Variable untersucht werden, ohne daß der Einfluß der anderen unabhängigen Variablen mitgemessen wird;
3) ein Signifikanztest bzw. Konfidenzaussagen für die unbekannten Regressionskoeffizienten berechnet bzw. gemacht werden.

Zur Frage nach der Repräsentativität und der Interpretation der empirischen Ergebnisse wird in einem späteren Aufsatz berichtet.

2. Regressionsschätzung mit unabhängigen qualitativen Variablen

Sollen in einer Regressionsanalyse die Einflüsse von qualitativen (diskreten) Variablen auf eine abhängige Variable untersucht werden, so verwendet man in der Regel die Technik der „Dummy-Variablen“[2]). Diese ist auf jede Variable anwendbar, deren Variation in sich gegenseitig ausschließende Klassen eingeteilt werden kann. Somit ist eine Beschränkung auf qualitative Variablen nicht notwendig; jedoch erlangt die noch zu beschreibende Technik bei Vorliegen qualitativer Variablen ihre hauptsächliche Bedeutung. Hierbei wird für jede Klasse bzw. Merkmalsausprägung eine Dummy-Variable eingeführt, welcher die Werte 1 oder 0 zugeordnet werden, je nachdem, ob die Beobachtungen in die betreffende Klasse fallen bzw. diese Merkmalsausprägung aufweisen oder nicht. Der zugehörige Regressionskoeffizient isoliert dann den Effekt der spezifischen Merkmalsausprägung einer bestimmten qualitativen Variablen auf die abhängige Größe.

2.1 Multikollinearität

2.1.1 Darstellung des Problems

Wird das klassische lineare Regressionsmodell auf unabhängige qualitative Variablen übertragen, so erhält man für den Spezialfall einer solchen Variablen mit k Ausprägungen den folgenden Ansatz:

$$(1)\quad Y_n = \beta_o + \beta_1 X_{1n} + \dots + \beta_{kn} X_{kn} + u_n$$

$$\text{mit } X_{in} = \begin{cases} 1 & \text{falls die j-te Merkmalsausprägung beobachtet wurde } (j \varepsilon \{1, \dots, k\}) \\ 0 & \text{sonst} \end{cases}$$

$$E(u_n) = 0 \quad \text{für } n = 1, \dots, N;\ N = \text{Zahl der Beobachtungen}$$

$$E(u_n \cdot u_{n'}) = \begin{cases} \sigma^2 & \text{falls } n = n' \\ 0 & \text{sonst} \end{cases}$$

Für m Variablen ($m \geq 1$) lautet Modell (1) entsprechend[3]):

$$(2)\quad Y = \beta_o + \sum_{j=1}^{k_1} \beta_{1j} X_{1j} + \dots + \sum_{j=1}^{k_m} \beta_{mj} X_{mj} + u$$

$$= \beta_o + \sum_{i=1}^{m} \sum_{j=1}^{k_i} \beta_{ij} X_{ij} + u$$

$$\text{mit } X_{ij} = \begin{cases} 1 & \text{falls die j-te Merkmalsausprägung der i-ten Variablen beobachtet wurde} \\ 0 & \text{sonst} \end{cases}$$

$$\text{und } E(u) = 0$$

$$E(uu') = \begin{cases} \sigma^2 & \text{falls } u = u' \\ 0 & \text{sonst} \end{cases}$$

Im folgenden wird der Einfachheit halber die Matrix-Schreibweise eingeführt. Modell (2) lautet dann:

$$(2')\quad \mathbf{Y} = \mathbf{X\beta} + \mathbf{U}$$

wobei $\mathbf{X}$ eine (N, K) Matrix ist mit
N = Zahl der Beobachtungen und

$$K = 1 + \sum_{i=1}^{m} k_i,$$

$\boldsymbol{\beta}$ ein K-Spaltenvektor, $\mathbf{Y}$ und $\mathbf{U}$ N-Spaltenvektoren mit folgen-

[1]) Im folgenden kurz „Denken im Beruf“ genannt.

den Voraussetzungen für die Störvariable **U**:

$E(\mathbf{U}) = \mathbf{0}$

$E(\mathbf{UU}') = \sigma^2 \mathbf{I}$, wobei **I** die (N, N) Einheitsmatrix und σ^2 eine nichtnegative Konstante ist.

Aus (2′) erhält man durch Anwendung der Methode der kleinsten Quadrate das folgende System der Normalgleichungen[4]).

(3) $\mathbf{X}'\mathbf{X}\beta = \mathbf{X}'\mathbf{Y}$

Hieraus ergeben sich eindeutige Schätzwerte $\hat{\beta}$ für die Parameter β, falls die Matrix $\mathbf{X}'\mathbf{X}$ regulär ist. Es kann nun jedoch gezeigt werden, daß auf der Grundlage von Modell (2′) die Beobachtungsmatrix **X** der exogenen Variablen linear abhängige Spalten besitzt, was gerade eine singuläre Matrix $\mathbf{X}'\mathbf{X}$ zur Folge hat[5]).

Schreibt man eine typische Beobachtungsmatrix von (2′) für den Fall N = 7, K = 6 mit m = 2, $k_1 = 2$ und $k_2 = 3$ ausführlich, so sieht man, daß Spalte 1 eine Linearkombination sowohl der Spalten 2 und 3 als auch von 4, 5 und 6 ist:

$$\mathbf{X} = \begin{array}{c} \\ \\ \\ \\ \\ \\ \\ \Sigma \end{array} \begin{array}{c} \\ \overbrace{}^{\mathbf{X}_1} \\ \\ \\ \\ \\ \\ \\ 7 \end{array} \left(\begin{array}{c|cc|ccc} 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ \hline 7 & 3 & 4 & 2 & 2 & 3 \end{array} \right) \begin{array}{c} \\ \\ \\ \\ \\ \\ \\ \\ \end{array} \begin{array}{c} \\ \\ \\ \\ \\ \\ \\ \overbrace{}^{\mathbf{X}_2} \\ \end{array}$$

Bei Vorliegen linear abhängiger Spalten der Beobachtungsmatrix **X** spricht man von *vollständiger Multikollinearität*. Diese läßt sich durch folgende duale Aussagen charakterisieren[6]):

1. Es existiert *keine erwartungstreue* lineare *Schätzung* $\hat{\beta}$ für die Parameter β des linearen Regressionsmodells, d. h., für alle $\hat{\beta}$ gilt

$E(\hat{\beta}) \neq \beta$,

denn die hierfür notwendige und hinreichende Bedingung

$(\mathbf{X}'\mathbf{X})^{-1} \cdot \mathbf{X}'\mathbf{X} = \mathbf{I}$

ist nicht erfüllt wegen

a) $\mathrm{Rg}\,(\mathbf{X}'\mathbf{X})^{-1} \cdot \mathbf{X}'\mathbf{X} \leq \min\{\mathrm{Rg}\,(\mathbf{X}'\mathbf{X})^{-1} \cdot \mathbf{X}', \mathrm{Rg}\,\mathbf{X}\} = \mathrm{Rg}\,\mathbf{X} = K - l$

wobei l gleich der Zahl der linear abhängigen Spalten der Beobachtungsmatrix **X** ist[7])

und b) $\mathrm{Rg}\,\mathbf{I} = K$

wobei **I** hier die (K,K) Einheitsmatrix ist.

Somit existiert bei Vorliegen von vollständiger Multikollinearität keine lineare unverzerrte Parameterschätzung, und man sagt, die Parameter seien *nicht schätzbar*.

2. Die Parameter sind *nicht identifizierbar*, denn für das zugrunde gelegte Modell

$M = (\mathbf{X}\beta, \Phi)$ mit $\mathbf{X}\beta$ = lineare Funktionsform

und Φ = Wahrscheinlichkeitsverteilung der Störvariablen **U**

gibt es mindestens zwei Strukturen

$S_1 = \mathbf{X}\hat{\beta}$ und
$S_2 = \mathbf{X}\hat{\beta}^*$ mit $S_1 = S_2$.

Man sagt, die Strukturen sind *beobachtungsäquivalent*, d.h., aufgrund der beobachteten Daten allein kann keine Differenzierung zwischen den Parameterschätzungen $\hat{\beta}$ und $\hat{\beta}^*$ vorgenommen werden.

Entsprechend den beiden Aspekten der Multikollinearität bieten sich auch zwei Lösungsmöglichkeiten für das Multikollinearitätsproblem an.

2.2.2 Lösung mit Hilfe von schätzbaren Funktionen

Der erste Aspekt führt zum *Konzept der schätzbaren Funktionen*[8]). Hierbei werden anstatt der nicht schätzbaren Parameter β Linearkombinationen $\mathbf{p}'\beta$, also gewichtete Regressionskoeffizienten, betrachtet. Diese sind unter bestimmten Bedingungen erwartungstreu schätzbar durch

$\mathbf{p}'\,(\mathbf{X}'\mathbf{X})^- \cdot \mathbf{X}'\mathbf{Y}$

dabei ist $(\mathbf{X}'\mathbf{X})^-$ die *verallgemeinerte Inverse* (g-Inverse) der singulären Matrix $\mathbf{X}'\mathbf{X}$ und $\mathbf{p}'$ ein K-Zeilen-Vektor von Gewichten,

mit

$\mathrm{Var}\,(\mathbf{p}'\hat{\beta}) = \sigma^2\,\mathbf{p}'\,(\mathbf{X}'\mathbf{X})^- \cdot \mathbf{p}$

falls die Varianz-Kovarianz-Matrix von **U** durch $\sigma^2 \mathbf{I}$ gegeben ist.

Im folgenden wird eine Definition der verallgemeinerten Inversen und einige ihrer (für die im empirischen Teil dieser Arbeit benötigten) Eigenschaften gegeben[9]).

Def.: Eine g-Inverse der (K, K) Matrix $\mathbf{X}'\mathbf{X}$ ist eine (K, K) Matrix $(\mathbf{X}'\mathbf{X})^-$ mit
$(\mathbf{X}'\mathbf{X})\,(\mathbf{X}'\mathbf{X})^-\,(\mathbf{X}'\mathbf{X}) = \mathbf{X}'\mathbf{X}$

Satz: 1 a) $(\mathbf{X}'\mathbf{X})^-$ existiert genau dann, wenn $\mathbf{H} = (\mathbf{X}'\mathbf{X})^- \cdot \mathbf{X}'\mathbf{X}$ idempotent ist, d. h. $\mathbf{H}^2 = \mathbf{H}$
und $\mathrm{Rg}\,\mathbf{H} = \mathrm{Rg}\,\mathbf{X}'\mathbf{X} = \mathrm{Sp}\,\mathbf{H}$

b) Eine allgemeine Lösung des Gleichungssystems $\mathbf{X}'\mathbf{X}\beta = \mathbf{X}'\mathbf{Y}$ ist $\hat{\beta} = (\mathbf{X}'\mathbf{X})^-\,\mathbf{X}'\mathbf{Y} + (\mathbf{I} - \mathbf{H})\,\mathbf{Z}$, wobei **Z** ein beliebiger Vektor ist.

c) $\mathbf{p}'\hat{\beta}$ ist genau dann eindeutig für alle Lösungen von $\mathbf{X}'\mathbf{X}\beta = \mathbf{X}'\mathbf{Y}$, wenn $\mathbf{p}'\mathbf{H} = \mathbf{p}'$

Mit c) steht uns somit eine notwendige und hinreichende Bedingung für die erwartungstreue Schätzbarkeit von $\mathbf{p}'\beta$ zur Verfügung.

Ein Verfahren zur Berechnung der g-Inversen wird durch folgenden Satz angegeben:

Satz 2: $(\mathbf{X}'\mathbf{X})^- = \mathbf{U}\,\mathbf{D}^{-1}\,\mathbf{V}'$ mit

$\mathbf{U}$ = Matrix der normierten Eigenvektoren von $\mathbf{X}'\mathbf{X}$

$\mathbf{V}'$ = Matrix der normierten Eigenvektoren von $\mathbf{X}\mathbf{X}'$

---

[3]) Siehe: *D. B. Suits*, "Use of Dummy Variables in Regression Equations" JASA, 52, 1957, S. 548 – 551.

[3]) Hierbei wurde der besseren Übersicht wegen die Indizierung der Beobachtung weggelassen.

[4]) Die Ableitung dieser Normalgleichungen findet man z. B. bei *H. Schneeweiß*: Ökonometrie, 1971, S. 94 f.

[5]) Die Einführung eines Systems von Dummy-Variablen führt nur dann nicht zu einer singulären Matrix $\mathbf{X}'\mathbf{X}$, falls von einer homogenen Regressionsgleichung mit einer qualitativen unabhängigen Variablen ausgegangen wird. Dies wäre bei Modell (1) der Fall, falls die Konstante β_0 weggelassen würde.

[6]) Vgl.: *P. Schönfeld*: Methoden der Ökonometrie, Bd. 1, S. 82 ff.

[7]) Im Normalfall ist l gleich der Anzahl der exogenen Variablen.

[8]) Vgl.: *M. L. Garg, B. R. Rao, S. Mazumdar*: On an Estimation Problem in Multiple Regression, Statistische Hefte 2/1972.

[9]) Zur Theorie und Anwendung der verallgemeinerten Inversen von Matrizen siehe: *Rao, Mitra*: Generalized Inverse of Matrices and its Applications N. Y., 1971.

$\mathbf{D}^{-1}$ = Inverse Matrix der aus den Wurzeln der Eigenwerte der Matrix $\mathbf{X}'\mathbf{X}$ bestehenden Diagonalmatrix.

2.2.3 Lösung mit Hilfe von a-priori-Restriktionen

Der zweite Aspekt des Multikollinearitätsproblems führt zum *Konzept der a-priori-Restriktion* von Parameterwerten[10]). Hierbei werden so vielen Parametern a priori Werte zugewiesen wie der Spaltenrang der Matrix $\mathbf{X}$ von der Anzahl der Spalten von $\mathbf{X}$ differiert (diese sind in der Regel gleich der Zahl der betrachteten unabhängigen qualitativen Variablen). Hat man keine externe Information über die Werte der a priori festzulegenden Parameter, so ist es zweckmäßig, ihnen den Wert Null zuzuordnen.

Für Modell (2) erhält man dann durch folgende Transformation eine eindeutige Lösung für die restlichen Parameter:

$$(4)\ Y - \sum_{\substack{i=1 \\ l \varepsilon k_i}}^{m} \beta_{il}\, X_{il} = \beta_o + \sum_{i=1}^{m} \sum_{\substack{j=1 \\ j \neq l}}^{k_i} \beta_{ij}\, X_{ij}$$

Die Zweckmäßigkeit dieser Null-Restriktion erkennt man daran, daß auf der linken Seite von (4) nach wie vor Y steht und auf der rechten Seite nur die jeweils l-te Variable gestrichen zu werden braucht. Die verbleibenden Parameter können nun eindeutig geschätzt werden und sind somit identifizierbar.

Diese Lösung des Multikollinearitätsproblems ist im Vergleich zum Konzept der schätzbaren Funktionen ohne größeren zusätzlichen Aufwand durchführbar. Es ist jedoch zu beachten, daß gegen dieses Verfahren einige theoretische Einwände vorgebracht werden können:

1. Es besteht eine gewisse Willkür bei der Auswahl der a priori festzulegenden Parameter.
2. Hat man eine Auswahl getroffen, so erhält man nur dann erwartungstreue Schätzungen für die übrigen Parameter, falls die ausgewählten Parameter tatsächlich den gemachten Restriktionen genügen. Die Parameterschätzwerte sind also nur *bedingt erwartungstreu.*

3. Regressionsschätzung mit abhängigen qualitativen Variablen

Bei der eingangs erwähnten Untersuchung tritt neben den unabhängigen qualitativen Variablen auch eine abhängige qualitative Variable auf. Es ist daher naheliegend, auch diese durch eine Dummy-Variable auszudrücken und dann das klassische lineare Regressionsmodell auf die vorliegenden Daten anzuwenden. Es ergeben sich jedoch hierbei folgende zusätzliche Probleme, die sich negativ auf die Parameterschätzung oder die Signifikanztests für die Parameter auswirken können[11]).

3.1 Heteroskedastische Varianz der Störvariablen

3.1.1 Darstellung des Problems

Wie nachfolgend gezeigt wird, ist die Varianz der Störvariablen $\mathbf{U}$ *nicht konstant* für alle Beobachtungen, falls für die abhängige Variable $\mathbf{Y_n} \varepsilon \{0,1\}$ und für die Störvariable $\mathbf{U}$ $E(\mathbf{U}) = O$ vorausgesetzt wird. Die Störvariable $\mathbf{U}$ kann wegen $\mathbf{Y_n} \varepsilon \{0,1\}$ nur zwei Werte annehmen, nämlich

$1 - \mathbf{X_n}\,\boldsymbol{\beta}$ falls $\mathbf{Y_n} = 1$
und $\quad -\mathbf{X_n}\,\boldsymbol{\beta}$ falls $\mathbf{Y_n} = 0$

Diese werden wegen $E(\mathbf{U})$
$= (1 - \mathbf{X_n}\,\boldsymbol{\beta})\, p + (-\mathbf{X_n}\,\boldsymbol{\beta})\,(1 - p) = 0$
mit den Wahrscheinlichkeiten $\mathbf{X_n}\,\boldsymbol{\beta}$ und $1 - \mathbf{X_n}\,\boldsymbol{\beta}$ angenommen.

Dadurch erhält man für die Varianz der Störvariablen folgenden Ausdruck:

$$\begin{aligned} E(\mathbf{U_n}^2) &= p\,(1 - \mathbf{X_n}\,\boldsymbol{\beta})^2 + (1 - p)\,(-\mathbf{X_n}\,\boldsymbol{\beta})^2 \\ &= \mathbf{X_n}\,\boldsymbol{\beta}\,(1 - \mathbf{X_n}\,\boldsymbol{\beta})^2 + (1 - \mathbf{X_n}\,\boldsymbol{\beta})\,(-\mathbf{X_n}\,\boldsymbol{\beta})^2 \\ &= \mathbf{X_n}\,\boldsymbol{\beta}\,(1 - \mathbf{X_n}\,\boldsymbol{\beta}) \\ &= E(\mathbf{Y_n}) \cdot \left[1 - E\,(\mathbf{Y_n})\right] \end{aligned}$$

Somit kompliziert sich das lineare Regressionsmodell (2′) bei Vorliegen einer abhängigen qualitativen Variablen wie folgt

$$(5)\ \mathbf{Y} = \mathbf{X}\,\boldsymbol{\beta} + \mathbf{U}$$
mit $\mathbf{Y_n} \varepsilon \{0,1\}$
$E(\mathbf{U}) = 0$
$E(\mathbf{U}\,\mathbf{U}') = \sigma_n^2 \cdot \mathbf{I}$, wobei

$$\sigma_n^2 = E\,(\mathbf{Y_n})\left[1 - E\,(\mathbf{Y_n})\right]$$

und die Varianzen für die Regressionskoeffizienten lassen sich nach der Formel

$$(6)\ (\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'E\,(\mathbf{U}\,\mathbf{U}')\,\mathbf{X}\,(\mathbf{X}'\mathbf{X})^{-1}$$

berechnen, im Gegensatz zu

$$(7)\ \sigma^2\,(\mathbf{X}'\mathbf{X})^{-1}$$

bei Gültigkeit von Modell (2′).

3.1.2 Lösung mit Hilfe der Aitkenschätzung (verallgemeinerte Methode der kleinsten Quadrate)[12])

Effiziente Parameterschätzwerte können mit der von Aitken entwickelten *verallgemeinerten* Methode der kleinsten Quadrate berechnet werden mit Hilfe von

$$(8)\ \hat{\boldsymbol{\beta}} = (\mathbf{X}^{*\prime}\,\mathbf{X}^*)^{-1}\,\mathbf{X}^{*\prime}\,\mathbf{Y}^*$$

wobei $\mathbf{X}^{*\prime}$, $\mathbf{X}^*$ und $\mathbf{Y}^*$ die mit den Hauptdiagonalelementen von $E\,(\mathbf{U}\,\mathbf{U}')$ von Modell (5) gewichteten ursprünglichen Matrizen $\mathbf{X}'$, $\mathbf{X}$ und $\mathbf{Y}$ sind

und

$$(9)\ \mathrm{Var}\,(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^{*\prime}\,\mathbf{X}^*)^{-1}$$

Diese Methode berücksichtigt also die heteroskedastischen Varianzen der Störvariablen; die Schwierigkeit ist nur, daß sie im allgemeinen nicht bekannt sind und zunächst mit Hilfe der gewöhnlichen Methoden der kleinsten Quadrate geschätzt werden müssen[13]).

3.2 Linearität des Regressionsmodells

3.2.1 Darstellung des Problems

Bildet man in Gleichung (5) auf der linken und der rechten Seite den Erwartungswert, so erhält man folgendes äquivalentes Regressionsmodell:

$$(5')\ E\,(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$$
mit $\mathbf{Y_n} \varepsilon \{0,1\}$
und $\mathrm{Var}\,(\mathbf{Y_n}) = E\,(\mathbf{Y_n})\left[1 - E\,(\mathbf{Y_n})\right]$

[10]) Vgl.: *H. Theil*: Principles of Econometrics, 1971, S. 152 ff.
[11]) Vgl.: *D. Huang*: Regression and Econometric Methods, 1969, S. 163 ff.
[12]) Vgl.: *A. Goldberger*: Economic Theory, 1964, S. 232 ff.
[13]) Bei der im empirischen Teil dieser Arbeit durchgeführten Untersuchung sind diese Varianzen jedoch bekannt.

Wegen $\quad \mathbf{Y_n} \in \{0,1\}$ folgt

$$E(\mathbf{Y_n}) = 1 \cdot p(\mathbf{Y_n} = 1) + 0 \cdot p(\mathbf{Y_n} = 0) = p(\mathbf{Y_n} = 1)$$

d. h., die Regressionsgleichung (5′) repräsentiert eine *lineare Wahrscheinlichkeitsfunktion.* Dabei kann die rechte Seite von Gleichung (5′) größer als 1 oder kleiner als 0 sein, falls für β die Schätzwerte $\hat{\beta}$ eingesetzt werden, während auf der linken Seite E (**Y**) wegen E (**Y**) = p (**Y** = 1) nicht außerhalb dieser Grenzen liegen kann. Tritt dieser Fall ein, so können die entsprechenden Ergebnisse nicht mehr sinnvoll interpretiert werden. Dies wird um so häufiger vorkommen, je näher die Wahrscheinlichkeit für das Auftreten der zu untersuchenden Merkmalsausprägung der abhängigen qualitativen Variablen bei Null oder Eins liegt.

3.2.2 Alternative zur linearen Funktionsform: Das „Logit"-Modell[14])

Das auf einer logistischen Funktionsform basierende *Logit-Modell*[15]) bietet die Gewähr dafür, daß die beim linearen Regressionsmodell möglichen nicht interpretierbaren Parameterschätzwerte nicht auftreten können.

Hierbei werden für jede Merkmalskombination der unabhängigen Variablen relative Häufigkeiten $\hat{p}$ für das Auftreten der zu untersuchenden Merkmalsausprägung der abhängigen Variablen berechnet. Diese werden dann einer monotonen Transformation derart unterzogen, daß die transformierten Werte zwischen $-\infty$ und $+\infty$ schwanken, wenn die ursprünglichen Werte zwischen 0 und 1 variieren.

Das Logit-Modell lautet

$$(10)\quad \mathbf{Y} = \ln\left(\frac{\hat{\mathbf{p}}}{1-\hat{\mathbf{p}}}\right) = \mathbf{X\beta} + \mathbf{U}$$

mit $E(\mathbf{U}) = \mathbf{0}$

$$E(\mathbf{UU'}) \approx \frac{1}{n_i \hat{p}_i (1-\hat{p}_i)}\,\mathbf{I}$$

und n_i = Anzahl der bei der i-ten Merkmalskombination der unabhängigen Variablen aufgetretenen Merkmalsausprägungen der abhängigen Variablen[16]).

Die Auflösung von (10) nach $\hat{\mathbf{p}}$ ergibt das äquivalente Modell

$$(11)\quad \hat{\mathbf{p}} = \frac{1}{1 + e^{-\mathbf{X\beta}}} + \mathbf{U}$$

Hierbei tritt die logistische Funktionsform deutlich in Erscheinung.

3.3 Signifikanztest für die unbekannten Parameter

Um Vertrauensaussagen für einzelne oder mehrere Regressionskoeffizienten, Linearkombinationen von Regressionskoeffizienten oder E (**Y**) selbst machen zu können, wird üblicherweise angenommen, daß

Y normalverteilt ist mit Erwartungswert $\mathbf{X\beta}$ und Varianz $\sigma^2 \mathbf{I}$.

Trifft diese Hypothese zu, so sind die Parameterschätzungen $\hat{\boldsymbol{\beta}}$ normalverteilt mit Erwartungswert $\boldsymbol{\beta}$ und Varianz $\sigma^2 (\mathbf{X'X})^{-1}$, da sie wegen $\hat{\boldsymbol{\beta}} = (\mathbf{X'X})^{-1} \cdot \mathbf{X'Y}$ eine Linearkombination der normalverteilten abhängigen Variablen sind. Mit der Formel

$$t_{N-K} = \frac{\hat{\beta} - \beta}{\sqrt{\mathrm{Var}\,(\hat{\beta})}}$$ steht dann eine Testgröße

zur Verfügung, die t-verteilt ist mit (N – K) Freiheitsgraden und mit der z. B. die Nullhypothese $\beta = 0$ (eine bestimmte Merkmalsausprägung einer unabhängigen Variablen hat keinen signifikanten Einfluß auf die abhängige Variable) überprüft werden kann.

Ist die abhängige Variable nicht normalverteilt (dies trifft bei Vorliegen einer qualitativen abhängigen Variablen zu, da sie, wie oben gezeigt, Null-Eins verteilt ist mit Erwartungswert $\mathbf{X\beta}$ und Varianz $\mathbf{X_n\beta} \cdot (1 - \mathbf{X_n\beta})$), so ist $\hat{\boldsymbol{\beta}}$ nicht notwendig normalverteilt und somit ist auch die Testgröße t_{N-K} nicht notwendig t-verteilt. Die Durchführbarkeit eines Signifikanztests für $\boldsymbol{\beta}$ hängt dann davon ab, ob $(\mathbf{X'X})^{-1}\,\mathbf{X'Y}$ trotzdem normalverteilt ist.

Die Lösung dieses Problems ist – im Gegensatz zu den beiden ersten Problemen im Zusammenhang mit abhängigen qualitativen Variablen – demnach eine Frage der empirischen Überprüfung der Normalverteilungsvoraussetzung für $\hat{\boldsymbol{\beta}}$.

Mit den bisher vorgestellten Methoden können nun im folgenden Abschnitt die empirischen Ergebnisse bereitgestellt und diskutiert werden.

4. Empirischer Teil

Zur Untersuchung der Frage nach der Abhängigkeit der Variablen „Denken im Beruf"[17]) von den Merkmalen „Geschlecht" und „Schulbildung" standen 20372 Beobachtungen zur Verfügung. Die jeweiligen Merkmalsausprägungen wurden durch folgende „Dummy-Variablen" gekennzeichnet:

Geschlecht: X_{11} = männlich
X_{12} = weiblich

Schulbildung[18]): X_{21} = Universität
X_{22} = Handelsschule
X_{23} = Fachschule
X_{24} = Berufsschule
X_{25} = Gymnasium
X_{26} = Realschule
X_{27} = Volksschule
X_{28} = ohne Angabe

Während man für die entsprechenden Regressionsparameter ohne vorherige Lösung des Multikollinearitätsproblems keine Schätzwerte erhält, berühren die im Zusammenhang mit der abhängigen Variablen auftretenden Probleme der Konfidenzaussagen sowie der linearen Funktionsform nicht die Schätzbarkeit der Parameter. Daher wird bei der folgenden Diskussion der Ergebnisse zunächst nach der Art der Lösung des Multikollinearitätsproblems unterschieden.

4.1 Parameterschätzung mit Hilfe von a-priori-Restriktionen

4.1.1 Interpretation der geschätzten Regressionskoeffizienten

Da für die Parameter des zunächst zugrunde gelegten

---

[14]) *H. Theil*: Principles of Econometrics, 1971, S. 632 ff.

[15]) Der Terminus „Logit" wurde von *Berkson* eingeführt und soll an „logistisch" erinnern. Das Logit-Modell ist nicht linear in den Variablen, jedoch linear in den Parametern.

[16]) Bei der Berechnung der Varianzen von $\ln \frac{\hat{p}}{1-\hat{p}}$ wird vorausgesetzt, daß die relativen Häufigkeiten von unabhängigen Stichproben binomialverteilter Grundgesamtheiten stammen.

[17]) Die entsprechende Merkmalsausprägung lautet: „Ja, ich möchte Probleme lösen und mir häufig über eine Sache den Kopf zerbrechen."

[18]) Die Frage lautete: „Welche Schule hast Du zuletzt besucht oder in welche Schule gehst Du noch?"

Modells (2) — wobei hier die Parameter für das Merkmal „Geschlecht“ mit α_1 (i = 1, 2) und diejenigen für das Merkmal „Schulbildung“ mit β_1 (i=1,...,8) bezeichnet wurden — keine externe Information vorlag, wurde unter Berücksichtigung der Transformation (4) mit $\alpha_1 := 0$ und $\beta_8 := 0$ folgende Schätzfunktion berechnet[19]):

$$
\begin{aligned}
(12)\ \hat{Y} = {} & \underset{(0{,}0328)}{0{,}502} \\
& \underset{(0{,}0077)}{-0{,}143 X_{12}} \quad \underset{(0{,}0714)}{+0{,}393 X_{21}} \quad \underset{(0{,}0355)}{+0{,}230 X_{22}} \\
& \underset{(0{,}0331)}{} \qquad \underset{(0{,}0078)}{} \qquad \underset{(0{,}0892)}{} \qquad \underset{(0{,}0359)}{} \\
& \underset{(0{,}0347)}{+0{,}235 X_{23}} \quad \underset{(0{,}0334)}{+0{,}147 X_{24}} \quad \underset{(0{,}0333)}{+0{,}206 X_{25}} \quad \underset{(0{,}0331)}{+0{,}123 X_{26}} \\
& \underset{(0{,}0351)}{} \qquad \underset{(0{,}0337)}{} \qquad \underset{(0{,}0336)}{} \qquad \underset{(0{,}0334)}{} \\
& \underset{(0{,}0330)}{+0{,}070 X_{27}} \qquad R^2 = 54{,}7\ \%
\end{aligned}
$$

Die geschätzten Regressionskoeffizienten geben hier im Gegensatz zur Regressionsschätzung mit quantitativen Variablen nicht den partiellen Differentialquotienten an, sondern stellen ein Maß für das relative Einflußgewicht einer spezifischen Merkmalsausprägung der unabhängigen Variablen auf die betrachtete Merkmalsausprägung der abhängigen Variablen dar. Als Bezugsbasis dienen die a priori festgelegten Parameterwerte. Demnach besagt der Wert — 0,143 in (12), daß der Einfluß der Merkmalsausprägung „weiblich“ auf die abhängige Variable „Denken im Beruf“ um 0,143 niedriger ist als der der Merkmalsausprägung „männlich“.

Auch folgende Interpretation ist zulässig und gibt den Koeffizienten eine anschaulichere Bedeutung:

Die Wahrscheinlichkeit, daß ein Mädchen die Merkmalsausprägung „Denken im Beruf“ wählt, ist um rd. 14 % niedriger als die entsprechende Wahrscheinlichkeit für Jungen, denn für jede Merkmalsausprägung X_{2j} mit $j \in \{1, \ldots, 8\}$ gilt:

$$
\begin{aligned}
& p\,(Y = 1/X_{12} = 1 \text{ und } X_{2j} = 1) - \\
& p\,(Y = 1/X_{11} = 1 \text{ und } X_{2j} = 1) \\
& = (0{,}502 - 0{,}143 + \hat{\beta}_j) - (0{,}502 + 0 + \hat{\beta}_j) = \\
& -0{,}143 \approx -14\ \%
\end{aligned}
$$

Entsprechende Vergleiche wurden auch für jeweils zwei Merkmalsausprägungen des Merkmals „Schulbildung“ berechnet (vgl. Tabelle 1). Dabei ergibt sich z. B. für die Merkmalsausprägung „Fachschule“ eine um 0,235 — 0,206 = 0,029 oder rd. 3 % größere Wahrscheinlichkeit gegenüber der Merkmalsausprägung „Gymnasium“.

Es wäre jedoch voreilig, aufgrund dieser Zahl auf einen signifikanten Unterschied zwischen diesen beiden Merkmalsausprägungen zu schließen. Um derartige Fragen zu beantworten, wurde ein Signifikanztest durchgeführt, bei dem für die Merkmalsausprägung X_{21} („Universität“) die Nullhypothese

H_0: $\beta_1 = \beta_j$ (j = 2, 3, ..., 7) der Alternativhypothese
H_1: $\beta_1 > \beta_j$
und für die Merkmalsausprägungen
$X_{22}, \ldots, X_{27}$ die Nullhypothese
H'_0: $\beta_i = \beta_j$ (i, j = 2, 3, ..., 7; i < j) der Alternative
H'_1: $\beta_i \neq \beta_j$ gegenübergestellt wurde.

Als Testgröße diente

$$
t = \frac{\hat{\beta}_i - \hat{\beta}_j}{\sqrt{\mathrm{Var}\,(\hat{\beta}_i - \hat{\beta}_j)}} \quad (\text{mit } \mathrm{Var}\,(\hat{\beta}_i - \hat{\beta}_j) = \mathrm{Var}\,(\hat{\beta}_i) + \mathrm{Var}\,(\hat{\beta}_j) - 2\,\mathrm{Cov}\,(\hat{\beta}_i, \hat{\beta}_j)),
$$

die t-verteilt ist mit Freiheitsgrad ∞, falls die Parameterschätzwerte normalverteilt sind (dies wird hier unterstellt) und die Nullhypothese zutrifft. Die Unterstellung der Normalverteilungsvoraussetzung ist nicht ganz unplausibel, da die geschätzten Regressionskoeffizienten hierbei im wesentlichen gewichtete Durchschnitte der beobachteten Werte der abhängigen Variablen sind und als solche dem zentralen Grenzwertsatz unterliegen.

In nachfolgender Tabelle sind die mit * bzw. ** gekennzeichneten prozentualen Unterschiede in den betreffen-

[19]) Die über und unter den Koeffizienten eingeklammerten Zahlen geben die in 4.1.3 näher erläuterten Standardfehler an.

Tabelle 1: Prozentuale Unterschiede in den Wahrscheinlichkeiten für das Auftreten der Merkmalsausprägung „Denken im Beruf“

von \ zu	Universität	Handelsschule	Fachschule	Berufsschule	Gymnasium	Realschule	Volksschule
Universität	–	+ 16,3*	+ 15,8*	+ 24,6**	+ 18,7*	+ 27,0**	+ 32,3**
Handelsschule		–	– 0,5	+ 8,3**	+ 2,4	+ 10,7**	+ 16,0**
Fachschule			–	+ 8,8**	+ 2,9	+ 11,2**	+ 16,5**
Berufsschule				–	– 5,9**	+ 2,4*	+ 7,7**
Gymnasium					–	+ 8,3**	+ 13,6**
Realschule						–	+ 5,3**

den Wahrscheinlichkeiten für die abhängige Variable signifikant auf dem 95-%- bzw. 99-%-Konfidenzniveau. Die nicht gekennzeichneten Werte sind nicht signifikant.

Überraschend sind hier auf den ersten Blick die „nur" mit 95 v. H. Sicherheitswahrscheinlichkeit signifikanten Unterschiede zwischen den Merkmalsausprägungen: „Universität" und „Handelsschule", „Universität" und „Fachschule" sowie „Universität" und „Gymnasium". Dies rührt jedoch von dem durchschnittlich 2,5mal höheren Standardfehler für den Parameter der Merkmalsausprägung „Universität" gegenüber den Parametern der anderen Ausprägungen des Merkmals „Schulbildung" her, was wiederum auf die extrem niedrige Besetzungszahl in dieser speziellen Gruppe zurückzuführen ist.

Zu beachten ist auch, daß eventuell eine Verfälschung der Ergebnisse durch die Nichtberücksichtigung weiterer unabhängiger Variablen, insbesondere des Merkmals „Alter", eingetreten ist.

4.1.2 Bedingte Erwartungstreue der geschätzten Regressionskoeffizienten

Wie in Abschnitt 2 bereits erwähnt, sind die Parameterschätzwerte für das Modell (2) bedingt erwartungstreu, d. h., sie sind erwartungstreu unter der Voraussetzung, daß die restringierten Parameter tatsächlich den festgelegten Beschränkungen genügen. Da dies in der Regel nicht der Fall ist, können durch Variation der Restriktionen eventuell bessere Schätzfunktionen erhalten werden. Eine Schätzfunktion soll im Vergleich zu einer anderen als „besser" bezeichnet werden, wenn die Summe der absoluten Abweichungen der relativen Häufigkeiten von den aus der Modellstruktur errechneten Wahrscheinlichkeiten für das Auftreten der untersuchten Merkmalsausprägung kleiner ist.

Die letzte Zeile der nachfolgenden Tabelle 2 gibt das Ergebnis einer sukzessiven Erhöhung des Parameters β_8 wieder. Wie man sieht, liegt bei $\beta_8 = 0{,}26$ ein relatives Minimum für die als Maß für die „Güte des Modells" anzusehende Summe der absoluten Abweichungen ($\Sigma|\Delta|$).

4.1.3 Effizienz der geschätzten Regressionskoeffizienten

Die in (12) unterhalb der Parameterschätzwerte angegebenen Standardfehler wurden unter der Vorausetzung homoskedastischer Varianzen der Störvariablen berechnet. Da diese Voraussetzung — wie in Abschnitt 3.1 gezeigt — nicht zutrifft, sind die mit Formel (7) berechneten Standardfehler nur näherungsweise richtig. Die „wahren" Standardfehler ergeben sich auf der Grundlage von Modell (5) mittels der komplizierteren Formel (6) und sind oberhalb der Parameterschätzwerte in (12) angegeben. Die „wahren" Standardfehler sind hierbei stets niedriger als ihre Näherungswerte; dies braucht jedoch nicht generell so zu sein.

Dagegen sind die mittels der in (9) angegebenen Aitkenschen Schätzformel berechneten Standardfehler *immer* niedriger als die oben betrachteten „wahren" Standardfehler. Die nachfolgende Schätzfunktion macht dies deutlich:

Tabelle 2: Veränderung der Parameter des linearen Modells ohne Berücksichtigung der Heteroskedastizität bei Variation des a priori restringierten Parameters β_8 (≙ ohne Angabe zur Schulbildung).

Mehrmalsausprägung	$\beta_8 := 0$	$\beta_8 := 0{,}1$	$\beta_8 := 0{,}2$	$\beta_8 := 0{,}26$	$\beta_8 := 0{,}3$
Geschlecht					
männlich	0,50109	0,46682	0,43255	0,41198	0,39828
weiblich	0,35844	0,32424	0,29004	0,26953	0,25585
Schulbildung					
Universität	0,39270	0,42693	0,46117	0,48171	0,49541
Handelsschule	0,23035	0,26457	0,29878	0,31931	0,33300
Fachschule	0,23474	0,26895	0,30317	0,32370	0,33739
Berufsschule	0,14717	0,18140	0,21562	0,23615	0,24984
Gymnasium	0,20577	0,23999	0,27421	0,29474	0,30843
Realschule	0,12306	0,15728	0,19149	0,21202	0,22570
Volksschule	0,07004	0,10425	0,13847	0,15899	0,17268
Ohne Angabe	0	0,1	0,2	0,26	0,3
Konstante	0	0	0	0	0
$\Sigma \mid \Delta \mid$	0,34950	0,28460	0,28436	0,28416	0,30574

$$
\begin{aligned}
(13)\ \hat{Y} = {} & 0{,}460 && -0{,}145\,X_{12} && +0{,}431\,X_{21} && +0{,}268\,X_{22} \\
& (0{,}0319) && (0{,}0077) && (0{,}0709) && (0{,}0346) \\
& +0{,}279\,X_{23} && +0{,}189\,X_{24} && +0{,}247\,X_{25} \\
& (0{,}0338) && (0{,}0326) && (0{,}0324) \\
& +0{,}165\,X_{26} && +0{,}111\,X_{27} && R^2 = 55{,}1\ \%. \\
& (0{,}0322) && (0{,}0319)
\end{aligned}
$$

Ein Vergleich mit den in (12) angegebenen Standardfehlern ergibt nur einen geringen Effizienzverlust, so daß der oben durchgeführte t-Test nicht seine Gültigkeit verliert.

Im allgemeinen wird jedoch der Unterschied zwischen den verschiedenen Standardfehlern um so größer sein, je stärker sich die einzelnen relativen Häufigkeiten für das Auftreten der zu untersuchenden Merkmalsausprägung unterscheiden, denn diese relativen Häufigkeiten gehen hauptsächlich in die Berechnung der Varianzen der Störvariablen ein. Daher wird sich der zusätzliche Aufwand, der mit der verallgemeinerten Methode der kleinsten Quadrate verbunden ist, vom Ergebnis her erst lohnen, wenn die Variation der relativen Häufigkeiten sehr groß ist.

4.1.4 Vergleich der Ergebnisse des logistischen und des linearen Modells

Auf der Grundlage des logistischen Modells (10) ergab sich mittels der verallgemeinerten Methode der kleinsten Quadrate folgende Schätzfunktion:

$$
\begin{aligned}
(14)\ \hat{Y} = \ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = {} & -0{,}162 && -0{,}618\,X_{12} && +2{,}038\,X_{21} && +1{,}127\,X_{22} \\
& (0{,}142) && (0{,}033) && (0{,}470) && (0{,}154) \\
& +1{,}151\,X_{23} && +0{,}785\,X_{24} && +1{,}034\,X_{25} \\
& (0{,}151) && (0{,}145) && (0{,}144) \\
& +0{,}694\,X_{26} && +0{,}485\,X_{27} && R^2 = 98{,}9\ \%. \\
& (0{,}143) && (0{,}142)
\end{aligned}
$$

Die Größenordnung sowie die Vorzeichen der Parameterschätzwerte geben Hinweise auf die Art des Einflusses einer bestimmten Merkmalsausprägung der unabhängigen Variablen auf die abhängige Variable. Jedoch lassen sich die Koeffizienten bzw. bestimmte Linearkombinationen davon nicht wie beim linearen Modell direkt als Wahrscheinlichkeiten interpretieren. Um dies auch hier tun zu können, muß zunächst die Transformation $\ln\left(\frac{\hat{p}}{1-\hat{p}}\right)$ nach $\hat{p}$ aufgelöst werden. Für die Wahrscheinlichkeit, daß etwa ein Mädchen mit Volksschule die Merkmalsausprägung „Denken im Beruf" wählt, ergibt sich demnach

$$
\hat{p} = \frac{e^{-0{,}162\ -0{,}618\ +0{,}485}}{1 + e^{-0{,}162\ -0{,}618\ +0{,}485}} = 0{,}427 \text{ oder } 42{,}7\ \%
$$

Zum Vergleich des linearen Modells und des logistischen Modells mit den empirischen Daten wurden in Tabelle 3 die jeweils errechneten Wahrscheinlichkeiten den relativen Häufigkeiten gegenübergestellt und die Summe der absoluten Abweichungen von den relativen Häufigkeiten ($\Sigma/\Delta/$) berechnet. Daran ist zu erkennen, daß das logistische Modell die beste Anpassung an die empirischen Daten liefert. Jedoch zeigt sich, daß dies nicht generell bessere Ergebnisse bringt und daß man mit dem linearen Modell insbesondere bei nicht extremen relativen Häufigkeiten eine durchaus brauchbare Anpassung erhält. Daß die Anpassung bei extremen relativen Häufigkeiten schlechter ist, überrascht nicht, wenn man berücksichtigt, daß das lineare Modell nicht notwendig die für relative Häufigkeiten vorliegende Beschränkung auf das Intervall [0,1] erfüllt.

Solange die relativen Häufigkeiten nicht extrem nahe bei null oder eins liegen, dürften auch kaum „Wahrscheinlichkeiten" auftreten, die größer als eins oder kleiner als null sind.

Wenn auch das mit Hilfe der gewöhnlichen Methode der kleinsten Quadrate geschätzte lineare Modell dem Theoretiker u. a. wegen der Möglichkeit des Auftretens nicht interpretierbarer Parameterschätzwerte unpassend erscheinen mag, kann es für den Praktiker dennoch gute Dienste leisten, wenn die Ansprüche an die Genauigkeit nicht zu hoch sind. Außerdem bietet es den Vorteil, daß die für das Aufstellen des Normal-Gleichungssystems notwendige Matrix $\mathbf{X'X}$ ohne zusätzlichen Aufwand direkt aus den Kreuztabellen übernommen werden kann und nicht auf das Urmaterial zurückgegriffen werden braucht, wie das beim logistischen, aber auch beim linearen Modell mit Berücksichtigung der Heteroskedastizität der Fall ist.

4.2 Parameterschätzung mit Hilfe von schätzbaren Funktionen

4.2.1 Ableitung einer schätzbaren Funktion

Will man den im Zusammenhang mit der a-priori-Restriktion von Parameterwerten genannten theoretischen Einwänden begegnen, so ist es zweckmäßig, anstatt der individuellen Regressionskoeffizienten Linearkombinationen, also gewichtete Regressionskoeffizienten, zu schätzen, deren Erwartungstreue mit dem in 2.2.2 angegebenen Kriterium nachgeprüft werden kann.

Im folgenden sollen als Beispiel einer schätzbaren Funktion die von Feldstein [20] — zur Untersuchung der Auswirkung von sozialen und biologischen Faktoren der perinatalen Mortalität — benutzten gewichteten Regressionskoeffizienten betrachtet werden. Darin werden die Gewichte der ursprünglichen Regressionskoeffizienten eines bestimmten Merkmals $\mathbf{X}$ mit r Ausprägungen wie folgt definiert:

$$
\mathbf{p'}_i = \left(-\frac{n_1}{N}, -\frac{n_2}{N}, \ldots, +\left(1-\frac{n_i}{N}\right), -, \ldots, -\frac{n_r}{N}\right)
$$

$$
\text{mit } \frac{n_i}{N} = \frac{\text{Zahl der Probanden mit Merkmalsauspr. i}}{\text{Gesamtzahl der Probanden}}
$$

Multipliziert man diesen Gewichtsvektor mit den ursprünglichen Regressionskoeffizienten, so ergeben sich für das Merkmal „Schulbildung" folgende schätzbaren Funktionen

$$
\begin{aligned}
(15)\ \mathbf{p'}_i\boldsymbol{\beta} &= -\frac{n_1}{N}\beta_1 - \frac{n_2}{N}\beta_2 - \ldots + \left(1-\frac{n_i}{N}\right)\beta_i \\
&= -\ \ldots\ - \frac{n_8}{N}\beta_8 \\
&= \beta_i\left(1-\frac{n_i}{N}\right) - \sum_{j\neq i}^{8}\beta_j\frac{n_j}{N} \\
&= \beta_i - \sum_{j=1}^{8}\beta_i\frac{n_j}{N} \\
&= \beta_i - \bar{\beta}
\end{aligned}
$$

[20] Vgl.: *M. S. Feldstein*: „A Binary Variable Multiple Regression Method of Analyzing Factors Affecting Peri-natal Mortality and Other Outcomes of Pregnancy", *J. Roy.* Stat.Soc., Series A, 1966, S. 61 – 73.

Tabelle 3: Vergleich der Anpassung verschiedener Modelle an die empirischen Daten

Merkmalsausprägungen	Rel. Häufigkeit in %	Wahrscheinlichkeiten in %		
	ohne Modell	Lineares Modell mit gewöhnlicher Methode der kleinsten Quadrate	Lineares Modell mit verallgemeinerter Methode der kleinsten Quadrate	Logistisches Modell mit verallgemeinerter Methode der kleinsten Quadrate
(X_{11}, X_{21})	84,21	89,40	89,12	86,71
(X_{11}, X_{22})	70,83	73,18	72,88	72,42
(X_{11}, X_{23})	77,15	73,62	73,91	72,89
(X_{11}, X_{24})	64,45	64,86	64,81	65,09
(X_{11}, X_{25})	69,87	70,72	70,71	70,52
(X_{11}, X_{26})	61,96	62,45	62,53	62,99
(X_{11}, X_{27})	58,06	57,15	57,18	58,01
(X_{11}, X_{28})	40,91	50,15	46,04	45,95
(X_{12}, X_{21})	81,25	75,08	74,57	77,87
(X_{12}, X_{22})	59,15	58,86	58,34	58,60
(X_{12}, X_{23})	57,66	59,30	59,36	59,17
(X_{12}, X_{24})	50,46	50,54	50,26	50,13
(X_{12}, X_{25})	56,64	56,40	56,17	56,33
(X_{12}, X_{26})	48,16	48,13	47,99	47,86
(X_{12}, X_{27})	42,37	42,83	42,46	42,69
(X_{12}, X_{28})	32,61	35,83	31,49	31,43
$\Sigma \mid \Delta \mid$	-	35,10	29,40	23,64

Entsprechend erhält man für das Merkmal „Geschlecht" die schätzbaren Funktionen

(16) $\mathbf{p}'_i\boldsymbol{\alpha} = \alpha_i - \bar{\alpha}$

Nach diesem Prinzip wurde die im Anhang angegebene Gewichtsmatrix **P** erstellt.

Um mittels des Produktes $\mathbf{P} \cdot \boldsymbol{\beta}$ Schätzwerte für die in (15) und (16) betrachteten gewichteten Regressionskoeffizienten zu erhalten, ist zuvor irgendeine Lösung für die ursprünglichen Parameter $\boldsymbol{\beta}$ erforderlich. Diese erhält man durch Multiplikation der verallgemeinerten Inversen der singulären Matrix $\mathbf{X}'\mathbf{X}$ mit dem Spaltenvektor $\mathbf{X}'\mathbf{Y}$. (Beide Matrizen sind im Anhang aufgeführt.) Schwierigkeiten bereitet dabei im allgemeinen die Berechnung der verallgemeinerten Inversen.

Für den vorliegenden Fall zweier Merkmale konnte die Struktur einer verallgemeinerten Inversen dem unter (8) zitierten Aufsatz entnommen werden.

Die Varianzen für die gewichteten Regressionskoeffizienten (15) und (16) errechnen sich durch

(17) $\mathbf{p}'\boldsymbol{\Sigma}\mathbf{p}$, wobei $\boldsymbol{\Sigma} = \sigma^2 (\mathbf{X}'\mathbf{X})^-$ die Varianz-Kovarianz-Matrix der ursprünglichen Regressionskoeffizienten für den Fall einer homoskedastischen Varianz der Störvariablen **U** ist.

4.2.2 Interpretation der gewichteten Regressionskoeffizienten

Die in (15) und (16) auftretenden Mittelwerte $\bar{\alpha}$ und $\bar{\beta}$ geben den „Durchschnittseffekt" der beiden Merkmale „Geschlecht" und „Schulbildung" auf die abhängige Variable an. Demnach sind die gewichteten Regressionskoeffizienten $\alpha_i - \bar{\alpha}$ bzw. $\beta_i - \bar{\beta}$ als lineare Abweichungen der spezifischen Merkmalsausprägungen vom „Durchschnittseffekt" zu interpretieren.

In Tabelle 4 sind diese linearen Abweichungen in Prozenten des Mittelwertes $\bar{Y}$ der abhängigen Variablen angegeben und den Ergebnissen einer herkömmlichen Analysemethode gegenübergestellt, bei der die Abweichungen der merkmalsspezifischen relativen Häufigkeiten auch in Prozenten von $\bar{Y}$ betrachtet wurden. Der Vergleich der Ergebnisse ist deshalb von Interesse, weil bei den auf der Regressionsanalyse basierenden Abweichungen der Einfluß des jeweils anderen Merkmals eliminiert ist, während dies bei den ohne Modell errechneten Zahlen nicht der Fall ist. Dies verdeutlichen die folgenden Berechnungsformeln:

(18) $\left(\frac{\alpha_i - \bar{\alpha}}{\bar{Y}}\right) \cdot 100$ bzw. $\left(\frac{\beta_i - \bar{\beta}}{\bar{Y}}\right) \cdot 100$ (Regressionsmodell)

Tabelle 4: Prozentuale Abweichung einer spezifischen Merkmalsausprägung vom Durchschnittseffekt

Merkmalsausprägung	**ohne** Elimination des Einflusses des jeweils anderen Merkmals	**nach** Elimination des Einflusses des jeweils anderen Merkmals
Geschlecht		
männlich	+ 20,99 (1,39)	+ 19,67 (1,08)
weiblich	- 7,68 (1,01)	- 7,18 (0,39)
Schulbildung		
Universität	+ 55,95 (12,01)	+ 48,57 (15,56)
Handelsschule	+ 17,64 (2,90)	+ 18,01 (2,79)
Fachschule	+ 19,01 (2,55)	+ 18,84 (2,42)
Berufsschule	+ 4,24 (1,80)	+ 2,35 (1,53)
Gymnasium	+ 14,97 (1,70)	+ 13,38 (1,47)
Realschule	- 3,30 (1,58)	- 2,18 (1,25)
Volksschule	- 13,07 (1,31)	- 12,16 (0,91)
ohne Angabe	- 27,35 (6,10)	- 23,35 (6,08)

$$(19)\quad \left(\frac{\bar{Y}_i - \bar{Y}}{\bar{Y}}\right) \cdot 100 \qquad \text{(ohne Modell)}$$

mit $\bar{Y}_i$ = Anteil der Probanden mit Merkmalsausprägung i der unabhängigen Variablen und Merkmalsausprägung „Denken im Beruf" an der Zahl der Probanden mit Merkmalsausprägung i

und $\bar{Y}$ = Anteil der Probanden mit Merkmalsausprägung „Denken im Beruf" an der Gesamtzahl der Probanden.

Die eingeklammerten Zahlen geben die Standardfehler an. Sie ergeben sich für die regressionsanalytischen Ergebnisse durch Multiplikation der

in (17) dargestellten Varianzen mit dem Faktor $\left(\frac{100}{\bar{Y}}\right)^2$.

Hierbei wurde angenommen, daß $\bar{Y}$ eine Konstante ist. Da der Standardfehler für $\bar{Y}$ bei 20372 Beobachtungen nur rund 0,35 % beträgt, werden die Ergebnisse durch diese Annahme nicht wesentlich verfälscht.

Die Standardfehler für die ohne Modell errechneten prozentualen Abweichungen wurden nach der Formel

$$\sqrt{\left(\frac{100}{\bar{Y}}\right)^2 \left[\text{Var}\ \bar{Y}_i + \text{Var}\ \bar{Y}\right]}$$

$$\text{mit} \quad \text{Var}\ \bar{Y}_i = \frac{\bar{Y}_i\,(1 - \bar{Y}_i)}{N_i}$$

$$\text{bzw.} \quad \text{Var}\ \bar{Y} = \frac{\bar{Y}\,(1 - \bar{Y})}{N}$$

berechnet.

Aus den in Tabelle 4 angegebenen Ergebnissen können unterschiedliche Schlußfolgerungen gezogen werden: So ergibt sich z. B. nach der herkömmlichen Analysemethode für die Merkmalsausprägung „Berufsschule" eine signifikant positive prozentuale Abweichung vom „Durchschnittseffekt" des Merkmals „Schulbildung" auf die abhängige Variable „Denken im Beruf", während sich bei der Regressionsanalyse, bei der der Einfluß des Merkmals „Geschlecht" eliminiert ist, keine signifikant positive Abweichung ergibt. Ein ähnliches Resultat erhält man für die Merkmalsausprägung „Realschule".

4.2.3 Erwartungstreue der geschätzten Koeffizienten

Es bleibt jetzt noch zu zeigen, daß die in (15) und (16) angegebenen linearen Abweichungen erwartungstreu sind. Nach dem in 2.2.2 angegebenen Satz 1 trifft dies zu, falls das Eindeutigkeitskriterium $\mathbf{p}' \cdot \mathbf{H} = \mathbf{p}'$ erfüllt ist. Durch Multiplikation der einzelnen Zeilen der Matrix **P** mit der ebenfalls im Anhang abgedruckten Matrix **H** kann obige Bedingung bestätigt werden. Die gewichteten Regressionskoeffizienten sind demnach erwartungstreu im Gegensatz zu der nur bedingten Erwartungstreue der mit Hilfe von a-priori-Restriktionen geschätzten Parameter.

5. Schlußbemerkungen

Der vorliegende Artikel gibt Hinweise, wie die bei einer multiplen Regressionsanalyse mit ausschließlich qualitativen Variablen auftretenden methodischen Probleme gelöst werden können. Je nach Genauigkeitsanforderung an die Ergebnisse können anspruchsvolle oder weniger anspruchsvolle Methoden verwendet werden.

Notwendige Bedingung für die Schätzbarkeit der Parameter ist die Lösung des Multikollinearitätsproblems. Begnügt man sich mit bedingt erwartungstreuen Schätzungen, so bietet sich zur Lösung des Multikollinearitätsproblems das Konzept der a-priori-Restriktionen an. Ist man besonders an erwartungstreuen Schätzungen interessiert, so ist es zweckmäßig, das Multikollinearitätsproblem mittels schätzbarer Funktionen zu lösen. Hierbei werden anstatt der individuellen Parameter geeignete Linearkombinationen geschätzt. Dieses Konzept bedient sich der verallgemeinerten Inversen einer singulären Matrix und führt immer zu einer erwartungstreuen Schätzung, falls eine solche existiert. Der Aufwand gegenüber dem Verfahren der a-priori-Restriktionen ist jedoch wesentlich größer.

Die im Zusammenhang mit den abhängigen qualitativen Variablen auftretenden Probleme der linearen Funktionsform sowie der Heteroskedastizität der Varianz der Störvariablen berühren nicht die Schätzbarkeit der Parameter. So erhält man — sofern das Multikollineari-

tätsproblem gelöst ist — Parameterschätzwerte, die im allgemeinen durchaus brauchbar sind. Die Anpassung an die empirischen Daten wird jedoch zunehmend schlechter, je näher die relative Häufigkeit für das Auftreten der untersuchten Merkmalsausprägung an den beiden Extremen Null oder Eins liegt. Insbesondere in diesen Fällen ist es zweckmäßig, die lineare Funktionsform durch eine logistische zu ersetzen.

Die Heteroskedastizität der Varianz der Störvariablen berührt die Effizienz der Parameterschätzwerte und beeinflußt damit die Signifikanztests für die Regressionskoeffizienten. Sie kann mittels der verallgemeinerten Methode der kleinsten Quadrate in der Regressionsanalyse berücksichtigt werden. Solange die merkmalsspezifischen relativen Häufigkeiten für das Auftreten der untersuchten Merkmalsausprägung jedoch nicht stark variieren, ist der Effizienzverlust bei einer Schätzung mit Hilfe der gewöhnlichen Methode der kleinsten Quadrate gering.

In jedem Fall steht mit der multiplen Regressionsanalyse ein multivariates Verfahren zur Verfügung, das unter Beachtung obiger Einschränkungen geeignet ist, lineare Abhängigkeiten auch zwischen qualitativen Variablen zu untersuchen.

6. Anhang

Matrix X′X

5 451	0	19	264	372	1 055	1 115	970	1 612	44	5 451
0	14 921	16	776	985	2 065	2 295	3 320	5 280	184	14 921
19	16	35	0	0	0	0	0	0	0	35
264	776	0	1 040	0	0	0	0	0	0	1 040
372	985	0	0	1 357	0	0	0	0	0	1 357
1 055	2 065	0	0	0	3 120	0	0	0	0	3 120
1 115	2 295	0	0	0	0	3 410	0	0	0	3 410
970	3 320	0	0	0	0	0	4 290	0	0	4 290
1 612	5 280	0	0	0	0	0	0	6 892	0	6 892
44	184	0	0	0	0	0	0	0	228	228
5 451	14 921	35	1 040	1 357	3 120	3 410	4 290	6 892	228	20 372

Matrix g-Inverse

0,004551	0,004346	–0,004458	–0,004398	–0,004403	–0,004416	–0,004413	–0,004393	–0,004394	0	0
0,004346	0,004395	–0,004369	–0,004353	–0,004382	–0,004379	–0,004379	–0,004384	–0,004384	0	0
–0,004458	–0,004369	0,032988	0,004391	0,004393	0,004399	0,004398	0,004389	0,004390	0	0
–0,004398	–0,004383	0,004391	0,005348	0,004387	0,004388	0,004388	0,004386	0,004387	0	0
–0,004403	–0,004382	0,004393	0,004387	0,005125	0,004389	0,004389	0,004387	0,004387	0	0
–0,004416	–0,004379	0,004399	0,004388	0,004389	0,004712	0,004391	0,004387	0,004387	0	0
–0,004413	–0,004379	0,004398	0,004388	0,004389	0,004391	0,004684	0,004387	0,004387	0	0
–0,004393	–0,004384	0,004389	0,004386	0,004387	0,004387	0,004387	0,004619	0,004386	0	0
–0,004394	–0,004384	0,004390	0,004387	0,004387	0,004387	0,004387	0,004386	0,004531	0	0
–1	–1	1	1	1	1	1	1	1	1	0
–1	–1	0	0	0	0	0	0	0	0	1

Matrix P

0,73243	–0,73243	0	0	0	0	0	0	0	0	0
–0,26757	0,26757	0	0	0	0	0	0	0	0	0
0	0	0,99828	–0,05105	–0,06661	–0,15351	–0,16739	–0,21058	–0,33831	–0,01119	0
0	0	–0,00172	0,94895	–0,06661	–0,15351	–0,16739	–0,21058	–0,33831	–0,01119	0
0	0	–0,00172	–0,05105	0,93339	–0,15351	–0,16739	–0,21058	–0,33831	–0,01119	0
0	0	–0,00172	–0,05105	–0,06661	0,84685	–0,16739	–0,21058	–0,33831	–0,01119	0
0	0	–0,00172	–0,05105	–0,06661	–0,15351	0,83261	–0,21058	–0,33831	–0,01119	0
0	0	–0,00172	–0,05105	–0,06661	–0,15351	–0,16739	0,78942	–0,33831	–0,01119	0
0	0	–0,00172	–0,05105	–0,06661	–0,15351	–0,16739	–0,21058	0,66169	–0,01119	0
0	0	–0,00172	–0,05105	–0,06661	–0,15351	–0,16739	–0,21058	–0,33831	0,98881	0
0	0	0	0	0	0	0	0	0	0	0

Matrix H

1	0	0	0	0	0	0	0	0	1	1
0	1	0	0	0	0	0	0	0	1	1
0	0	1	0	0	0	0	0	0	– 1	0
0	0	0	1	0	0	0	0	0	– 1	0
0	0	0	0	1	0	0	0	0	– 1	0
0	0	0	0	0	1	0	0	0	– 1	0
0	0	0	0	0	0	1	0	0	– 1	0
0	0	0	0	0	0	0	1	0	– 1	0
0	0	0	0	0	0	0	0	1	– 1	0
0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0