

Regressionsverfahren in der Bevölkerungswissenschaft *

Josef Brüderl
Institut für Soziologie
München

Langversion eines Artikels aus:
Müller, U.; B. Nauck und A. Diekmann (Hrsg.)(2000)
Handbuch der Demographie 1: Modelle und Methoden.
Berlin: Springer, S. 589-642.

Zitieren als:

Brüderl, J. (2000) Regressionsverfahren in der Bevölkerungswissenschaft.
München: unv. Manuskript. [Internetquelle]

*Für hilfreiche Hinweise danke ich Henriette Engelhardt, Axel Franzen, Wolfgang Ludwig-Mayerhofer, Peter Preisendörfer, Rainer Winkelmann, Rolf Ziegler, Klaus F. Zimmermann und einem anonymen Gutachter.

Inhaltsverzeichnis

1	Einleitung	1
2	Regressionsverfahren für metrische abhängige Variablen	2
2.1	Das lineare Regressionsmodell	2
2.1.1	Einfache Regression	2
2.1.2	Multiple Regression	6
2.2	Regressionsdiagnostik	12
2.2.1	Nicht-Linearität	12
2.2.2	Heteroskedastizität	15
2.2.3	Nicht-normalverteilte Fehler	17
2.2.4	Einflußreiche Datenpunkte	18
2.2.5	Anwendungsbeispiel: Scheidungsraten in 50 Ländern	22
2.3	Mehrgleichungssysteme	31
2.3.1	Schätzverfahren für nicht-rekursive Simultangleichungssysteme	35
2.3.2	Pfadanalyse	38
2.3.3	Anwendungsbeispiel: Scheidungsraten in 50 Ländern	39
2.4	Begrenzte abhängige Variablen	41
2.4.1	Zensierte abhängige Variable: Das Tobit-Modell	43
2.4.2	Stichprobenauswahl: Das Heckit-Verfahren	45
2.4.3	Exkurs: Fitmaße für Maximum-Likelihood Modelle	48
2.4.4	Anwendungsbeispiel: Scheidungsraten in 50 Ländern	50
2.5	Modelle für Zähldaten	51
2.5.1	Drei Zähldatenmodelle	52
2.5.2	Anwendungsbeispiel: Die Kinderzahl deutscher Frauen	55
3	Regressionsverfahren für qualitative abhängige Variablen	58
3.1	Binäre Modelle	59
3.1.1	Logit- und Probit-Modelle	59
3.1.2	Das bivariate Probit-Modell	65
3.1.3	Anwendungsbeispiel: Nichteheliche Lebensgemeinschaften und Scheidung	66
3.2	Multinomiale Modelle	71
3.2.1	Das multinomiale Logit-Modell	72
3.2.2	Exkurs: Entscheidungstheoretische Begründung	74
3.2.3	Das ordinale Probit-Modell	76
3.2.4	Anwendungsbeispiel: Die Einstellung zur Ehe	78
	Literatur	81

1 Einleitung

Multivariate Verfahren sind statistische Verfahren zur Analyse der Beziehungsstruktur *mehrerer* Variablen. Man kann multivariate Verfahren in drei große Gruppen unterteilen: i) Verfahren, die mehrdimensionale Datenstrukturen auf wenige Dimensionen reduzieren (multidimensionale Skalierung, Faktorenanalyse). ii) Verfahren, die Beobachtungen anhand mehrerer Variablen klassifizieren (Clusteranalyse, Diskriminanzanalyse). iii) Verfahren, die den Einfluß von unabhängigen Variablen auf die abhängigen Variablen erfassen (Regressionsanalyse, Varianzanalyse). In diesem Beitrag können aus Platzgründen nur die Verfahren der Regressionsanalyse behandelt werden. Der Grund für die Auswahl der Regressionsverfahren ist, daß die große Mehrzahl der bevölkerungswissenschaftlichen Forschungsarbeiten an der Identifikation „kausaler“ Effekte anhand nicht-experimenteller Daten interessiert ist, wozu Regressionsverfahren besonders geeignet sind. Insbesondere die Entwicklung von Regressionsverfahren auch für qualitative abhängige Variablen in den letzten Jahren hat zur Folge, daß für die meisten Fragestellungen der Sozialforschung die in diesem Beitrag vorgestellten Regressionsverfahren ausreichen. Bezüglich der weiteren multivariaten Verfahren muß auf die einschlägigen Lehrbücher verwiesen werden (z.B. Dillon und Goldstein, 1984; Fahrmeir und Hamerle, 1984; Tabachnick und Fidell, 1989; Morrison, 1990; Backhaus et al., 1994). Desweiteren werden in diesem Beitrag nur Regressionsverfahren für Querschnittdaten besprochen. Entsprechende Verfahren für Längsschnittdaten findet man in dem Beitrag von Diekmann in diesem Band.

Regressionsmodelle mit einer abhängigen und mehreren unabhängigen Variablen bezeichnet man als multiple Modelle. Von multivariaten Modellen spricht man, wenn mehrere abhängige Variablen gemeinsam betrachtet werden. Nach dieser Unterscheidung sind die meisten Modelle dieses Beitrags multipler Natur, nur an einigen Stellen werden wir multivariate Modelle behandeln. In der Literatur wird die Unterscheidung in multiple und multivariate Regressionsmodelle allerdings nur selten beachtet, weshalb auch in diesem Beitrag meist nur pauschal von multivariaten Modellen gesprochen wird.

Der Artikel ist an den Bedürfnissen von Anwendern ausgerichtet, die sich einen Überblick auch über die neueren Verfahren der Regressionsanalyse verschaffen wollen. Es wurde versucht, die wichtigsten Modelle in einfacher Art und Weise vorzustellen und deren Umsetzung anhand von ernsthaften Anwendungen aus der Bevölkerungswissenschaft zu demonstrieren. Ein besonderes Gewicht wird auf die Interpretation der Modellparameter gelegt. Die zugrundeliegende Schätztheorie dagegen wird nur am Rande gestreift. Der Leser, der einen „tieferen“ Überblick wünscht, sei auf die Literatur zu generalisierten linearen Modellen (McCullagh und Nelder, 1989; Fahrmeir und Tutz, 1994; Arminger, 1995) bzw. auf die ökonomische Literatur (Amemiya, 1985; Judge et al., 1985; Greene, 1993) verwiesen.

Vorausgesetzt werden die Grundzüge der statistischen Schätz- und Testtheorie, wie sie etwa in den einführenden Kapiteln von Greene (1993) gelegt werden. Ebenfalls vorausgesetzt werden elementare Grundregeln der Matrix-Algebra (gute Einführungen in die Matrix-Algebra findet man z.B. in Fahrmeir und Hamerle, 1984; Morrison, 1990).

2 Regressionsverfahren für metrische abhängige Variablen

In diesem Kapitel werden Regressionsmodelle für metrische abhängige Variablen vorgestellt. Die Abschnitte 2.1 und 2.3 behandeln die klassischen multiplen und multivariaten Regressionsmodelle. Die anderen drei Abschnitte gehen auf relativ neue Entwicklungen ein: In 2.2 werden Verfahren zur Regressionsdiagnostik vorgestellt, in 2.4 Verfahren für begrenzte abhängige Variablen und in 2.5 Verfahren für Zähldaten.

2.1 Das lineare Regressionsmodell

Das klassische, lineare Regressionsmodell ist sicher das am häufigsten verwendete Verfahren der multivariaten Analyse. Gründe dafür sind die einfache Interpretierbarkeit dieses Modells, seine relativ schwachen Annahmen und die einfache Schätzmöglichkeit. Entsprechend oft wurde das lineare Regressionsmodell in der statistischen Literatur besprochen (eine gute, anwendungsorientierte Einführung gibt Weisberg, 1985), weshalb sich die folgenden Ausführungen auf das Wesentliche beschränken sollen. Im ersten Unterabschnitt wird die einfache Regression mit nur einer unabhängigen Variable behandelt, im zweiten Unterabschnitt dann die multiple Regression.

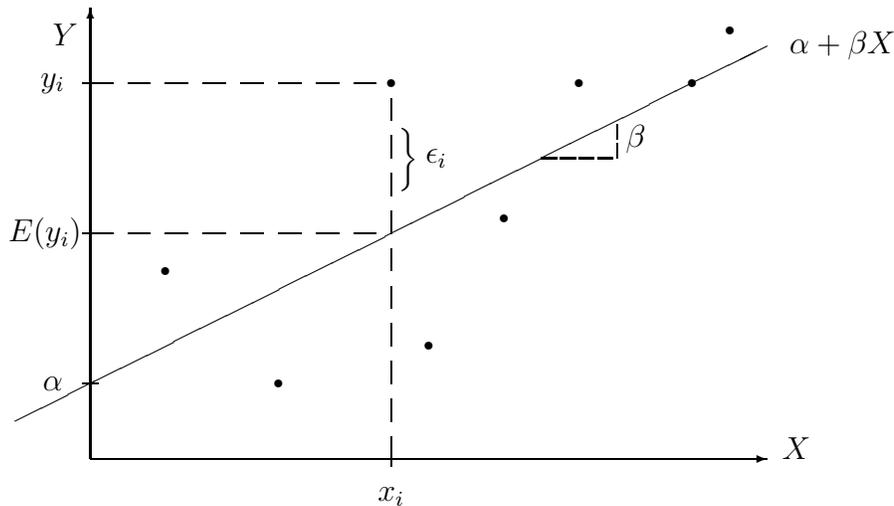
2.1.1 Einfache Regression

Angenommen wir wollen den Zusammenhang zwischen einer abhängigen Variable Y und *einer* unabhängigen Variable X (synonym: Regressor oder Kovariate) untersuchen. Beide Variablen seien metrisch. Wir beobachten bei n Personen die Datenpaare (y_i, x_i) . Das klassische Regressionsmodell geht nun davon aus, daß diese Daten mittels eines linearen Modells beschreibbar sind:

$$\bullet \text{ A1: } y_i = \alpha + \beta x_i + \epsilon_i, \quad i = 1, \dots, n.$$

α und β sind zu schätzende Parameter (Regressionskoeffizienten) und ϵ ist ein Fehlerterm. A1 enthält insbesondere folgende Annahmen: Die Beziehung zwi-

Abbildung 1: Das lineare Regressionsmodell



schen X und Y ist linear und die Parameter α und β sind identisch für alle Beobachtungen.

Das *Streudiagramm* in Abbildung (1) gibt eine graphische Darstellung des Modells. Die Punkte repräsentieren jeweils eine Beobachtung. Die Gerade ist die (Modell-)Regressionsgerade mit Achsenabschnitt α und Steigung β . α gibt den Y -Wert an, wenn $X = 0$. β gibt an, um wieviele Einheiten sich Y verändert, wenn X um eine Einheit steigt. Meist werden die Daten natürlich nicht exakt auf der Regressionsgeraden liegen, so wie in Abbildung (1) z.B. Beobachtung i . Der laut Modell zu erwartende Wert ($E(y_i|x_i) = \alpha + \beta x_i$), stimmt nicht mit dem beobachteten Wert (y_i) überein. Die Differenz ist der Fehler ϵ_i .

Über die stochastischen Eigenschaften dieses Fehlerterms macht man einige weitere *Annahmen*:

- A2: $E(\epsilon_i) = 0$, für alle i ; im „Mittel“ ist der Fehler null
- A3: $V(\epsilon_i) = \sigma^2$, für alle i ; die Fehlervarianz ist konstant (Homoskedastizität)
- A4: $Cov(\epsilon_i, \epsilon_j) = 0$, für alle $i \neq j$; die Fehlerkovarianzen sind null (keine Autokorrelation)
- A5: $Cov(x_i, \epsilon_j) = 0$, für alle i und j ; Regressor und Fehler sind unkorreliert

A5 wird oft verschärft, indem man annimmt, daß X nicht-stochastisch ist (die Werte der unabhängigen Variable sind fest, wie es etwa bei einem Experiment der Fall ist). Geht man von einem stochastischen Regressor aus, was bei quasi-experimentellen Designs sicher realistischer ist, so impliziert A5 unter anderem:

der Regressor darf keine Meßfehler enthalten und er darf mit keinen weiteren unbeobachteten Variablen (die ja laut A1 im Fehlerterm zusammengefaßt sind) korreliert sein.

Zusammen mit A1 sind dies die Annahmen des klassischen, linearen Regressionsmodells. Sie sind nötig, damit die Parameterschätzungen gewissen Anforderungen genügen (s.u.). Sind sie nicht erfüllt, müssen die Schätzverfahren modifiziert werden (s. Abschnitte 2.2 bis 2.5). Will man Hypothesen über die Parameter des Modells A1 testen, so ist eine weitere Annahme nötig:

- A6: $\epsilon_i \sim N(0, \sigma^2)$; die Fehler sind normalverteilt

Es ist zu beachten, daß die Normalverteilungsannahme *nicht* zur Schätzung der Parameter erforderlich ist. Da in der Sozialforschung aber praktisch immer Hypothesentests durchgeführt werden, muß auch A6 üblicherweise gelten.

Dieses Modell enthält die unbekannt Parameter α , β und σ^2 . Anhand der beobachteten Daten können die Parameter geschätzt werden. Die *Schätzer* notieren wir mit $\hat{\alpha}$, $\hat{\beta}$ und $\hat{\sigma}^2$. Insbesondere $\hat{\beta}$ interessiert den Sozialforscher, weil dieser Schätzer Auskunft über den Zusammenhang von X und Y gibt. Zur Schätzung wird im klassischen Regressionsmodell die Methode der kleinsten Quadrate („Ordinary Least Squares“, OLS) eingesetzt. Dazu werden $\hat{\alpha}$ und $\hat{\beta}$ so bestimmt, daß die Summe der quadrierten Residuen (der geschätzten Fehler) minimiert wird:

$$\min_{\hat{\alpha}, \hat{\beta}} \sum_i \hat{\epsilon}_i^2 = \min_{\alpha, \beta} \sum_i (y_i - \alpha - \beta x_i)^2.$$

Graphisch interpretiert bedeutet dies, daß $\hat{\alpha}$ und $\hat{\beta}$ so gewählt werden, daß die resultierende Regressionsgerade die Summe der quadrierten Residuen minimiert. Daraus ergeben sich folgende OLS-Schätzer für α , β und σ^2 :

$$\begin{aligned} \hat{\alpha} &= \bar{y} - \hat{\beta} \bar{x}, \\ \hat{\beta} &= \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}, \\ \hat{\sigma}^2 &= \frac{\sum_i \hat{\epsilon}_i^2}{n - 2}. \end{aligned}$$

$\hat{\beta}$ ergibt sich aus dem um die Mittelwerte korrigiertem Kreuzprodukt von y_i und x_i , dividiert durch die quadrierte Summe der korrigierten x_i . Hiermit kann sofort $\hat{\alpha}$ berechnet werden. Mit diesen beiden Schätzern können dann die Residuen und schließlich $\hat{\sigma}^2$ errechnet werden ($n - 2$ ist die Zahl der Freiheitsgrade der Residuen: Zahl der Beobachtungen minus Zahl der Modellparameter, die zu ihrer Berechnung nötig sind).

Die OLS-Schätzer haben bei Gültigkeit von A1-A5 gewisse wünschenswerte Eigenschaften: Sie sind unverzerrt (erwartungstreu: $E(\hat{\beta}) = \beta$) und in der Klasse der

linearen, unverzerrten Schätzer die mit der kleinsten Stichprobenvarianz (*Gauss-Markov Theorem*). Will man einen linearen und unverzerrten Schätzer verwenden, so besagt das Gauss-Markov Theorem salopp ausgedrückt, daß die OLS-Schätzer in dieser Klasse die präzisesten sind. Außerdem sind sie bei Gültigkeit der Normalverteilungsannahme die Maximum-Likelihood (ML) Schätzer und besitzen somit auch deren Eigenschaften.

Neben der Schätzung der Parameter ist es weiterhin wichtig zu wissen, wie gut das Modell die Daten anpaßt. Darüber gibt das *Bestimmtheitsmaß* R^2 Auskunft. Grundlage für die Berechnung von R^2 ist die sogenannte Varianzzerlegung: Die Varianz von Y ergibt sich aus der Summe der durch das Modell erklärten Varianz und der Residuen-Varianz. R^2 wird nun als Quotient von erklärter zu gesamter Varianz definiert. Analog ist die Definition über eins minus den Quotient von unerklärter zu gesamter Varianz:

$$R^2 = 1 - \frac{\sum_i \hat{\epsilon}_i^2}{\sum_i (y_i - \bar{y})^2}.$$

Ist $R^2=0$, so bedeutet dies, daß das Modell nichts zur Erklärung der Varianz von Y beiträgt. $R^2=1$ dagegen zeigt an, daß das Modell die Daten vollkommen beschreibt (sie liegen alle auf der Regressionsgeraden). $R^2=0,3$ etwa ist folgendermaßen zu interpretieren: 30% der Varianz von Y sind durch das Modell erklärt, oder andersherum, 70% sind durch unbekannte Faktoren (die in den Residuen enthalten sind) bestimmt. In der Praxis ist ein $R^2=0,3$ bei Individualdaten bereits als guter Fit anzusehen. Höhere R^2 können nur erzielt werden, wenn eine gute Theorie vorliegt, die die wichtigsten Faktoren benennt. Dennoch sind bei der unvermeidlichen Individualität demographischer Entscheidungen Bestimmtheitsmaße nahe eins wohl nie erreichbar (glücklicherweise). Ferner ist zu beachten, daß R^2 den Fit eines *linearen* Modells beschreibt. Ein niedriges R^2 kann somit auch auf die Nicht-Linearität der Beziehung hindeuten (s.u.).

Schließlich will man noch *Hypothesen über die Schätzer testen*. Im einfachsten Fall wird die Nullhypothese $H_0 : \beta = \beta^0$ gegen die Alternative $H_1 : \beta \neq \beta^0$ getestet (zweiseitiger Test). Dazu muß zuerst der Standardfehler von $\hat{\beta}$ berechnet werden ($\hat{\sigma}_{\hat{\beta}} = \hat{\sigma} / \sqrt{\sum_i (x_i - \bar{x})^2}$). Dann wird die Testgröße $(\hat{\beta} - \beta^0) / \hat{\sigma}_{\hat{\beta}}$ errechnet. Sie ist t-verteilt mit $n - 2$ Freiheitsgraden. Die Nullhypothese wird abgelehnt, wenn der Betrag der Testgröße größer als der kritische Wert der t-Verteilung ist (bei gegebenem Signifikanzniveau). Meist jedoch ist der Sozialforscher nur daran interessiert, ob X überhaupt einen Einfluß auf Y ausübt ($H_0 : \beta = 0$). Dann lautet die Testgröße (der sogenannte t-Wert) $t = \hat{\beta} / \hat{\sigma}_{\hat{\beta}}$. Auf dem 95%-Niveau bei großen Fallzahlen ($n > 500$) beträgt der kritische Wert 1,96. Ist somit $|t| > 1,96$, können wir die Nullhypothese ablehnen und sprechen von einem signifikanten Effekt der Variable X .¹

¹Wenn im folgenden von „Signifikanz“ gesprochen wird, so ist natürlich „statistische Signifikanz“ gemeint, nicht „inhaltliche Signifikanz“. Diese so grundlegende wie wichtige Unterschei-

2.1.2 Multiple Regression

Das einfache Regressionsmodell ist kein multivariates Verfahren, da nur der lineare Zusammenhang zwischen einer abhängigen und *einer* unabhängigen Variable untersucht wird. Das Modell ist jedoch problemlos um weitere unabhängige Variablen erweiterbar. Man spricht dann von multipler Regression:

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \dots + \beta_k x_{ik} + \epsilon_i.$$

Dieses Modell beschreibt eine $(k - 1)$ -dimensionale Regressionsebene, um die herum die Beobachtungen mehr oder weniger stark streuen. Es kann in Matrix-Schreibweise notiert werden als

$$y_i = \boldsymbol{\beta}' \mathbf{x}_i + \epsilon_i,$$

wobei $\boldsymbol{\beta}$ ein k -elementiger Spaltenvektor der zu schätzenden Parameter und \mathbf{x}_i ebenfalls ein k -elementiger Spaltenvektor der unabhängigen Variablen ist (das erste Element von \mathbf{x}_i ist eine Eins, was der Konstanten entspricht).

Ein β_j kann in diesem Modell ebenso wie bei der einfachen Regression als Effekt der Variable X_j interpretiert werden. Der wichtigste Unterschied zum einfachen Regressionsmodell ist, daß ein bestimmter Regressionskoeffizient β_j nun den Effekt von X_j *bei Konstanthaltung* der anderen unabhängigen Variablen wiedergibt (man sagt auch „deren Effekte wurden herauspartialisiert“). Dies ist generell der Hauptvorteil multivariater Verfahren: Man erhält Effekte, die von den Einflüssen aller weiteren im Modell enthaltenen Kovariaten befreit sind. Multivariate Regressionsmodelle ermöglichen es somit, auf einfache Art und Weise für die störenden Einflüsse von Drittvariablen zu kontrollieren.

Zur genaueren Interpretation der Regressionskoeffizienten β_j ist es hilfreich, den bedingten Erwartungswert der abhängigen Variable gegeben \mathbf{x}_i zu betrachten:

$$E(y_i | \mathbf{x}_i) = \boldsymbol{\beta}' \mathbf{x}_i.$$

Nun kann man die Frage stellen, um wieviel sich dieser bedingte Erwartungswert verändert, wenn man X_j um eine Einheit erhöht (*Einheitseffekt*). Um diesen Einheitseffekt zu erhalten, muß man vom bedingten Erwartungswert an der Stelle $x_j + 1$ den bedingten Erwartungswert an der Stelle x_j abziehen. Es ergibt sich, daß der Einheitseffekt im multiplen Regressionsmodell gleich β_j ist, was eine anschauliche Interpretation der Regressionskoeffizienten ermöglicht. Eleganter kann

die Interpretation wird leider selbst in der Forschungsliteratur kaum beachtet (Deal und Anderson, 1995; McCloskey und Ziliak, 1996). Darum sei betont, daß eine Ergebnisinterpretation, die nur auf die t -Werte abstellt, ungenügend ist. Die Größe und Bedeutsamkeit der Regressionseffekte ist wichtiger.

man den Einheitseffekt näherungsweise bestimmen, indem man den bedingten Erwartungswert nach X_j ableitet (*Marginaleffekt*):

$$\frac{\partial E(y|\mathbf{x})}{\partial x_j} = \beta_j.$$

Offensichtlich ist der Marginaleffekt identisch mit dem Einheitseffekt, weil das Modell linear ist. Weiter unten werden wir sehen, daß diese nützliche Eigenschaft in nicht-linearen Regressionsmodellen nicht mehr gilt.

Zur weiteren Behandlung des Modells wollen wir die Regressionsgleichung nicht nur für eine, sondern für alle Beobachtungen niederschreiben. \mathbf{X} sei die $n \times k$ Matrix der unabhängigen Variablen. $\boldsymbol{\beta}$ ist ein k -elementiger Spaltenvektor der Parameter. \mathbf{y} und $\boldsymbol{\epsilon}$ sind n -elementige Spaltenvektoren der abhängigen Variable bzw. der Fehler. Damit kann das multiple Regressionsmodell geschrieben werden als

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

Die Modellannahmen sind identisch zur einfachen Regression. A2, A3, A4 und A6 lauten in Matrix Notation: $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$. Die Fehler sind unabhängig, identisch normalverteilt mit Erwartungswert 0 und konstanter Varianz σ^2 . $\sigma^2 \mathbf{I}$ ist die Kovarianzmatrix der Fehler, wobei \mathbf{I} die Einheitsmatrix ist (eine Matrix mit Einsen auf der Diagonalen und Nullen sonst). Weiterhin muß analog zu A5 angenommen werden, daß die Regressoren nicht mit dem Fehlerterm korreliert sind (oder \mathbf{X} nicht-stochastisch ist). Die einzige zusätzliche Annahme gegenüber dem einfachen Regressionsmodell ist, daß \mathbf{X} vollen Spalten-Rang hat (die unabhängigen Variablen dürfen nicht linear abhängig sein). Als *OLS-Schätzer* von $\boldsymbol{\beta}$ ergibt sich unter diesen Annahmen:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y}.$$

Das *Bestimmtheitsmaß* R^2 kann analog zum einfachen Regressionsmodell errechnet werden und gibt Auskunft, wieviel der Varianz von Y durch die Regressoren „erklärt“ ist. Fügt man einen weiteren Regressor hinzu, so ist das Bestimmtheitsmaß des erweiterten Modells mindestens genauso groß wie zuvor. Ist allerdings die Erklärungskraft der hinzugefügten Variable, gegeben die bereits im Modell enthaltenen Variablen, gering, so wird sich R^2 nur minimal erhöhen (was man sich bei Verfahren zur Variablenselektion zunutze macht, s.u.). Das Hinzufügen weiterer Variablen verbessert das Modell somit nur, wenn diese Variablen einen „eigenständigen“ Erklärungsbeitrag leisten. Häufig wird empfohlen, ein „adjustiertes“ R^2 zu berichten, bei dem die Zahl der Freiheitsgrade in die Berechnung Eingang findet. Es kann bei Hinzufügen einer weiteren, erklärungschwachen Variable auch kleiner werden. Das Argument für diese Empfehlung ist, daß R^2 durch die Einführung vieler Variablen beliebig nahe an eins gebracht werden kann, und

man deshalb Modelle mit vielen Variablen „bestrafen“ muß. Dieses Argument ist offensichtlich nicht schlüssig. Außerdem kann das adjustierte R^2 nicht mehr im Sinne von erklärter Varianz interpretiert werden, weshalb es nur für die Variablenselektion eingesetzt werden sollte.

Die Kovarianzmatrix der Schätzer lautet:

$$\hat{V}(\hat{\beta}) = \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}, \quad \text{wobei} \quad \hat{\sigma}^2 = \frac{\sum_i \hat{\epsilon}_i^2}{n - k}.$$

Tests von Hypothesen über einzelne Koeffizienten können mit diesen Größen wie im vorigen Unterabschnitt beschrieben durchgeführt werden (t-Test). Im multiplen Fall sind auch Tests über komplexe Linearkombinationen von Parametern möglich (F-Test). Ein Beispiel für einen F-Test ist der Test auf die Nullhypothese, daß alle Koeffizienten (außer der Konstanten) gleich null sind. Die Testgröße für diesen Fall läßt sich unter anderem aus R^2 errechnen:

$$F = \frac{R^2/(k - 1)}{(1 - R^2)/(n - k)}.$$

Sie ist F-verteilt mit $k - 1$ und $n - k$ Freiheitsgraden. Ist der F-Wert größer als der entsprechende kritische Wert aus der F-Verteilung, so spricht man oft salopp von der „Signifikanz des Regressionsmodells“.²

Bisher gingen wir davon aus, daß die Regressoren metrische Variablen sind. Doch auch nicht-metrische Regressoren können problemlos im linearen Regressionsmodell berücksichtigt werden. Dies geschieht mittels sogenannter *Dummy-Variablen* (ausführlich behandelt dieses Thema Hardy, 1993). Hat ein qualitativer Regressor drei Kategorien, so kann man drei Dummies bilden, wobei eine Dummy nur dann eins ist, wenn eine Beobachtung in die zugehörige Kategorie fällt, sonst ist sie immer null. Enthält das Regressionsmodell eine Konstante, so muß man eine Referenzkategorie wählen und die zugehörige Dummy aus dem Modell nehmen (sonst sind die Regressoren linear abhängig). Ein Beispiel ist eine Regression auf das Einkommen (y_i), mit den Regressoren Alter (x_i) und Schultyp (Hauptschule, Realschule und Gymnasium). Wählt man „Hauptschule“ als Referenzgruppe, so benötigt man eine Dummy für „Realschule“ (D_{i1}) und eine weitere für „Gymnasium“ (D_{i2}). Die Einkommensregression lautet damit:

$$y_i = \beta_1 + \beta_2 x_i + \delta_1 D_{i1} + \delta_2 D_{i2} + \epsilon_i.$$

²Eine verbreitete Fragestellung ist, ob sich Effekte verändern, wenn man dem Regressionsmodell weitere Variablen hinzufügt. Man könnte etwa bei der im folgenden Absatz beschriebenen Einkommensregression daran interessiert sein, ob sich die Bildungseffekte signifikant verändern, wenn man die Variable „Intelligenz“ berücksichtigt. Wäre dies der Fall, so müßte man daraus schließen, daß Bildungseffekte in Einkommensregressionen ohne die Intelligenzvariable verzerrt sind (man spricht vom sogenannten „ability-bias“). Tests für diese Art von Fragestellungen findet man bei Clogg et al. (1995).

δ_1 gibt an, um wieviel im Schnitt ein Realschüler mehr verdient als ein Hauptschüler (bei Konstanz des Alters!). δ_2 spiegelt den Einkommensvorsprung der Gymnasiasten gegenüber den Hauptschülern wieder. Für jeden dieser beiden Koeffizienten kann ein t-Wert errechnet werden, der dann Aufschluß darüber gibt, ob die Einkommensdifferenz zwischen den entsprechenden Schultypen signifikant ist. Die Signifikanz der Dummy-Effekte hängt natürlich entscheidend von der Wahl der Referenzgruppe ab. Es macht deshalb keinen Sinn, nur die signifikanten Effekte als bedeutsam herauszustellen. Sinnvoll ist nur eine Aussage über die Bedeutsamkeit der kategorialen Variable insgesamt. Dazu kann der folgende F-Test eingesetzt werden. Will man etwa feststellen, ob der Schultyp insgesamt einen Einfluß auf das Einkommen hat, so muß die Nullhypothese $\delta_1 = \delta_2 = 0$ getestet werden. Bezeichnen wir die Regression ohne die Dummies mit 1, die mit den Dummies mit 2, so ergibt sich die zugehörige F-Testgröße als

$$F = \frac{(R_2^2 - R_1^2)/(k_2 - k_1)}{(1 - R_2^2)/(n - k_2)}.$$

Überschreitet diese Testgröße den kritischen Wert aus der F-Verteilung, so ist der Einfluß der kategorialen Variable insgesamt signifikant. Wie bedeutsam der Einfluß der kategorialen Variable ist, kann man sehr anschaulich aus dem Zuwachs an erklärter Varianz $R_2^2 - R_1^2$ ersehen.

Schließlich kann der Fall auftreten, daß die verwendete Theorie *Interaktionseffekte* voraussagt (ausführlich behandeln dieses Thema Jaccard et al., 1990). Der häufigste Fall ist, daß sich die Effekte bestimmter Variablen in verschiedenen Gruppen unterscheiden. Zum Beispiel könnte man vermuten, daß die Bildungserträge für Männer und Frauen unterschiedlich sind (S_i , 0=Frau, 1=Mann). Dazu nimmt man S_i in das Modell auf. Der Parameter dieser Variable gibt dann den Einkommensvorsprung (wenn er positiv ist) der Männer an. Zusätzlich multipliziert man S_i mit den beiden Schultyp-Dummies und nimmt auch diese Interaktionsvariablen in das Modell mit auf. Die zugehörigen Effekte zeigen dann, ob die Männer andere Bildungserträge haben als die Frauen. Ebenso kann man eine multiplikative Alters-Geschlechts Interaktionsvariable bilden.³ Deren Effekt zeigt, ob das Alters-Einkommensprofil der Männer anders verläuft als bei den Frauen. Um die Signifikanz von Interaktionseffekten zu testen, gibt es zwei verschiedene Möglichkeiten.

(1) *Fehlervarianz gleich*: Wenn man von der Annahme ausgeht, daß Männer und Frauen identische Fehlervarianzen haben ($\sigma_m^2 = \sigma_f^2$), so kann man die Daten der beiden Gruppen „poolen“ und *eine* Einkommensregression mit dem Geschlechts-Haupteffekt und allen Interaktionseffekten schätzen. Für jeden einzelnen Interaktionseffekt läßt sich sodann ein t-Wert berechnen und der Signifikanztest durchführen. Die Signifikanz aller Interaktionen zusammen läßt sich mit

³Die multiplikative Verknüpfung der beiden Variablen ist zwar die am häufigsten zu findende, ist aber nicht zwingend. Auch nicht-multiplikative Interaktionen sind möglich.

einem F-Test, wie er oben beschrieben wurde, ermitteln. R_2^2 ist dann das Bestimmtheitsmaß in dem unrestringierten Modell mit allen Interaktionen, R_1^2 ist das Bestimmtheitsmaß des restringierten Modells ohne Interaktionen.

(2) *Fehlervarianz verschieden*: In diesem Fall muß man *getrennte* Regressionen für Frauen und Männer schätzen. In großen Stichproben kann die Signifikanz der Differenz einzelner Koeffizienten mit folgender z-Teststatistik überprüft werden (s. Clogg et al., 1995):

$$z = \frac{\hat{\beta}_m - \hat{\beta}_f}{\sqrt{\hat{V}(\hat{\beta}_m) + \hat{V}(\hat{\beta}_f)}}.$$

Unter der Nullhypothese der Gleichheit der Koeffizienten ist diese Teststatistik standardnormalverteilt. Die Berechnung des Standardfehlers der Koeffizientendifferenz (Ausdruck unter der Wurzel) beruht allerdings auf der Annahme, daß die beiden Stichproben unabhängig sind. Bei Zufallsstichproben aus großen Bevölkerungen sollte diese Annahme unproblematisch sein. Sie wäre aber sicher verletzt, wenn wir die Daten unseres Beispiels aus der Befragung von Ehepaaren gewonnen hätten. Zum Test der Signifikanz aller Differenzen zusammen wurde wiederum ein F-Test vorgeschlagen (Chow-Test; s. Chow, 1960). Die F-Testgröße läßt sich am einfachsten aus den Summen der quadrierten Residuen errechnen. Bezeichnen wir mit RSS_2 die Summe der aufsummierten quadrierten Residuen der getrennten Regressionen, mit RSS_1 die aufsummierten quadrierten Residuen der restringierten (gepoolten) Regression, so lautet die F-Testgröße:

$$F = \frac{(RSS_1 - RSS_2)/k_1}{RSS_2/(n_1 + n_2 - 2k_1)}.$$

Liegt diese Testgröße über dem kritischen Wert aus der F-Verteilung, so unterscheiden sich die Effekte von Männern und Frauen signifikant.

Neben den üblichen Schätz- und Testproblemen steht man bei jeder statistischen Datenanalyse vor dem Problem, welche Variablen das Modell enthalten soll. Betreibt man Kausalanalyse, so gibt es nur eine sinnvolle Antwort: Theoretische Überlegungen müssen die *Variablenselektion* bestimmen. Alle anderen Vorgehensweisen produzieren nur nicht interpretierbare ad-hoc Resultate. Will man dagegen seine Modelle für prognostische Zwecke einsetzen, so mag es sinnvoll erscheinen, ein „bestes“ Modell zu suchen. Hierfür gibt es verschiedene Variablen-Selektionsverfahren. Die meisten beruhen auf schrittweisem Vorgehen, wo z.B. bei jedem Schritt die Variable in das Modell eingeführt wird, die von den verbliebenen Variablen den höchsten R^2 -Zuwachs erzeugt. Dies wird solange fortgesetzt, bis alle noch nicht im Modell enthaltenen Variablen einen R^2 -Zuwachs aufweisen, der unter einer vorgegebenen Schwelle liegt. Alternativ kann auch ein Algorithmus verwendet werden, der unter allen möglichen Variablenkombinationen das Modell sucht, das ein adjustiertes R^2 -Maß maximiert. Nach einer Variablenselektion sind

natürlich Signifikanztests sinnlos, da ja bewußt nur die stärksten Prädiktoren selektiert wurden. Dasselbe gilt für „manuell“ selektierte Modelle (z.B. wenn man nur Variablen, deren t-Wert größer als eins ist, in das Modell nimmt).

Eine vielversprechende Methode der Modellselektion, die in der bayesianischen Wahrscheinlichkeitstheorie begründet ist, schlägt Raftery (1995) vor. Ein Algorithmus bestimmt die Modelle, die das BIC-Maß (s. Abschnitt 2.4.3) minimieren. Über alle diese Modelle, die sich im sogenannten „Fenster von Occam“ befinden, wird der mittlere Effekt einer jeden Variable und der dazugehörige Standardfehler errechnet. Diese Methode ermöglicht mithin auch nach einer Modellselektion die Durchführung von Signifikanztests.

In diesem Zusammenhang taucht manchmal noch ein weiteres Problem auf: Die im Modell enthaltenen Variablen können hoch miteinander korreliert sein. Man spricht von *Multikollinearität*. Korrelieren zwei Regressoren vollständig, so liegt lineare Abhängigkeit vor und $(\mathbf{X}'\mathbf{X})^{-1}$ existiert nicht. Dieses Problem tritt bereits bei nicht perfekter Korrelation auf (bei $r > 0,99$ wird es kritisch). Bei extremer Multikollinearität sind somit die OLS-Schätzer nicht berechenbar. Bei geringerer Multikollinearität sind die OLS-Schätzer schätzbar und auch konsistent, allerdings erhöht Multikollinearität die Standardfehler der OLS-Schätzer, die Schätzungen sind weniger „präzise“. Dies sieht man, wenn man die geschätzte Varianz für $\hat{\beta}_j$ schreibt als:

$$\hat{V}(\hat{\beta}_j) = \frac{\hat{\sigma}^2}{(n-1)\sigma_{x_j}^2} \frac{1}{1-R_j^2},$$

wobei R_j^2 das Bestimmtheitsmaß einer Regression aller anderen Regressoren auf X_j ist. Korreliert X_j hoch mit den anderen Kovariaten (R_j^2 nahe eins), so wird der zweite Faktor sehr groß (Varianz-Inflations-Faktor, VIF) und der Schätzfehler wächst an.⁴ Ist z.B. $R_j=0,9$, so ist $\sqrt{\text{VIF}}=2,29$, der Standardfehler des Schätzers erhöht sich um etwas mehr als das Doppelte und der t-Wert wird halbiert. Da allerdings solch hohe Korrelationen unter den Regressoren bei Individualdaten eher selten sind, stellt Multikollinearität im Normalfall kaum ein Problem dar. In der Forschungsliteratur wird ihr dennoch große Aufmerksamkeit zuteil, weil viele Forscher auf der „Jagd nach signifikanten Ergebnissen“ hier eine Möglichkeit sehen, das Resultat zu verbessern. Demgegenüber ist festzuhalten, daß nur bei extremen Multikollinearitäten Gegenmaßnahmen ergriffen werden müssen. Betroffene Variablen einfach wegzulassen, ist keine gute Lösung. Besser erscheint es, aus den multikollinearen Variablen einen Index zu bilden, denn meist werden diese Variablen sowieso ähnliche Konstrukte messen.

⁴Es ist zu beachten, daß nicht die bivariaten Korrelationen unter den Regressoren, sondern die *multiple* Korrelation entscheidend ist. $1/\text{VIF}$ wird von manchen Programmen als „Tolerance“ ausgewiesen.

2.2 Regressionsdiagnostik

Wie bei jedem statistischen Modell, beruht auch die Konsistenz der OLS-Schätzer auf der Gültigkeit der getroffenen Annahmen. Deshalb empfiehlt es sich, die Gültigkeit dieser Annahmen zu testen. In diesem Abschnitt sollen Verfahren vorgestellt werden, mit denen Annahmeverletzungen diagnostiziert werden können. Dabei können nur besonders einfache und wichtige Diagnoseverfahren berücksichtigt werden. Der Leser findet in jedem Ökonometrie-Lehrbuch viele weitere Verfahren. Speziell mit dieser Thematik befassen sich z.B. Krämer und Sonnberger (1986), Chatterjee und Hadi (1988) und Fox (1991). Moderne Regressionsdiagnostik ist ohne graphische Verfahren nicht mehr vorstellbar. Überblicke zu graphischen Diagnoseverfahren findet man bei Cook und Weisberg (1994) und Schnell (1994).

Zuerst werden diagnostische Verfahren zur Aufdeckung von Nicht-Linearität diskutiert. Nicht-Linearität verletzt A1 (und als Folge A2), weil das Modell fehlspezifiziert ist. Im zweiten Unterabschnitt befassen wir uns mit Heteroskedastizität (A3 nicht gültig). Bei Vorliegen von Heteroskedastizität sind zwar die Koeffizientenschätzer nicht verzerrt, aber sie sind ineffizient. Außerdem sind die geschätzten Varianzen der Regressionskoeffizienten verzerrt und damit die üblichen Signifikanztests ungültig. Im dritten Unterabschnitt werden Verletzungen der Normalverteilungsannahme (A6) diskutiert. Sowohl Koeffizienten- als auch Varianzschätzer sind in diesem Fall asymptotisch unverzerrt, sie sind aber nicht mehr die ML-Schätzer. Schließlich befassen wir uns mit einflußreichen Datenpunkten, die die Ergebnisse einer Regression stark beeinflussen können.

Auch die Verfahren, die später in den Abschnitten (2.3) bis (2.5) vorgestellt werden, können als „Heilung“ von Annahmeverletzungen betrachtet werden. Mehrgleichungssysteme führen dazu, daß Regressoren und Fehler korreliert sind (Verletzung von A5). Begrenzte abhängige Variablen und Zähldaten verletzen die Normalverteilungsannahme. Für diese Fälle existieren spezielle Schätzverfahren, die in den entsprechenden Abschnitten besprochen werden. Mit Autokorrelation (A4) befassen wir uns im folgenden nicht, da dieses Problem überwiegend bei Zeitreihendaten auftritt, aber nicht bei den in der Demographie meist verwendeten Individualdaten (s. z.B. Greene, 1993: Kap. 15). Ebenso unberücksichtigt bleiben die inzwischen verfügbaren Tests auf allgemeine Fehlspezifikation des Modells (s. hierzu Long und Trivedi, 1992).

2.2.1 Nicht-Linearität

Eine oft nicht mehr bewußt wahrgenommene Annahme des Regressionsmodells ist die Linearität der Beziehung zwischen unabhängigen und abhängiger Variable. Trotzdem wird gerade diese Annahme häufig verletzt sein. Die Anwender

hoffen meist nur, daß eine lineare Funktion eine gute Näherung ist. Solche Gewohnheiten können aber in die Irre führen, weshalb es sich immer empfiehlt, die Linearität der Beziehung zu testen. Nicht-Linearität in der Beziehung von Y und X führt insbesondere dazu, daß die Beobachtungen in systematischer Art und Weise von der Regressionsebene abweichen (die Fehler haben nicht mehr einen Erwartungswert von null; Verletzung von A2).

Viele nicht-lineare Beziehungen können durch Hinzufügen von Potenzen höherer Ordnung approximiert werden. Der häufig angewandte *RESET-Test* („Regression Specification Error Test“) beruht auf dieser Idee. Vermutet man z.B. eine quadratische Beziehung zwischen Y und X , also

$$y_i = \beta_1 + \beta_2 x_i + \beta_3 x_i^2 + \epsilon_i,$$

so ist ein Test auf $H_0 : \beta_3 = 0$ ein Linearitätstest.⁵

Der RESET-Test setzt allerdings voraus, daß man zumindest eine ungefähre Vorstellung über die funktionale Form der Beziehung hat. Dazu kann man *Residuen-Plots* einsetzen, wobei man für alle Regressoren die Residuen $\hat{\epsilon}_i$ gegen x_{ij} aufträgt. Abbildung (2) enthält ein Streudiagramm und einen Residuen-Plot (für eine einfache Regression; im multiplen Fall geben diese Diagramme nur Aufschluß über die marginale, nicht die partielle Beziehung, weshalb man im multiplen Fall Partielle-Residuen-Plots verwendet; s.u.). In diesem speziellen Fall liegt eine konvexe Beziehung zwischen Y und X vor, die mittels eines linearen und quadratischen X -Terms gut approximiert werden kann. Bei komplexeren Mustern kann es nötig werden, weitere Potenzen höherer Ordnung in das Modell aufzunehmen.

Aber nicht nur Potenzfunktionen sind zur Beschreibung nicht-linearer Beziehungen geeignet. So könnte das Muster von Abbildung (2) auch mittels einer *Exponentialfunktion* modelliert werden:

$$y_i = e^{\beta_1 + \beta_2 x_i + \epsilon_i}.$$

Diese Funktion läßt sich linearisieren und damit mit OLS schätzen, indem man beide Seiten der Gleichung logarithmiert:

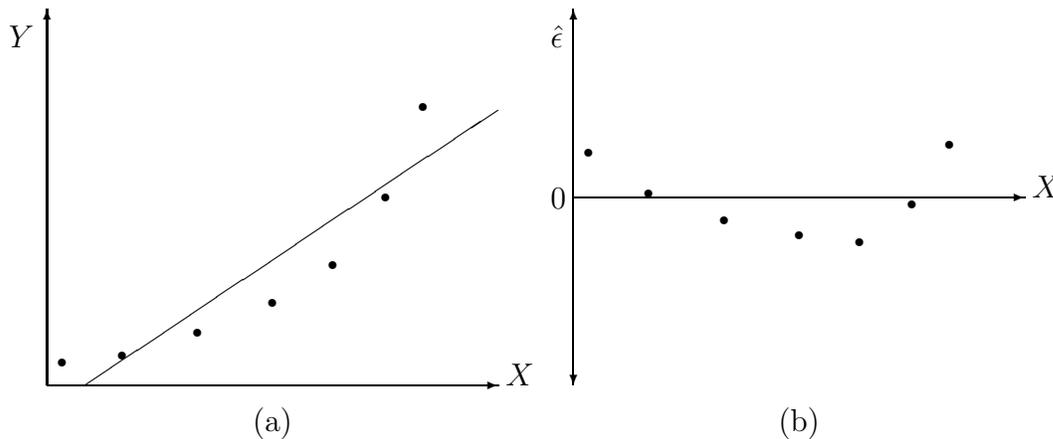
$$\ln(y_i) = \beta_1 + \beta_2 x_i + \epsilon_i.$$

Einkommensfunktionen werden häufig in dieser Form geschätzt. In diesem Modell ist ohne weiteres nur das Vorzeichen von β_2 interpretierbar, welches die Einflußrichtung von X angibt. Den Marginaleffekt erhält man aus

$$\frac{d E(y|x)}{d x} = E(y|x)\beta_2.$$

⁵Bei diskretem X mit nicht zu vielen Ausprägungen bietet es sich an, anstelle der Potenzen eine Dummy für jede Ausprägung (bis auf eine natürlich) in das Modell aufzunehmen. Ein F-Test gibt dann Auskunft, ob die Linearitätsannahme gerechtfertigt ist. Zeigen sich Nicht-Linearitäten, so können die Dummies eventuell eine Vorstellung über die funktionale Form vermitteln.

Abbildung 2: Ein Streudiagramm (a) und der dazugehörige Residuen-Plot (b)



Er hängt offensichtlich von X ab. Meist berichtet man ihn am Mittelwert von X oder an einer sonstigen Stelle, die aus inhaltlichen Gründen von besonderem Interesse ist. Da es sich hier um ein nicht-lineares Modell handelt, ist der Marginaleffekt aber nur eine Näherung des Einheitseffektes. Den exakten Einheitseffekt erhält man aus

$$E(y|x+1) - E(y|x) = E(y|x)(e^{\beta_2} - 1).$$

Ist $\beta_2 < 0,1$, so gilt $e^{\beta_2} - 1 \approx \beta_2$. Mithin sind nur für kleine β_2 Marginal- und Einheitseffekt gleich. Bei größeren Regressionskoeffizienten sollte man deshalb besser den exakten Einheitseffekt berichten. Auch der Einheitseffekt ist jedoch von X abhängig, was die Interpretation erschwert. Eine von X unabhängige Interpretation ist jedoch ebenfalls möglich: Wenn man die obigen Formeln durch $E(y|x)$ dividiert, erkennt man, daß $(e^{\beta_2} - 1) \cdot 100$ (bzw. $\beta_2 \cdot 100$ für kleine Koeffizienten) die prozentuale Veränderung von Y bei Erhöhung von X um eine Einheit angibt. Diese Interpretation der Koeffizienten als „Ertragsrate“ ist sehr anschaulich, und deshalb recht häufig in der Literatur zu finden.

Manchmal findet man *log-lineare Beziehungen* der Form:

$$y_i = B_1 x_i^{\beta_2} e^{\epsilon_i}.$$

Auch dieses Modell läßt sich linearisieren, indem man beide Seiten der Gleichung logarithmiert:

$$\ln(y_i) = \ln(B_1) + \beta_2 \ln(x_i) + \epsilon_i.$$

Man könnte wiederum Marginal- bzw. Einheitseffekte berechnen, aber in diesem Modell kann der Regressionskoeffizient einfacher interpretiert werden: β_2 ist eine Elastizität, denn

$$\frac{d E(\ln(y)|x)}{d \ln(x)} = \beta_2.$$

β_2 gibt an, um wieviel Prozent sich Y verändert, wenn sich X um ein Prozent erhöht.

Die Zahl der möglichen Modelle ist meist sehr groß. Im besten Fall gibt die Theorie oder die vorliegende Forschungsliteratur Aufschluß über das „richtige“ Modell. Oft muß der Forscher jedoch anhand der Daten das am besten geeignete Modell finden. Das einfachste Vorgehen besteht darin, daß man den R^2 -Fit verschiedener funktionaler Beziehungen vergleicht. Eine ausgereifere Methode ist die Konstruktion eines „*Supermodells*“, das verschiedene Transformationen und auch das lineare Modell als Spezialfälle umfaßt. Die ML-Schätzer dieses Supermodells zeigen dann, welche Transformation angebracht ist. Am bekanntesten ist die Box-Cox Transformation der abhängigen Variable (s. Fox, 1991: 68ff; DeVeaux, 1990).

Manche Modelle sind jedoch inhärent *nicht-linear*. Zum Beispiel kann

$$y_i = \alpha + \beta x_i^\gamma + \epsilon_i$$

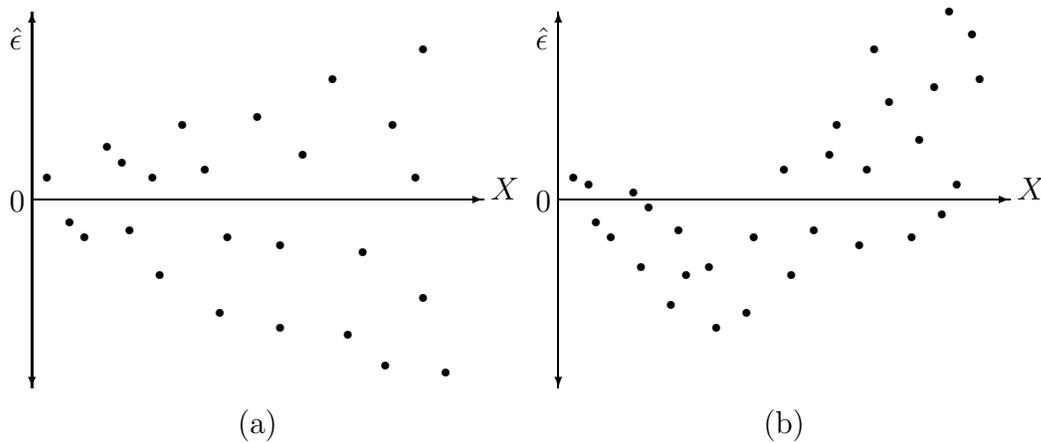
nicht linearisiert werden. Linearität heißt „linear in den Parametern“ und im Gegensatz zu den obigen Modellen können wir dies hier nicht durch eine Transformation erreichen. Die Parameter eines solchen Modells können mittels der nicht-linearen Kleinste-Quadrate-Methode (NLS) geschätzt werden (ausführlich hierzu Seber und Wild, 1989). Das Prinzip ist wie bei OLS: Minimiere die Summe der quadrierten Residuen $\sum_i \hat{\epsilon}_i^2 = \sum_i (y_i - \hat{\alpha} - \hat{\beta} x_i^\gamma)^2$. Die notwendigen Bedingungen für ein Minimum werden allerdings hoch nicht-linear sein, was den Einsatz spezieller iterativer Minimierungs-Algorithmen erforderlich macht. Deshalb ist es meist effizienter, Annahmen über die Verteilung von ϵ_i zu treffen und die ML-Methode zu verwenden. Hat man die Regressionskoeffizienten geschätzt, so kann in unserem Beispiel schließlich ein Test auf $H_0 : \gamma = 1$ Aufschluß bringen, ob tatsächlich ein nicht-lineares Modell erforderlich ist.

2.2.2 Heteroskedastizität

A3 besagt, daß die Variation der abhängigen Variable um die Regressionsebene herum (die Fehlervarianz) an jeder Stelle gleich ist. Ist die Variation nicht konstant, so spricht man von Heteroskedastizität. Die OLS-Schätzer sind zwar bei Verletzung von A3 unverzerrt, aber sie sind nicht mehr effizient (es existieren andere Schätzer mit kleinerer Varianz). Darüberhinaus ist die Formel für die Berechnung der Standardfehler der Koeffizienten nicht mehr korrekt, und damit sind auch die t-Werte für die Signifikanztests falsch.

Das einfachste Prüfverfahren bedient sich der *Residuen-Plots*. „Bildlich“ gesprochen bedeutet Heteroskedastizität, daß die Residuen, trägt man sie gegen X_j oder auch \hat{Y} auf, unterschiedlich weit um die Null-Linie streuen. Abbildung (3)

Abbildung 3: Residuen-Plots mit Heteroskedastizität



veranschaulicht zwei mögliche Fälle. Abbildung (3a) zeigt den Fall, daß die Fehlervarianz mit dem Regressor zunimmt. Dies kann z.B. dann auftreten, wenn Y und X positiv korrelieren, Y aber nach unten beschränkt ist. Dann können die Y -Werte bei kleinem X notwendigerweise nicht so stark variieren (dies wird besonders deutlich, wenn man die Residuen gegen \hat{Y} aufträgt). Entsprechend erhält man bei abhängigen Variablen, die sowohl nach unten wie nach oben beschränkt sind (Prozentwerte, Indexwerte), Residuen-Plots, deren Streuung zuerst zunimmt, dann aber wieder abnimmt. Abbildung (3b) zeigt ebenfalls zunehmende Fehlervarianz, aber nun zusätzlich auch noch Nicht-Linearität. Dies verdeutlicht, daß Residuen-Plots zur gleichzeitigen Überprüfung beider Annahmen eingesetzt werden können.

Leider ist die Situation bei der praktischen Arbeit nur selten graphisch so deutlich zu erkennen, wie in Abbildung (3). Insbesondere bei großen Datensätzen kann man mit dem freien Auge kaum mehr erkennen, ob die Fehler konstant streuen. Deshalb wurden Tests entwickelt, mit deren Hilfe man Heteroskedastizität feststellen kann. Am bekanntesten ist der *Breusch-Pagan Test*. Dieser Test geht davon aus, daß die Fehlervarianz von bestimmten Variablen Z abhängt: $\sigma_i^2 = g(\gamma' z_i)$, wobei die Funktion g beliebig ist. Die Variablen Z können etwa eine Teilmenge der Regressoren sein. Die Nullhypothese konstanter Fehlervarianz ist in diesem Modell äquivalent mit $H_0 : \gamma = 0$. Vereinfachend wird häufig angenommen, daß g linear ist. Dann kann der Breusch-Pagan Test durchgeführt werden, indem man eine OLS-Regression mit den quadrierten und standardisierten Residuen als abhängige Variable berechnet. Mittels eines F-Tests kann dann obige Nullhypothese getestet werden (s. genauer Fox, 1991: 73).

Hat man Heteroskedastizität gefunden, so stellt sich die Frage, was getan werden kann. Als erstes kann man versuchen, die Variablen zu transformieren.

Dies bietet sich insbesondere dann an, wenn man zusätzlich Nicht-Linearitäten oder nicht-normalverteilte Residuen festgestellt hat (einige solche Transformationen sind in den Unterabschnitten 2.2.1 und 2.2.3 beschrieben). In vielen Fällen ist dieses einfache Mittel ausreichend. Darüberhinaus wurden aber eigene Schätzverfahren für Heteroskedastizität entwickelt. Unter Heteroskedastizität kann man die Fehler-Kovarianzmatrix allgemein schreiben als $V(\epsilon) = \sigma^2 \Omega$. Ω ist dabei eine Diagonalmatrix, deren Diagonalelemente die Varianzgewichte sind (unter Homoskedastizität gilt $\Omega = \mathbf{I}$). Ist Ω bekannt, so ist die *generalisierte Methode der kleinsten Quadrate* (GLS) ein effizientes Schätzverfahren, mit $\hat{\beta}_{\text{GLS}} = (\mathbf{X}'\Omega^{-1}\mathbf{X})^{-1}\mathbf{X}'\Omega^{-1}\mathbf{y}$ als Schätzformel. Das Problem ist aber normalerweise, daß Ω unbekannt ist. Dann sind Annahmen zu treffen, damit Ω schätzbar wird. Eine in der Praxis häufig getroffene Annahme ist, daß die Fehlervarianz proportional zum Quadrat eines Regressors j ist: $\sigma_i^2 = \sigma^2 x_{ij}^2$. Dann ist ein effizienter Schätzer leicht zu erhalten, indem man OLS auf die mit $1/x_{ij}$ gewichteten Daten anwendet (sowohl die abhängige Variable, als auch alle Regressoren einschließlich der Konstanten sind durch x_{ij} zu dividieren). Man spricht dann auch von der *gewichteten Methode der kleinsten Quadrate* (WLS). Das Problem von WLS ist offensichtlich: Selten ist die Gültigkeit der Annahme über die Fehlervarianz bekannt. Deshalb hat White (1980) ein Verfahren vorgeschlagen, das einen konsistenten Schätzer der Kovarianzmatrix der OLS-Schätzer liefert, auch wenn Heteroskedastizität unbekannter Form vorliegt. Dieser *White-Schätzer* lautet:

$$\hat{V}_w(\hat{\beta}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1},$$

wobei $\mathbf{D} = \text{diag}(\hat{\epsilon}_1^2, \dots, \hat{\epsilon}_n^2)$. Damit kann auch bei Vorliegen von Heteroskedastizität unbekannter Form OLS verwendet werden (man nimmt dabei allerdings einen Effizienzverlust in Kauf). Der White-Schätzer der Kovarianzmatrix gewährleistet, daß die Signifikanztests dennoch gültig sind. Dies legt es nahe, standardmäßig auch die White-Kovarianzmatrix zu berechnen (was z.B. in LIMDEP leicht möglich ist), um vor Fehlschlüssen aufgrund von Verletzungen der Annahme A3 sicher zu sein.

2.2.3 Nicht-normalverteilte Fehler

Das Gauss-Markov Theorem sagt uns, daß auch bei nicht-normalverteilten Fehlern die OLS-Schätzer unverzerrt sind. Auch die Tests sind asymptotisch gültig, d.h. bei großen Stichproben können wir die Signifikanztests auch bei nicht-normalverteilten Fehlern weiter einsetzen. Verletzungen der Normalverteilungsannahme sind folglich im Forschungsalltag kein gravierendes Problem. Eher ist dies der Fall, wenn man mit seinen Schätzungen Prognosen tätigen will, denn bei Verletzung der Normalverteilungsannahme sind die OLS-Schätzer keine ML-Schätzer mehr und nicht mehr effizient. Andere Schätzer sind in diesem Fall genauer.

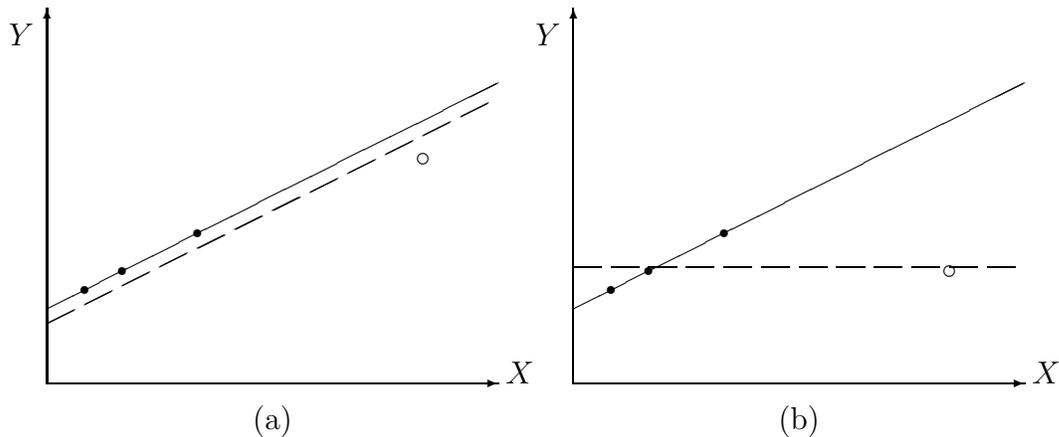
Zur Überprüfung der Verteilungsannahme seien zwei graphische Verfahren vorgestellt. Das naheliegendste Verfahren besteht darin, ein *Histogramm der standardisierten Residuen* zu erstellen. Viele Programme zeichnen in solch ein Histogramm zusätzlich die exakte Normalverteilungskurve ein, so daß eventuelle Abweichungen leicht zu erkennen sind. Zeigen die Residuen etwa eine starke Schiefe, so empfiehlt es sich, die Daten zu transformieren (s.u.). Eine Residuen-Verteilung mit mehreren Modalwerten deutet darauf hin, daß wichtige qualitative Regressoren übersehen wurden. Besonders bedeutsam ist es Abweichungen in den Verteilungsrändern zu erkennen, denn dort verbergen sich „Ausreißer“ (s. Abschnitt 2.2.4). Um solche Abweichungen zu erkennen, ist das *Normal-Probability Plot* besonders geeignet. Hier wird im Prinzip die kumulierte Verteilung der standardisierten Residuen gegen die kumulierte Standardnormalverteilung aufgetragen. Folgen die Residuen einer Normalverteilung, so sollte sich eine Gerade ergeben. Überbesetzungen in den Rändern drückt sich darin aus, daß das Normal-Probability Plot bei betragsmäßig großen Residuen von der Geraden abweicht.

Stellt man fest, daß die Residuen von der Normalverteilung abweichen, so besteht die theoretisch beste Vorgehensweise darin, den entsprechenden ML-Schätzer zu verwenden. Allerdings setzt dies voraus, daß man die Verteilung der Residuen beschreiben kann. Dies wird so gut wie nie der Fall sein. Die graphischen Checks werden nur Aufschluß über ins „Auge springende“ Charakteristika der Verteilung liefern. Dies reicht jedoch meist aus, um mittels Transformation die gravierendsten Abweichungen zu beseitigen. Rechtsschiefe Residuen etwa kann man oft dadurch beseitigen, daß man die Y-Werte logarithmiert. Analog empfiehlt es sich bei linksschiefen Residuen die Y-Werte zu potenzieren. Handelt es sich bei Y um Prozentwerte, die ja nach oben und unten beschränkt sind, so hilft meist eine Logit-Transformation $y^* = \ln[y/(1-y)]$. Bei „gestutzten“ Verteilungen allerdings, wo viele Y-Werte (und damit auch Residuen) auf einen Wert fallen, helfen Transformationen nicht mehr weiter. In diesem Fall muß man auf die speziellen Verfahren für begrenzte Daten zurückgreifen (s. Abschnitt 2.4).

2.2.4 Einflußreiche Datenpunkte

Ein weiteres Problem sind *einflußreiche Datenpunkte*. Ein Datenpunkt ist einflußreich, wenn er die Regressionsschätzer (Koeffizienten und/oder Standardfehler) stärker beeinflusst als es die anderen Beobachtungen tun. Dies stellt natürlich keine Verletzung irgendwelcher Annahmen des Regressionsmodells dar, aber die meisten Forscher betrachten einflußreiche Datenpunkte als problematisch, weil einige wenige, eventuell meßfehlerbehaftete Beobachtungen das Ergebnis über Gebühr beeinflussen. So können etwa einige „extreme“ Beobachtungen einen Zusammenhang zwischen Y und X erzeugen, während die „Masse“ der anderen Beobachtungen keinen Zusammenhang anzeigt. Das ist insbesondere bei kleinen

Abbildung 4: Ausreißer und Einfluß im einfachen Regressionsmodell



Stichproben ein Problem, aber selbst sehr große Stichproben sind nicht davor gefeit. So berichten Kahn und Udry (1986) von einer Studie über die Koitushäufigkeit (pro Monat) in 2063 Ehen. Ein erstaunliches Ergebnis dieser Studie war, daß die Koitushäufigkeit mit dem Alter der Frau ansteigt. Wie sie in einer Replikation zeigen, ist dieses Ergebnis darauf zurückzuführen, daß vier Ehen mit eigentlich fehlenden Werten auf der abhängigen Variable (88!) irrtümlich in die Analyse aufgenommen wurden. Dieses Beispiel zeigt, daß es sich auch bei großen Datensätzen lohnt, die folgenden Diagnoseverfahren einzusetzen.

Einflußreiche Datenpunkte sind immer *Ausreißer*.⁶ Ein Ausreißer ist definiert als eine Beobachtung, die bezüglich Y und/oder X einen „ungewöhnlichen“, d.h. weit vom Mittelwert entfernten Wert aufweist. Die Streudiagramme in Abbildung (4) zeigen zwei mögliche Konstellationen (für eine einfache Regression).⁷ Die Ausreißer sind als Kreis gekennzeichnet, die „normalen“ Beobachtungen als Punkte. Die ohne Ausreißer geschätzten Regressionsgeraden sind durchgezogen, die mit Ausreißer geschätzten sind gestrichelt.

Abbildung (4a) zeigt den Fall, in dem der Ausreißer einen ungewöhnlichen X -Wert hat, der Y -Wert aber dennoch nahe bei der ohne Ausreißer geschätzten Regressionsgerade liegt. $\hat{\beta}$ wird von dem Ausreißer kaum beeinflusst, aber der Standardfehler wird kleiner sein. Die Präzision des Koeffizientenschätzers ist höher, weil die Varianz von X größer ist. Problematischer ist der Fall, der in Abbildung (4b)

⁶Umgekehrt muß dies nicht der Fall sein. Ein „ Y -Ausreißer“ etwa, der nahe bei \bar{x} liegt, beeinflusst die Regressionschätzer kaum.

⁷Im multiplen Fall verwendet man sogenannte „Partielle-Regression Streudiagramme“ (s.u.). Man beachte, daß zur Aufdeckung von Nicht-Linearität und Heteroskedastizität Residuen-Plots besser geeignet sind, Ausreißer aber besser in Streudiagrammen erkannt werden können.

dargestellt ist: Hier beeinflusst der Ausreißer $\hat{\beta}$ dramatisch, er „zerstört“ den Zusammenhang zwischen Y und X (der umgekehrte Fall kann auch auftreten, wie man sich leicht klarmachen kann).

Wegen solcher Effekte ist es sinnvoll, sich einflußreiche Datenpunkte näher anzusehen. Als erstes sollte überprüft werden, ob keine Fehler bei der Messung oder Verkodung gemacht wurden. Es wäre fatal, wenn durch solche Fehler Ausreißer erzeugt werden, die das Ergebnis stark verändern. Stellt sich allerdings heraus, daß die Ausreißer korrekt sind, so liegt das Problem bei der Modellspezifikation. Weglassen der Ausreißer ist sicherlich keine Lösung. Der Forscher muß sich vielmehr überlegen, ob nicht wichtige Einflußgrößen übersehen wurden. Hierfür ist es hilfreich, sich alle Informationen über die Ausreißer anzusehen. Hat man mehrere Ausreißer identifiziert, so kann man statistische Kennzahlen (Mittelwerte etwa) berechnen und mit den „unauffälligen“ Beobachtungen vergleichen. Dabei kann man oft Merkmale identifizieren, die die Ausreißer von den anderen Beobachtungen unterscheiden. Diese Merkmale sind dann als Regressoren in das Modell mit aufzunehmen (ausführlich demonstrieren dieses Vorgehen Bollen und Jackman, 1990).⁸ Auf jeden Fall hüte man sich davor, die von den im folgenden vorzustellenden Einfluß-Maßen identifizierten Beobachtungen routinemäßig aus der Analyse auszuschließen (wie es inzwischen in manchen Arbeiten leider gemacht wird). Dies ist nichts anderes als die Anpassung der (Daten-)Realität an das (statistische) Modell.

Im einfachsten Fall kann man Ausreißer über die Inspektion der Streudiagramme ausfindig machen. Aber auch numerische Kennziffern wurden vorgeschlagen. Beobachtungen mit ungewöhnlichen X -Werten („high leverage points“) können über die sogenannten *Hat-Werte* (h_i) gefunden werden. h_i ist ein Diagonalelement der Hat-Matrix

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'.$$

Der Name kommt davon, daß $\hat{\mathbf{y}}$ („y-hat“) sich ergibt aus:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{H}\mathbf{y}.$$

Je größer h_i ist, desto stärker ist der Einfluß der Beobachtung i auf die Prognose von Y . Es gilt $1/n \leq h_i \leq 1$, wobei h_i im Mittel gleich k/n ist. Die in der Literatur meist geäußerte Faustregel lautet, alle Beobachtungen, für die $h_i > 2k/n$ gilt, als Ausreißer zu betrachten. Alternativ wird häufig auch vorgeschlagen,

⁸Eine weitere Möglichkeit stellt die Verwendung von robusten Regressionstechniken dar (s. z.B. Rousseeuw und Leroy, 1987; Berk, 1990; Härdle, 1990). Bei OLS haben abweichende Beobachtungen starkes Gewicht, weil ihre Residuen quadriert in die Schätzgleichungen eingehen. Bei robusten Verfahren wird deshalb großen Residuen weniger Gewicht gegeben. Die bekannteste robuste Alternative zu OLS ist die Methode der kleinsten absoluten Abweichungen, bei der die absoluten Residuen in die Berechnungen eingehen.

studentisierte Residuen

$$\epsilon_i^* = \frac{\hat{\epsilon}_i}{\sqrt{\hat{\sigma}^2(1 - h_i)}}$$

zu betrachten. Beobachtungen, deren ϵ_i^* eine festzulegende Schwelle (meist $|\epsilon_i^*| > 2$) überschreiten, sollten näher betrachtet werden. Allerdings ist es nicht unproblematisch, die Ausreißerdiagnose nur auf die Inspektion der Residuen zu stützen. Wenn nämlich Ausreißer die Regressionsebene stark an sich „heranziehen“ können (wie in Abbildung (4b)), so werden deren Residuen klein sein. Dann wird nur der Hat-Wert die Ausreißer finden können. Aus diesem Grund liegt es nahe, Maße zu konstruieren, die die Hat-Werte und die studentisierten Residuen kombinieren. Ein solches Einfluß-Maß ist *Cook's D*:

$$D_i = \frac{\epsilon_i^{*2}}{k} \left(\frac{h_i}{1 - h_i} \right).$$

Beobachtungen mit relativ großem D_i sind einflußreiche Datenpunkte. Da in großen Stichproben der absolute Einfluß einer Beobachtung normalerweise gering ist, wird die fallzahlabhängige Grenze $4/(n - k)$ empfohlen.

Ein anderer Ansatz versucht, den Einfluß einer Beobachtung direkt zu messen. Dabei wird untersucht, wie sich $\hat{\beta}_j$ verändert, wenn Beobachtung i weggelassen wird ($\hat{\beta}_{j(-i)}$).⁹ Das Maß

$$\text{DFBETAS}_{ij} = \frac{\hat{\beta}_j - \hat{\beta}_{j(-i)}}{\hat{\sigma}_{\hat{\beta}_{j(-i)}}}$$

zeigt an, wie groß der (standardisierte!) Einfluß der Beobachtung i auf einen Koeffizienten j ist.¹⁰ Da in großen Stichproben eine einzelne Beobachtung kaum große Veränderungen der Schätzer bewirken kann, wird in der Literatur eine fallzahlabhängige Schwelle vorgeschlagen: Ist der Betrag von DFBETAS_{ij} größer $2/\sqrt{n}$, so ist Vorsicht geboten. Weil für jeden Koeffizienten (einschließlich der Konstanten) und für jede Beobachtung ein DFBETAS_{ij} berechnet werden kann, empfiehlt sich ein graphisches Vorgehen. Für jede Variable wird ein Plot erstellt, in dem DFBETAS gegen die Fallnummer aufgetragen wird (ein sogenannter *Indexplot*). Zeichnet man dann die Schwelle als Linie in dieses Diagramm, so kann man schnell prekäre Fälle ausfindig machen.

Keines dieser Verfahren bietet die absolute Gewähr dafür, daß alle einflußreichen Datenpunkte gefunden werden. In der Praxis sollte man deshalb mehrere der

⁹Dieses Maß betrachtet nur den Einfluß einer Beobachtung auf die Koeffizientenschätzer. Es gibt weitere Maße für den Einfluß auf den Standardfehler (s. Fox, 1991: 31f). Ein solches Maß würde auch den Ausreißer in Abbildung (4a) entdecken.

¹⁰Auch Cook's D kann in dieser Form geschrieben werden, mißt aber im Unterschied zu DFBETAS nicht den Einfluß auf einen einzelnen Koeffizienten, sondern auf den gesamten Koeffizienten-Vektor.

obigen Verfahren einsetzen. Hat man die Ausreißer erkannt und entweder die Daten korrigiert oder das Modell verändert, sollten diese Diagnoseverfahren auch auf das neue Modell angewandt werden, denn es ist leicht möglich, daß durch die Maßnahmen gegen die zuerst entdeckten Ausreißer andere Beobachtungen zu Ausreißern avancieren. Insofern sollte Ausreißerdiagnostik ein iterativer Prozeß sein.

2.2.5 Anwendungsbeispiel: Scheidungsraten in 50 Ländern

Um das multiple Regressionsmodell und das Vorgehen bei der Regressionsdiagnostik zu illustrieren, wollen wir im folgenden Scheidungsraten untersuchen. Dazu verwenden wir einen Datensatz von Diekmann (1992), der für 146 Länder versucht hat, Angaben über Scheidungen zu sammeln. Bei 69 Ländern war es ihm möglich, eine Scheidungsrate zu berechnen (die Angaben beziehen sich auf jeweils ein Jahr aus dem Zeitraum 1980-88). Man beachte, daß es sich hierbei nicht um Individualdaten handelt, sondern um Aggregatdaten. Dies hat verschiedene Konsequenzen: (1) Der Fit von Regressionsmodellen für Aggregatdaten ist meist erheblich besser. Dies liegt daran, daß durch die Aggregation viel individuelles „Rauschen“, das den Fit auf der Individualebene schlechter macht, „herausgefiltert“ wurde. (2) Den gefundenen Beziehungen auf der Aggregatebene stehen nicht unbedingt auch gleichlaufende Beziehungen auf der Individualebene gegenüber. Das Problem des „ökologischen Fehlschlusses“ mahnt zur Vorsicht bei der Interpretation der Ergebnisse.¹¹ (3) Die Fallzahlen sind recht niedrig, weshalb einzelne Ausreißer einen großen Einfluß auf die Ergebnisse haben können. Deshalb ist es unerlässlich, Verfahren zur Regressionsdiagnostik einzusetzen.

Inhaltlich beschränken sich die folgenden Analysen auf nur drei Variablen (Diekmann (1992) verwendet noch einige weitere Variablen): die Sex-Ratio, die Erwerbsquote der Frauen und das Modernisierungsniveau. Frauenerwerbstätigkeit sollte die Scheidungsrate erhöhen, weil für unzufriedene Ehefrauen Erwerbsarbeit eine Alternative zur Ehe ist. Moderne Gesellschaften sollten allein schon wegen der höheren Lebenserwartung höhere Scheidungsraten aufweisen. Für die Sex-Ratio (definiert als Verhältnis der Zahl der Männer zur Zahl der Frauen) formulierten Guttentag und Secord (1983) folgende Hypothese: Hohe Sex-Ratios (also Knappheit von Frauen) gehen einher mit eher patriarchalischen Lebensformen.

¹¹Mit der Analyse von Daten mehrerer Ebenen befaßt sich die *Mehrebenenanalyse*, wobei insbesondere zwei Fragestellungen verfolgt werden: (1) Verfahren der ökologischen Analyse beschäftigen sich damit, wie man aus Analysen auf der Aggregatebene Schlüsse auf die Individualebene ziehen kann. Das Problem des ökologischen Fehlschlusses ist hierbei die Hauptschwierigkeit (Achen und Shively, 1995). (2) Die Kontextanalyse dagegen beschäftigt sich mit Modellen, die es erlauben, den Einfluß von Aggregatmerkmalen (Kontexten) auf Prozesse auf der Individualebene zu untersuchen. Bei diesen Verfahren muß man über Daten auf beiden Ebenen verfügen (DiPrete und Forristal, 1994; Hox und Kreft, 1994).

Tabelle 1: Regressionsmodelle auf die Scheidungsrate

Variable	(1)	(2)	(3)	(4)	(5)
Konstante	-504,0* (2,34)	-351,7 (0,55)	-351,7 (0,54)	155,1 (0,20)	931,3 (1,06)
Sex-Ratio	4,1* (2,94) [1,21]	1,4 (0,23)	1,4 (0,22)	-3,0 (0,49) [1,12]	-6,9 (1,05)
Erwerbsquote	7,9* (2,94) [1,32]	10,7* (3,31)	10,7* (3,31)	9,0 (1,13) [2,27]	3,2 (0,35)
Modernisierungsindex	52,2* (2,75) [1,19]	75,2* (2,86)	75,2* (3,05)	82,6 (1,87) [2,36]	20,4 (0,44)
λ					-167,2* (2,36)
N	50	44	44	43	43
R^2	0,46	0,55	0,55	0,57	0,58
F	13,2	16,6	16,6	17,5	13,0

* signifikant auf dem 5%-Niveau. Absolute t-Werte in runden Klammern. $\sqrt{\text{VIF}}$ in eckigen Klammern.

(1) OLS mit allen Daten.

(2) OLS ohne Ausreißer.

(3) OLS mit White-Schätzer der Kovarianzmatrix.

(4) ILS mit Sex-Ratio, Heiratsalter und Modernisierungsindex als Instrumente.

(5) 2SLS mit Heckit-Korrektur für den Stichprobenauswahl-Fehler. Auswahlgleichung mit der Variable Bruttonationaleinkommen pro Kopf.

Quelle: Eigene Berechnungen mit Daten aus Diekmann (1992).

In solchen Gesellschaften sollte, so die Guttentag-Secord Hypothese, die sexuelle Freiheit von Frauen stärker eingeschränkt und die Scheidungsrate niedriger sein.

Die Scheidungsrate ist definiert als die Zahl der Scheidungen pro 1000 Eheschließungen in einem Jahr. Sie hat in unseren Daten einen Wertebereich von 25,6 bis 494,9 (Mittel 226,3). Die Sex-Ratio ist definiert als das Verhältnis von Zahl der Männer der Altersgruppe 15 bis 49 zu Zahl der Frauen dieser Altersgruppe (multipliziert mit 100). Ihr Wertebereich ist 92,9 bis 182,8 (Mittel 104,1). Die Erwerbsquote der Frauen ist in Prozent gemessen und reicht von 19,3 bis 48,0 (Mittel 37,3). Das Modernisierungsniveau wird durch einen additiven, mit Faktorladungen gewichteten Index repräsentiert, in den das Volkseinkommen pro

Kopf, die Kindersterblichkeit, die Lebenserwartung und der Verstärkerungsgrad eingehen. Der Modernisierungsindex reicht von -2,8 bis 1,3 (Mittel 0,2). Wegen fehlender Werte basieren die folgenden Auswertungen nur auf 50 Fällen.

In Tabelle (1) sind die Ergebnisse der OLS-Regression festgehalten (Spalte 1).¹² Der Fit des Modells ist mit $R^2=0,46$ als gut zu bezeichnen. Wie der F-Wert von 13,2 zeigt, ist das Modell „signifikant“ (Der kritische Wert beträgt bei 3 und 46 Freiheitsgraden auf dem 1%-Niveau 4,2). Die negative Konstante ist der Wert der Scheidungsraten, den das Modell prognostiziert, wenn alle Kovariate gleich null wären. Da die beobachteten Sex-Ratios und die Erwerbsquote weit jenseits von null liegen, ist in unserem Fall die Konstante nicht sinnvoll zu interpretieren. Die negative Konstante deutet aber ein Problem an: Unser Modell ist so formuliert, daß für extreme Datenkonstellationen der Vorhersagewert der Scheidungsrate negativ werden kann, was ja eigentlich durch die Definition dieser Größe ausgeschlossen ist. Wollte man dieses Modell für Prognosezwecke einsetzen, so müßte die Scheidungsrate einer geeigneten Transformation (z.B. einer logarithmischen) unterzogen werden, um dies zu verhindern. Der Koeffizient für die Sex-Ratio besagt, daß eine Erhöhung des Geschlechterverhältnisses um einen Prozentpunkt, 4,1 zusätzliche Scheidungen erzeugt. Dies gilt „ceteris paribus“, d.h. für Länder, die dieselbe Erwerbsquote und denselben Modernisierungsgrad haben. Der Koeffizient ist auch signifikant von null verschieden, weshalb die Guttentag-Secord Hypothese durch unsere Daten widerlegt ist. Signifikant und positiv sind auch die Koeffizienten der beiden anderen Variablen, was den oben geäußerten Hypothesen entspricht. Besonders anschaulich ist der Effekt der Frauenerwerbsquote interpretierbar: Ein Prozentpunkt mehr Erwerbstätigkeit der Frauen zieht 7,9 zusätzliche Scheidungen nach sich.

Diese Ergebnisse sind allerdings noch als vorläufig zu betrachten, solange wir das Modell nicht den wichtigsten diagnostischen Checks unterzogen haben. Zur Multikollinearitäts-Diagnostik sind in Spalte (1) unter den Koeffizienten und den t-Werten auch die \sqrt{VIF} -Werte angeführt. Wie man sieht, werden die Standardfehler durch die Korrelation der unabhängigen Variablen untereinander kaum erhöht. Multikollinearität ist bei diesem Modell mithin kein Problem.

Bei der Vielzahl der oben vorgestellten Diagnoseverfahren stellt sich sofort die Frage, wie man nun weiter vorgeht. Eine Möglichkeit sieht so aus: (1) Zuerst spürt man einflußreichen Datenpunkten nach. Dazu verwendet man Cook's D und DFBETAS. Zur Absicherung betrachtet man aber auch die Streudiagramme. Entdeckt man einflußreiche Datenpunkte und entschließt sich zur Modifikation des Ausgangsmodells, so muß das Modell neu geschätzt werden. Dann kann eine weitere Ausreißerdiagnose erfolgen. Es ist jedoch davor zu warnen, diesen itera-

¹²Die Analysen dieses Abschnitts wurden mit SPSS für Windows durchgeführt. Dieses Programm verfügt (ebenso wie SYSTAT und SAS) über ausgezeichnete Diagnosemöglichkeiten und bietet auch viele graphische Checks an (s. Norusis, 1993: Kap. 18).

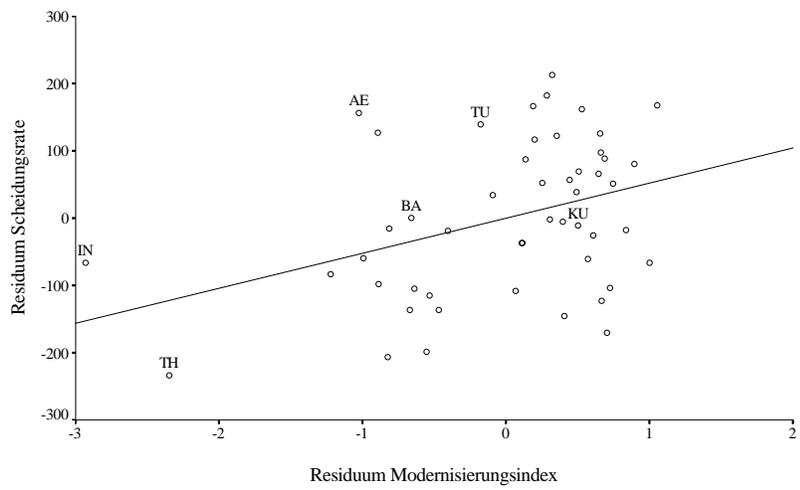
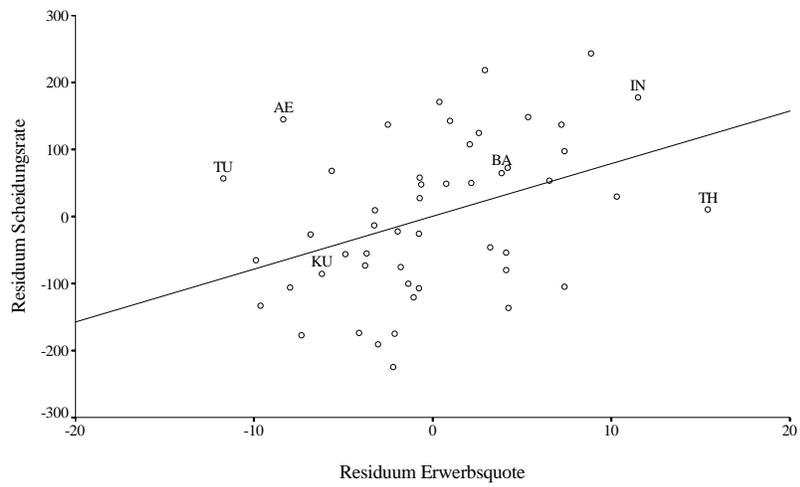
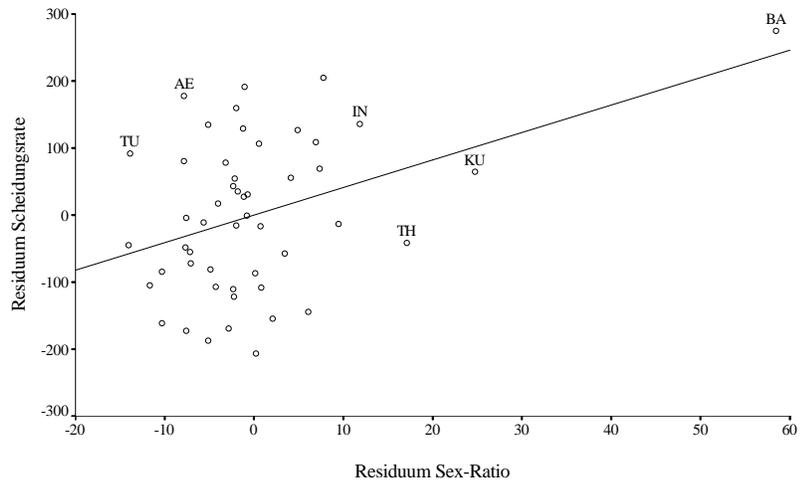


Abbildung 5: Partielle-Regressions Streudiagramme (N=50)

tiven Prozeß zu weit zu führen, denn jeder weitere Schritt geht zu Lasten der Sparsamkeit des Modells. (2) Mit den Residuen des modifizierten Modells erstellt man Residuen-Plots, und versucht Nicht-Linearitäten und Heteroskedastizität zu entdecken. Dies führt eventuell zu weiteren Modifikationen, worauf man wieder ein Residuen-Plot erstellen sollte, denn manche „Heilungsversuche“ werden nicht das gewünschte Ergebnis hervorbringen oder führen zur Verletzung anderer Annahmen. (3) Schließlich wird man die Normalverteilungsannahme überprüfen (mittels Normal-Probability-Plot etwa). Ergibt sich hier die Notwendigkeit zu Transformationen, so muß der zweite Schritt wiederholt werden.

(1) *Ausreißerdiagnose*: Als erstes betrachten wir Cook's D. Über der Grenze von 0,087 (4/46) liegen folgende fünf Länder: Bahrain, Ägypten, Tunesien, Thailand und Indonesien. Diese fünf Länder werden auch durch die standardisierten DFBETAS als auffällig identifiziert (die Grenze ist hier 0,28). Bezüglich des Sex-Ratio Koeffizienten zeigt nur Bahrain einen auffälligen Einfluß. Bei der Erwerbsquote fallen Ägypten, Tunesien und Thailand auf, beim Modernisierungsindex Ägypten, Thailand und Indonesien. Beruhigenderweise kommen also beide Maße zu den selben Schlußfolgerungen.

Um dieses Ergebnis graphisch abzusichern, betrachten wir nun die Streudiagramme. Im multiplen Fall wird man nicht die bivariaten Streudiagramme betrachten, sondern die sogenannten Partiellen-Regressions Streudiagramme. Man trägt nicht Y gegen X_j , sondern das Residuum aus der Regression von Y auf alle anderen X gegen das Residuum aus der Regression von X_j auf alle anderen X auf. Bildlich gesprochen wird dadurch aus Y und X_j der Effekt der anderen Variablen „herausgerechnet“. ¹³ Ausreißer im Partiellen-Regressions Streudiagramm sind somit ungewöhnliche Datenpunkte, selbst wenn man für die anderen Variablen kontrolliert. Im bivariaten Streudiagramm könnten manche Beobachtungen als Ausreißer erscheinen, nur weil man hier noch nicht für die anderen Variablen kontrolliert hat. Die Steigung der Regressionsgerade dieser beiden Residuen ist im übrigen identisch mit dem multiplen Regressionskoeffizienten der jeweiligen Variable. Zeichnet man diese Gerade in das Streudiagramm ein, so kann man leicht erkennen, welche Beobachtungen einen großen Einfluß auf die Steigung der Regressionsgeraden ausüben.

Abbildung (5) enthält die drei Streudiagramme für unser Modell. Die von den Maßzahlen identifizierten Länder sind mit einem Kürzel versehen (ebenso Kuwait). Bezüglich der Erwerbsquote erkennt man sofort die von DFBETAS identifizierten einflußreichen Länder: Ägypten, Tunesien und Thailand. Jedes dieser Länder flacht die Regressionsgerade ab. Noch deutlicher ist beim Modernisierungsindex der starke Einfluß von Thailand und Indonesien zu erkennen. Aber auch der Einfluß Ägyptens ist offensichtlich. Jedes dieser drei Länder „zieht den

¹³In SPSS heißen sie „partial plots“ und werden automatisch erstellt. In der Literatur findet sich häufig auch die Bezeichnung „added-variable plot“.

Anfang der Regressionsgerade hoch“, weshalb der Schätzer für den Modernisierungsgrad niedriger ausfällt. Schließlich wird bei der Sex-Ratio der extreme Einfluß von Bahrain deutlich: ohne diese Beobachtung wäre die Regressionsgerade annähernd waagrecht. Die Streudiagramme bestätigen also die Ergebnisse, die wir mit den Einfluß-Maßen erhalten haben. Zusätzlich erkennen wir aber einen weiteren einflußreichen Datenpunkt: Kuwait hätte, wenn Bahrain nicht enthalten wäre, einen deutlichen Einfluß auf den Koeffizienten der Sex-Ratio. Man kann es auch so ausdrücken: Bahrain und Kuwait üben *gemeinsam* einen starken Einfluß aus. Deshalb wurde Kuwait auch nicht von den Einfluß-Maßen entdeckt, weil diese nur *einzelne* einflußreiche Datenpunkte entdecken. Läßt man die im ersten Durchgang gefundenen einflußreichen Datenpunkte weg (also auch Bahrain) und berechnet für das modifizierte Modell erneut Cook’s D, so zeigt Kuwait den mit Abstand größten D-Wert. Das Problem mit Kuwait erkennt man allerdings viel einfacher in einem Partiellen-Regressions Streudiagramm.

Wir haben somit sechs einflußreiche Datenpunkte ausgemacht. Nun stellt sich die Frage, was zu tun ist mit diesen sechs Ländern. Bahrain und Kuwait fallen auf durch ihre extremen Sex-Ratios (Kuwait 150, Bahrain 183, das Maximum der anderen Länder ist 109). Diese sind nicht „natürlich“, sind vielmehr durch die vielen Gastarbeiter, die in diesen Ländern arbeiten, verursacht. Insofern spiegeln die Sex-Ratios dieser beiden Länder auch keine strukturellen Heiratsmarktbarrieren wieder. Da die „natürlichen“ Sex-Ratios nicht verfügbar sind, müssen diese beiden Länder aus der Analyse ausgeschlossen werden. Bei den anderen vier Ländern fällt auf, daß es sich überwiegend um islamische Staaten handelt. Dies legt es nahe, dem Problem mit einer zusätzlichen Dummy für islamische Länder zu begegnen. Es zeigt sich auch, daß der Einfluß der vier Länder in solch einem Modell tatsächlich verschwindet. Allerdings übt dann die Türkei einen starken Einfluß aus. Da es sich hier um ein Demonstrationsbeispiel handelt, soll dieser langwierige Weg zu einem besser spezifizierten Modell nicht beschritten werden. Wir schließen für die folgenden Analysen die sechs einflußreichen Länder einfach aus.

Die Ergebnisse nach Ausschluß der sechs Länder sind in Tabelle (1) festgehalten (Spalte 2). Wie aufgrund der Partiellen-Regressions Streudiagramme zu erwarten war, zeigt sich nun, daß die Sex-Ratio keinen signifikanten Effekt hat. Dies ist auf den Ausschluß von Bahrain und Kuwait zurückzuführen. Die Effekte für die Erwerbsquote und das Modernisierungsniveau fallen dagegen deutlich stärker aus. Man beachte auch, daß R^2 deutlich ansteigt, was natürlich darauf zurückzuführen ist, daß die Ausreißer ausgeschlossen wurden.

(2) *Residuen-Plots*: Nun kommen wir zum zweiten Schritt: Mittels der Residuen-Plots soll die Linearität der Beziehungen im modifizierten Modell überprüft werden. Wie beim ersten Schritt empfiehlt es sich auch hier, nicht einfach die Residuen gegen X_j aufzutragen. Man erstellt besser sogenannte Partielle-Residuen-

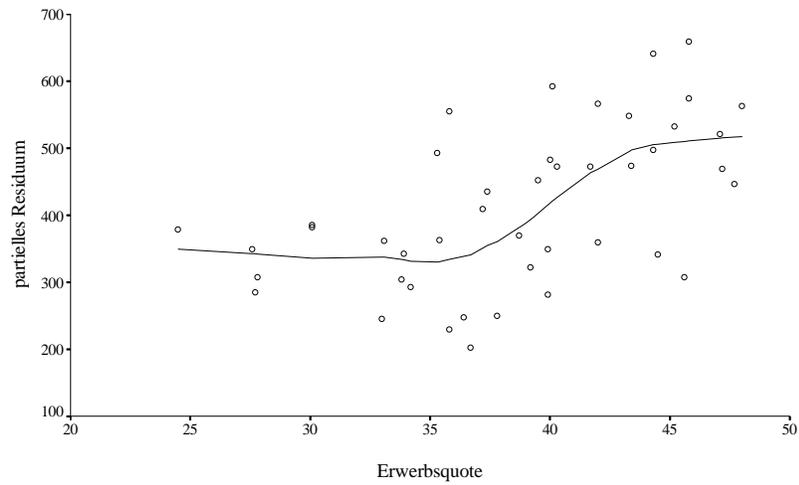
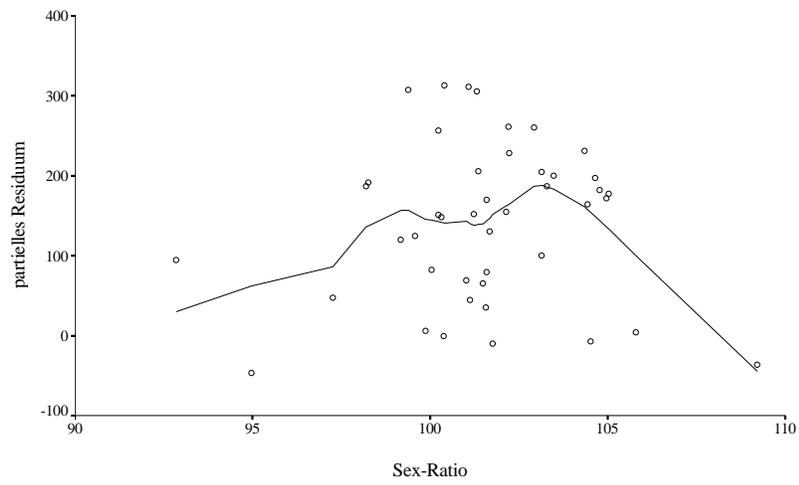
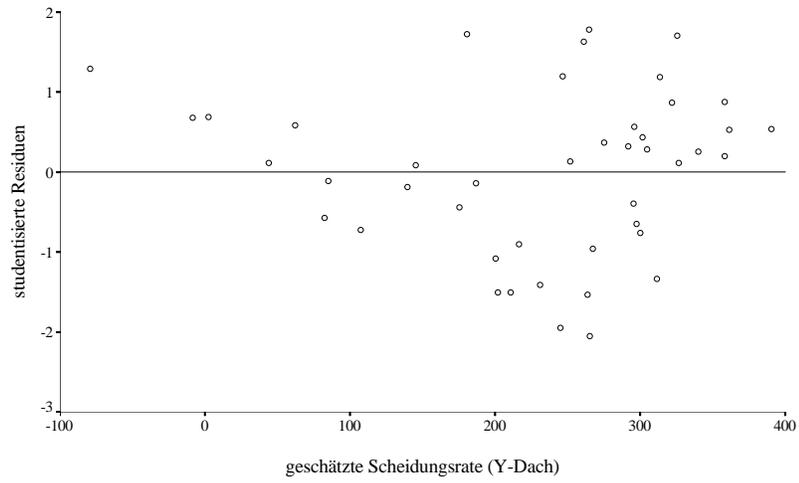


Abbildung 6: Partielle-Residuen-Plots (N=44)

Plots. Bei diesen Plots wird zu den (unstandardisierten) Residuen jeweils $\hat{\beta}_j x_{ij}$ hinzuaddiert. Diese partiellen Residuen trägt man dann gegen X_j auf. Man „addiert zu den Residuen die Regressionsgerade hinzu“, um die Form einer eventuellen Nicht-Linearität besser erkennen zu können.¹⁴ Abbildung (6) enthält die drei Partielle-Residuen-Plots. In jede dieser Graphiken ist zusätzlich eine sogenannte Lowess-Kurve eingezeichnet. Diese Kurve wird mittels eines Glättungsalgorithmus errechnet, bei dem an jeder Stelle aus einer lokalen, gewichteten Regression ein Vorhersagewert errechnet wird. Die Verbindung dieser Werte ergibt die Lowess-Kurve (s. Goodall, 1990; Härdle, 1990). Sie erleichtert die visuelle Erkennung des Musters einer eventuellen Nicht-Linearität.

Im ersten Plot erkennen wir, daß der Zusammenhang zwischen Sex-Ratio und Scheidungsrate weitgehend linear ist. Drei Länder mit hoher Sex-Ratio und ungewöhnlich niedriger Scheidungsrate ziehen allerdings die Lowess-Kurve nach unten. Dies könnte mit einem zusätzlichen quadratischen Sex-Ratio-Term modelliert werden, man muß sich jedoch bewußt sein, daß ein eventuelles negatives Vorzeichen dieses Terms auf nur drei Beobachtungen zurückzuführen ist (ein positiver linearer und ein negativer quadratischer Koeffizient beschreiben eine umgekehrt U-förmige Kurve). Besser abgesichert erscheint die Einführung eines quadratischen Terms bei der Erwerbsquote. Hier spricht doch der Großteil der Daten für eine konvexe Beziehung. Ähnlich verhält es sich beim Modernisierungsindex. Allerdings ist hier anstatt einer stetigen Modellierung der Nicht-Linearität (mittels quadratischem Term) eine diskrete Modellierung angebracht, da der Residuen-Plot eine sprunghafte Veränderung des Zusammenhangs ab einem Modernisierungsgrad von etwa 0,5 andeutet. An dieser Stelle ist auch eine „Lücke“ in den Daten zu erkennen, weshalb es sich anbietet, die Länder in zwei Gruppen zu teilen und für jede Gruppe den Modernisierungsindex in das Modell aufzunehmen (die Länder der anderen Gruppe werden jeweils auf null gesetzt).

Die Residuen-Plots können auch zur Erkennung von Heteroskedastizität eingesetzt werden. Bei der Erwerbsquote und dem Modernisierungsindex finden wir Anzeichen für ein Zunehmen der Streuung der Residuen. Besser kann Heteroskedastizität allerdings diagnostiziert werden, wenn wir die studentisierten Residuen gegen \hat{Y} auftragen (in SPSS die Residuen und Vorhersagewerte abspeichern und ein Streudiagramm erstellen). In Abbildung (7) erkennt man deutlich, daß mit zunehmender Scheidungsrate die Residuen stärker streuen. Dies ist das typische, trichterförmige Muster, das immer dann zu erwarten ist, wenn die abhängige Variable nach unten beschränkt ist. In solchen Fällen bietet sich eine logarithmische Transformation an, die außerdem das Problem mit den negativen vorhergesagten Scheidungsraten behebt (man erkennt 2 Länder mit negativen Vorhersagewerten).

¹⁴Diese Plots sind in SPSS nicht automatisch erhältlich. Man kann aber die multiple Regression berechnen, die unstandardisierten Residuen abspeichern und jeweils $\hat{\beta}_j x_{ij}$ addieren ($\hat{\beta}_j$ entnimmt man dem Regressionsoutput). Für jede unabhängige Variable läßt man sich dann ein Streudiagramm ausgeben.

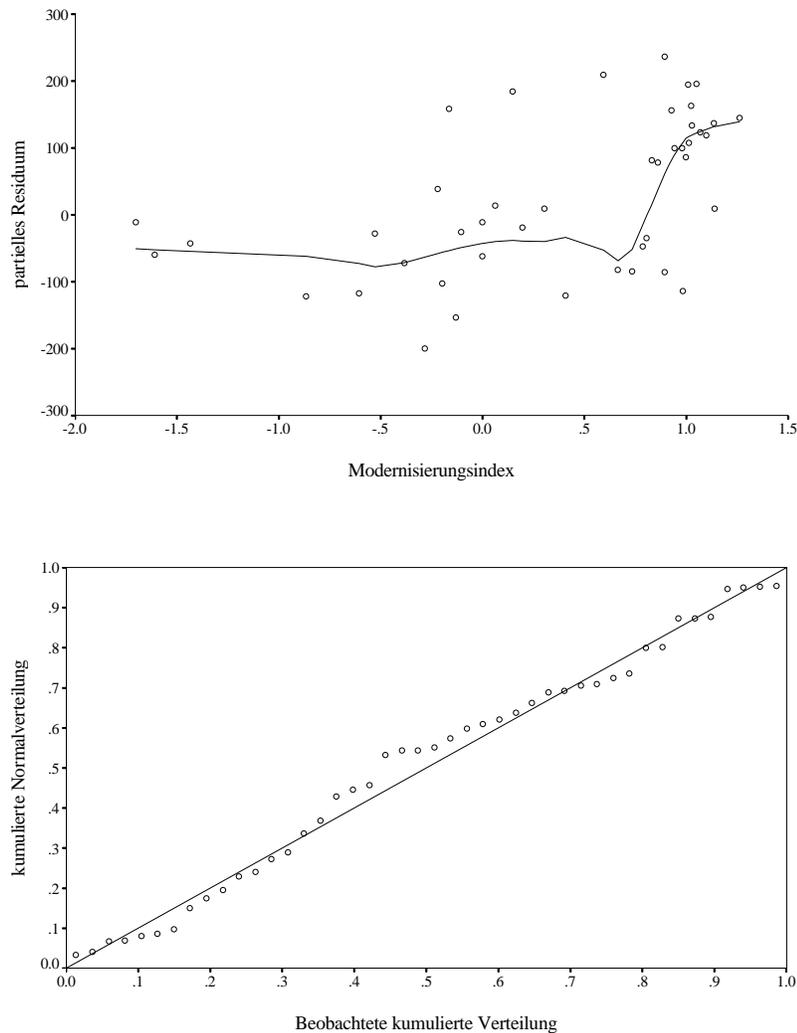


Abbildung 7: Residuen-Plot gegen \hat{Y} und Normal-Probability-Plot (N=44)

Es sei nochmals betont, daß nach einer solchen Transformation die Linearitätsannahme erneut untersucht werden muß. Wir machen es uns hier einfacher und berechnen die White-Kovarianzmatrix (mit LIMDEP; s. Greene, 1995: Kap. 15). Die Ergebnisse in Tabelle (1) in Spalte 3 zeigen, daß die Verzerrung der Standardfehler durch das Vorliegen der Heteroskedastizität offensichtlich kein großes Problem bei unserer Analyse darstellt. Im Gegenteil, die t-Werte der Erwerbsquote und des Modernisierungsindex werden sogar etwas größer.¹⁵

(3) *Normalverteilungsannahme:* Zum Schluß soll die Normalverteilungsannahme

¹⁵Man beachte, daß sich bei Anwendung des White-Verfahrens nur die Standardfehler des Modells verändern. Die Koeffizienten bleiben gleich. Dies ist bei einer Transformation nicht der Fall!

überprüft werden. Dazu betrachten wir das Normal-Probability-Plot in Abbildung (7) (von SPSS automatisch erstellt). Im Großen und Ganzen scheinen die Residuen einer Normalverteilung zu folgen. Dies bestätigt auch ein Kolmogorov-Smirnov-Test, der keine signifikante Abweichung feststellt. Eine Unregelmäßigkeit ist allerdings zu erkennen: zwischen 0,4 und 0,6 sind die Daten „unterbesetzt“ (dies erkennt man noch deutlicher, wenn man sich ein Histogramm der Residuen zeichnen läßt). Der Grund hierfür ist, daß die Länder in zwei Gruppen mit besonders niedrigen bzw. hohen Scheidungsraten zerfallen. Die Länder mit niedriger Scheidungsrate sind ausschließlich katholische Industriestaaten und Entwicklungsländer. Dies legt es nahe, eine Variable für den Katholikenanteil in das Modell mit aufzunehmen (für den Entwicklungsgrad haben wir ja bereits kontrolliert).

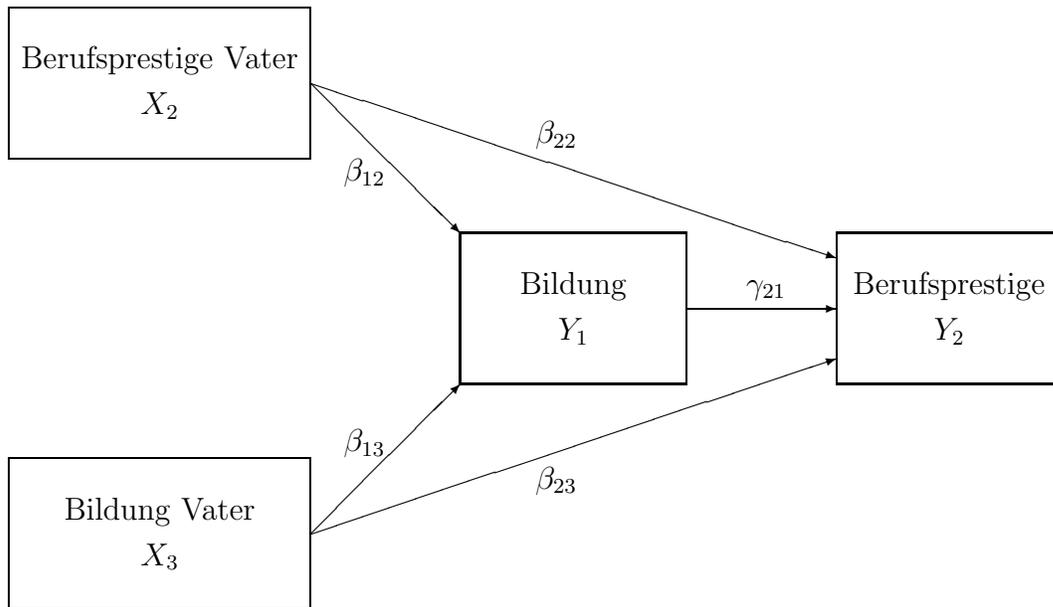
Nach all diesen Schritten hat man schließlich (hoffentlich) ein akzeptables Modell gefunden. Die obige Demonstration hat gezeigt, daß man im Verlauf einer Diagnose wesentlich besser mit den Daten und ihren Problemen vertraut wird. Besonders wichtig sind Erkenntnisse über einflußreiche Datenpunkte, Hinweise auf eine bessere Modellspezifikation und Hinweise auf zusätzliche Variablen, die in das Modell einbezogen werden sollten. Zum Schluß nochmals eine Warnung: Man hüte sich davor, jede Besonderheit, die man in den Daten entdeckt, zu modellieren. Viele dieser Besonderheiten sind nur von zufälliger Natur, und „Overfitting“ führt zu Modellen, die zwar die Stichprobendaten sehr gut anpassen, aber keine Entsprechung in der Grundgesamtheit haben.

2.3 Mehrgleichungssysteme

Bisher haben wir nur Schätzgleichungen für einzelne Modellgleichungen diskutiert. Solche Einzelgleichungsmodelle unterstellen eine sehr einfache Kausalstruktur: einige unabhängige Variablen zeigen jeweils einen kausalen Effekt auf *eine* abhängige Variable.¹⁶ Doch die soziale Welt ist komplex und entsprechend postulieren theoretische Ansätze oft wesentlich differenziertere Kausalstrukturen, die nicht mehr nur mit einer Gleichung modelliert werden können. Man benötigt Mehrgleichungssysteme. Ein Mehrgleichungssystem besteht aus mindestens zwei sogenannten Strukturgleichungen, die die theoretische Kausalstruktur abbilden. Die Literatur zu Mehrgleichungssystemen ist äußerst umfangreich. Besonders zu empfehlen sind Greene (1993: Kap. 20) und Judge et al. (1985: Kap. 14 und 15).

¹⁶Wir haben es bisher vermieden, von „Kausaleffekten“ zu sprechen. Wir verwendeten die schwächeren Begriffe „Effekt“, „Zusammenhang“ und „Einfluß“. Die Bedingungen, unter denen von einem Regressionseffekt auf einen Kausaleffekt geschlossen werden kann, sind kompliziert und auch umstritten (s. Sobel, 1995). Deshalb ist der leichtfertige und routinemäßige Gebrauch dieses Begriffs problematisch. Im Zusammenhang mit Mehrgleichungssystemen ist es allerdings üblich von Kausaleffekten zu sprechen, weshalb ich mich in diesem Abschnitt dieser Gepflogenheit anschließe.

Abbildung 8: Ein rekursives Simultangleichungssystem



Man kann vier Fälle unterscheiden. Im einfachsten Fall ist keine einzige abhängige Variable gleichzeitig unabhängige Variable in einer anderen Strukturgleichung. Das System besteht also aus mehreren (scheinbar) getrennten Regressionsgleichungen. Man spricht auch von *multivariater Regression* (in der ökonomischen Literatur auch „seemingly unrelated regression“, SURE). Sind die Fehlerterme der Gleichungen unkorreliert, so ist eine getrennte OLS-Schätzung jeder einzelnen Gleichung konsistent und effizient. Meist werden die Fehlerterme über die Gleichungen hinweg jedoch korreliert sein, weil etwa eine unbeobachtete Variable mehrere der abhängigen Variablen beeinflusst. Dann ist OLS nicht mehr effizient, denn man verschenkt bei der getrennten Schätzung der Gleichungen Information. Eine effiziente Schätzung ist allerdings mittels eines speziellen GLS-Verfahrens möglich (s. Greene, 1993: Kap. 17). Hat man also mehrere scheinbar getrennte Regressionsmodelle vorliegen, ein Fall der in der Forschungspraxis häufiger auftritt, so sollte man sich immer überlegen, ob die standardmäßige Verwendung von OLS sinnvoll ist.¹⁷

Komplizierter sind die folgenden drei Fälle, bei denen mindestens eine abhängige Variable in einer anderen Strukturgleichung als unabhängige Variable auftaucht. Man spricht dann von einem Simultangleichungssystem.¹⁸ Im einfachsten Fall

¹⁷Sind allerdings die Regressoren aller Gleichungen identisch, so entspricht OLS dem GLS-Verfahren (s. Greene, 1993: 488).

¹⁸Man beachte, daß ein solches System immer durch wechselseitiges Einsetzen der Gleichungen in die Form des ersten Falles gebracht werden kann (die sogenannte „reduzierte Form“). Theoretisch interessant sind allerdings die Parameter der Strukturform, nicht die der reduzier-

zeigt das Simultansystem einen „stufenförmigen“ Aufbau. Dies sei anhand des klassischen Statuszuweisungsmodells von Blau und Duncan (1967) in Abbildung (8) veranschaulicht. Ein Pfeil symbolisiert in dieser Abbildung einen gerichteten kausalen Effekt von einer Variable auf eine andere. Das Statuszuweisungsmodell postuliert, daß das Berufsprestige und die Bildung des Vaters einen Einfluß auf das Bildungsniveau einer Person haben. Dieses wiederum bestimmt zusammen mit den beiden Herkunftsvariablen das Berufsprestige. Dieses Modell ist ein sogenanntes *rekursives Simultangleichungssystem*, weil der erste Kausalprozeß (Bildungsteilnahme) zwar den zweiten Kausalprozeß (Statuszuweisung) beeinflusst, nicht aber umgekehrt. Das Modell in Abbildung (8) wäre nicht-rekursiv, wenn zusätzlich ein Pfeil von Y_2 nach Y_1 enthalten wäre. Der stufenförmige Aufbau eines rekursiven Systems wird besonders deutlich, wenn man die Regressionsgleichungen niederschreibt (Personenindex i weggelassen):

$$\begin{aligned} y_1 &= \beta_{11} + \beta_{12}x_2 + \beta_{13}x_3 + \epsilon_1 \\ y_2 &= \gamma_{21}y_1 + \beta_{21} + \beta_{22}x_2 + \beta_{23}x_3 + \epsilon_2 \quad . \end{aligned}$$

Dieses Modell ist ein Simultangleichungssystem, weil die abhängige Variable Y_1 gleichzeitig unabhängige Variable ist. Y_1 selbst hängt aber nur von exogenen Größen ab.¹⁹ Ist deshalb Y_1 bestimmt, so ergibt sich Y_2 rekursiv, indem man die erste Gleichung in die zweite einsetzt.

In der Forschungsliteratur wird zur Schätzung der Parameter eines rekursiven Systems standardmäßig OLS eingesetzt. Man muß sich allerdings bewußt sein, daß die Konsistenz der OLS-Schätzer auf einer wichtigen Annahme beruht: Die Fehlerterme der Gleichungen müssen unkorreliert sein (die Kovarianzmatrix Σ der Fehlerterme des Systems ist eine Diagonalmatrix). Wir haben aber bereits oben argumentiert, daß diese Annahme häufig nicht gelten wird, die Regressionsgleichungen nur „seemingly unrelated“ sind. Ist Σ keine Diagonalmatrix, so sind die OLS-Schätzer eines rekursiven Systems nicht konsistent. Man kann aber konsistente und effiziente Schätzer mittels GLS erhalten (s. Greene, 1993: 600f).²⁰

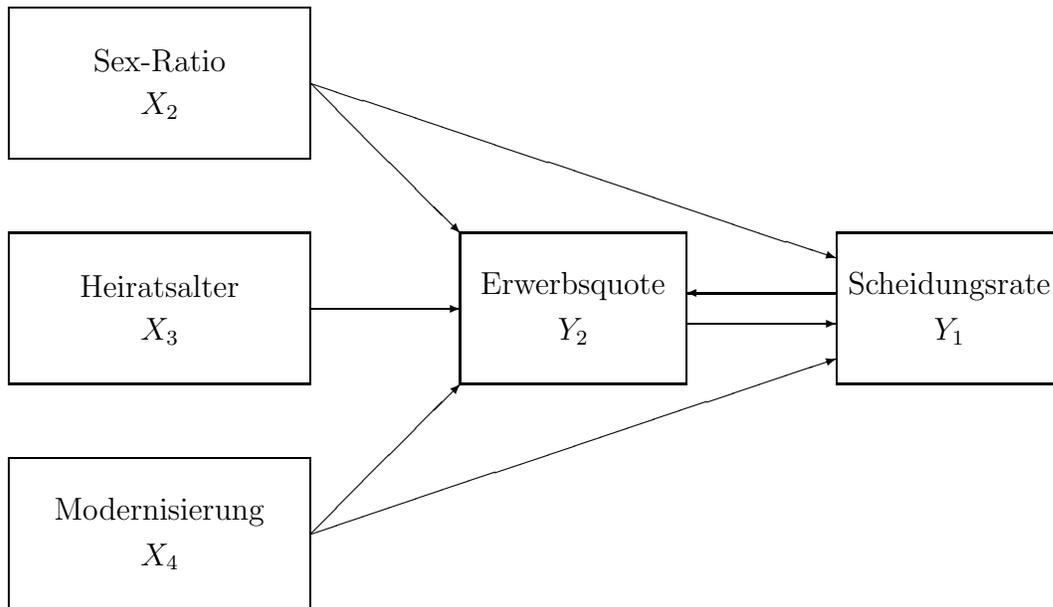
Ein *nicht-rekursives Simultangleichungssystem* ergibt sich, wenn die Kausaleffekte der endogenen Variablen nicht stufenförmig angeordnet sind. Betrachten wir unser Anwendungsbeispiel aus dem letzten Abschnitt. Dort haben wir argumentiert, daß die Sex-Ratio, die Erwerbsquote und das Modernisierungsniveau die Scheidungsrate beeinflussen. Nun gibt es aber gute Gründe für die Annahme, daß die Erwerbsquote endogen ist, d.h. ebenfalls von der Sex-Ratio und dem Modernisierungsniveau beeinflusst wird. Zusätzlich berücksichtigen wir eine weitere exogene

ten Form, weshalb die Fallunterscheidung anhand der Strukturform durchaus sinnvoll ist.

¹⁹Die X -Variablen nennt man „exogen“, weil sie nicht durch eine Strukturgleichung bestimmt werden. Die Y -Variablen dagegen sind „endogen“, durch das System bestimmt.

²⁰Ist allerdings Σ keine Diagonalmatrix, so ist ein rekursives System, in dem alle exogenen Variablen in jeder Gleichung enthalten sind, nicht mehr identifiziert (s.u.). Man muß in diesem Fall einige Parameter restringieren.

Abbildung 9: Ein nicht-rekursives Simultangleichungssystem



Variable (das durchschnittliche Heiratsalter von Frauen), weil zu vermuten ist, daß mit höherem Heiratsalter die Frauenerwerbsquote ansteigt. Die bisherigen Überlegungen führen zu einem rekursiven System. Weiterhin kann jedoch vermutet werden, daß die Scheidungsrate auch die Erwerbsquote beeinflusst, denn bei hohen Scheidungsraten werden Frauen vermehrt in der Erwerbstätigkeit verbleiben, um gegen das erhöhte Scheidungsrisiko abgesichert zu sein. Mit dieser zusätzlichen Hypothese wird unser System nicht-rekursiv. Dieses Strukturmodell ist in Abbildung (9) dargestellt. Die Nicht-Rekursivität erkennt man an den wechselseitigen Kausaleffekten der beiden endogenen Variablen. Zur Schätzung der Parameter nicht-rekursiver Systeme benötigt man spezielle Schätzverfahren, von denen wir einige in Abschnitt 2.3.1 kennenlernen werden.

Eine weitere Verkomplizierung resultiert, wenn man in Simultangleichungssystemen latente Variablen, d.h. nicht direkt meßbare Variablen, berücksichtigt. Man bezeichnet solche Strukturgleichungsmodelle mit latenten Variablen häufig als *LISREL-Modelle* (nach dem am weitesten verbreiteten Schätzprogramm für diese Modelle). Der LISREL-Ansatz ist allerdings zu komplex, um hier vorgestellt werden zu können. Das Standardwerk zum LISREL-Ansatz ist Bollen (1989). Eine nicht-technische Einführung findet man bei Hoyle (1995).

2.3.1 Schätzverfahren für nicht-rekursive Simultangleichungssysteme

Bevor man die Parameter eines nicht-rekursiven Systems schätzen kann, muß sichergestellt sein, daß es *identifiziert* ist. Was damit gemeint ist, wollen wir anhand des Modells von Abbildung (9) verdeutlichen (ausführlich hierzu Berry, 1984). Die Strukturform dieses Modells lautet (wir lassen X_2 und X_4 der Einfachheit halber weg):

$$\begin{aligned} y_1 &= \gamma_{11}y_2 + \beta_{11} + \epsilon_1 \\ y_2 &= \gamma_{21}y_1 + \beta_{21} + \beta_{23}x_3 + \epsilon_2 \end{aligned} .$$

Das Modell enthält fünf Parameter, die es zu schätzen gilt. Ein Modell heißt identifiziert, wenn die Parameter der Struktur anhand der vorliegenden Information über die exogenen und endogenen Variablen eindeutig bestimmbar sind. Die Identifiziertheit eines Modells kann man überprüfen, indem man die Gleichungen wechselseitig ineinander einsetzt und so umformt, daß auf der rechten Seite nur noch exogene Variablen stehen. In unserem Beispiel lautet diese sogenannte reduzierte Form des Modells:

$$\begin{aligned} y_1 &= \frac{\beta_{11} + \gamma_{11}\beta_{21}}{1 - \gamma_{11}\gamma_{21}} + \frac{\gamma_{11}\beta_{23}}{1 - \gamma_{11}\gamma_{21}}x_3 + \frac{\epsilon_1 + \gamma_{11}\epsilon_2}{1 - \gamma_{11}\gamma_{21}} = \pi_{11} + \pi_{12}x_3 + \nu_1 \\ y_2 &= \frac{\gamma_{21}\beta_{11} + \beta_{21}}{1 - \gamma_{11}\gamma_{21}} + \frac{\beta_{23}}{1 - \gamma_{11}\gamma_{21}}x_3 + \frac{\gamma_{21}\epsilon_1 + \epsilon_2}{1 - \gamma_{11}\gamma_{21}} = \pi_{21} + \pi_{22}x_3 + \nu_2. \end{aligned}$$

Die reduzierte Form hat nur vier Parameter. Damit ist klar, daß die fünf Parameter der Strukturform nicht bestimmbar sind, das Modell ist nicht identifiziert. Allerdings sieht man, daß für die Parameter der ersten Strukturgleichung gilt:

$$\gamma_{11} = \frac{\pi_{12}}{\pi_{22}} \quad \text{und} \quad \beta_{11} = \pi_{11} - \pi_{21} \frac{\pi_{12}}{\pi_{22}}.$$

Das bedeutet, daß die erste Strukturgleichung identifiziert ist. γ_{11} und β_{11} können geschätzt werden, nicht jedoch die drei Parameter der zweiten Strukturgleichung.

Allgemein ist eine Struktur nur dann identifiziert, wenn einige der γ 's und β 's restringiert sind. Meist wird die Identifikation über Nullrestriktionen sichergestellt, d.h. endogene oder exogene Variablen werden aus Gleichungen ausgeschlossen. Welche Variablen in einer Gleichung nicht berücksichtigt werden, folgt im Idealfall aus der zugrundeliegenden Theorie. Ob die Zahl der Restriktionen ausreicht, darüber gibt das Abzählkriterium Auskunft: Eine Gleichung ist identifiziert, wenn die Zahl der Nullrestriktionen mindestens genauso groß ist, wie die Zahl der Gleichungen im System minus eins.²¹ Wenden wir das Abzählkriterium auf das obige Beispiel an, so sehen wir, daß in der ersten Gleichung X_3 ausgeschlossen wurde. Mithin haben wir eine Restriktion, was der Gleichungszahl

²¹Das Abzählkriterium ist nur eine notwendige Bedingung für die Identifikation. Es gibt weitere, kompliziertere Kriterien, die notwendig und hinreichend sind. In fast allen praktischen Situationen ist jedoch das Abzählkriterium auch hinreichend.

minus eins entspricht. Die zweite Gleichung ist allerdings mit keiner Restriktion belegt, weshalb dieselbe auch nicht identifiziert ist. Hilfreich ist manchmal auch die folgende Daumenregel: Hat jede Gleichung des Systems eine *eigene* exogene Variable (die exogene Variable taucht in keiner anderen Gleichung auf), so ist das ganze System identifiziert.²² Betrachten wir das volle Modell aus Abbildung (9). Die Erwerbsquotengleichung hat eine eigene exogene Variable (das Heiratsalter), nicht aber die Scheidungsratengleichung. Deshalb ist das gesamte System nicht identifiziert.²³ Würden wir aber etwa die Sex-Ratio aus der Erwerbsquotengleichung herausnehmen, so hätte auch die Scheidungsratengleichung eine eigene exogene Variable, beide Gleichungen wären identifiziert. In größeren Systemen ist es oft nicht einfach, die Identifikationsbedingungen zu überprüfen. Ein einfacher Weg zur Sicherstellung der Identifikation besteht nach unserer Daumenregel aber darin, jeder Gleichung ihre eigene exogene Variable zu geben.

Betrachten wir schließlich noch das rekursive System aus Abbildung (8). Die erste Gleichung weist eine Restriktion auf und ist deshalb identifiziert. Die zweite Gleichung ist nicht identifiziert, weil sie keine Restriktion aufweist. Dennoch werden solche rekursiven Systeme standardmäßig geschätzt. Wie ist dies möglich, wenn Gleichungen nicht identifiziert sind? Der Grund liegt in der oben erwähnten Annahme über Σ : Die zusätzliche Restriktion, die das System identifizierbar macht, ist die Annahme, daß die Fehlerterme nicht korreliert sind. Mithin ermöglichen manchmal auch Restriktionen über Σ die Modellidentifikation.

Wenden wir uns nun der *Schätzung der Parameter* einer einzelnen, identifizierten Gleichung zu.²⁴ Wir notieren die l -te Gleichung in Matrixschreibweise:

$$\mathbf{y}_l = \mathbf{Y}_l \boldsymbol{\gamma}_l + \mathbf{X}_l \boldsymbol{\beta}_l + \boldsymbol{\epsilon}_l.$$

Fassen wir die endogenen und exogenen Variablen auf der rechten Seite der Gleichung in der Matrix \mathbf{Z}_l und die beiden Parametervektoren im Vektor $\boldsymbol{\delta}_l$ zusammen, so können wir auch schreiben:

$$\mathbf{y}_l = \mathbf{Z}_l \boldsymbol{\delta}_l + \boldsymbol{\epsilon}_l.$$

Der OLS-Schätzer $\hat{\boldsymbol{\delta}}_l = (\mathbf{Z}_l' \mathbf{Z}_l)^{-1} \mathbf{Z}_l' \mathbf{y}_l$ ist inkonsistent, weil die endogenen Variablen auf der rechten Seite der Gleichung mit dem Fehlerterm korreliert sind (nicht

²²Diese Daumenregel ist sehr „streng“: es gibt natürlich auch identifizierte Systeme, bei denen nicht jede Gleichung eine eigene exogene Variable aufweist.

²³Man beachte, daß in diesem Beispiel gerade die Gleichung identifiziert ist, die keine eigene exogene Variable hat.

²⁴Wir stellen im folgenden nur sogenannte „Einzelgleichungsverfahren“ vor. Verfahren, die die Parameter des ganzen Systems gemeinsam schätzen („full information estimators“), diskutieren wir nicht (nachzulesen z.B. bei Greene, 1993: 610ff). Diese Verfahren haben zwar Effizienzvorteile gegenüber den Einzelgleichungsverfahren, sind aber zu komplex, um hier dargestellt zu werden.

nur $\hat{\gamma}_l$ ist inkonsistent, auch $\hat{\beta}_l$). Dies verletzt die Annahme A5, die Voraussetzung für die Konsistenz des OLS-Schätzers ist. Man erkennt das Problem, wenn man die reduzierte Form unseres obigen Beispiels betrachtet: Jeder Fehlerterm der reduzierten Form ist eine Funktion aller Fehlerterme der Strukturgleichungen. Mithin korrelieren alle endogenen Kovariate mit den Fehlertermen der Strukturgleichungen.

Anhand der reduzierten Form ergibt sich eine indirekte Möglichkeit, zu konsistenten Schätzern zu gelangen: man schätzt mittels OLS die reduzierte Gleichung und errechnet aus dem so gewonnenen $\hat{\pi}_l$ sowohl $\hat{\gamma}_l$ als auch $\hat{\beta}_l$. Dies ist allerdings nur dann eindeutig möglich, wenn die Gleichung exakt identifiziert ist. Ist sie überidentifiziert, d.h. die Zahl der Restriktionen übersteigt die Zahl der Gleichungen minus eins, so gibt es mehrere Lösungen. Überidentifikation von Systemen ist ein durchaus häufiger Fall, weshalb dieses indirekte Verfahren nur selten angewendet wird.

Einen konsistenten Schätzer von δ_l erhält man auch mit der sogenannten *Instrumentvariablen-Methode* (IV-Schätzer). Man ersetzt bei dieser Methode jede endogene Variable in \mathbf{Z}_l durch jeweils eine Instrumentvariable, die nicht mit dem Fehlerterm, aber möglichst hoch mit der zu ersetzenden Variable korreliert ist. Faßt man diese Instrumentvariablen und die exogenen Variablen in der Matrix \mathbf{W}_l zusammen, so ist der folgendermaßen definierte IV-Schätzer konsistent und effizient:

$$\hat{\delta}_{l,IV} = (\mathbf{W}_l' \mathbf{Z}_l)^{-1} \mathbf{W}_l' \mathbf{y}_l.$$

Woher bekommt man aber die Instrumentvariablen? Ist die l -te Gleichung exakt identifiziert, so folgt aus dem Abzählkriterium, daß die Zahl der in Gleichung l auf der rechten Seite enthaltenen endogenen Variablen gleich der Zahl der ausgeschlossenen exogenen ist. Mithin kann jede endogene Variable durch genau eine exogene Instrumentvariable ersetzt werden. Die Matrix \mathbf{W}_l besteht somit aus allen exogenen Variablen \mathbf{X} des Systems. Der „*instrumental least squares*“ (ILS) Schätzer lautet damit:

$$\hat{\delta}_{l,ILS} = (\mathbf{X}' \mathbf{Z}_l)^{-1} \mathbf{X}' \mathbf{y}_l.$$

Oft wird aber die zu schätzende Gleichung überidentifiziert sein, d.h. man hat mehr ausgeschlossene exogene Variablen zur Verfügung, als endogene in der Gleichung sind. In diesem Fall könnten mehrere, unter Umständen verschiedene ILS-Schätzer konstruiert werden. Eine naheliegende Idee ist es deshalb, als Instrument eine Linearkombination der exogenen Variablen zu verwenden. Eine Möglichkeit ist der Vorhersagewert der zu ersetzenden endogenen Variable aus einer Regression auf die exogenen Variablen. Man kann zeigen, daß ein asymptotisch effizienter Schätzer resultiert, wenn man \mathbf{Y}_l mit $\hat{\mathbf{Y}}_l$ instrumentiert ($\hat{\mathbf{Y}}_l$ erhält man aus den Regressionen auf *alle* exogenen Variablen). Dieser „*two stage least squares*“

(2SLS) Schätzer lautet somit:

$$\hat{\delta}_{l,2SLS} = (\hat{\mathbf{Z}}_l' \mathbf{Z}_l)^{-1} \hat{\mathbf{Z}}_l' \mathbf{y}_l.$$

$\hat{\mathbf{Z}}_l$ notiert dabei die Matrix $(\hat{\mathbf{Y}}_l, \mathbf{X}_l)$. Die asymptotische Kovarianzmatrix der Schätzer lautet

$$\hat{V}(\hat{\delta}_{l,2SLS}) = \hat{\sigma}_l^2 (\hat{\mathbf{Z}}_l' \hat{\mathbf{Z}}_l)^{-1},$$

wobei zu beachten ist, daß $\hat{\sigma}_l^2$ nicht mit $\hat{\mathbf{Z}}_l$, sondern mit den Originaldaten \mathbf{Z}_l zu berechnen ist.

Der Name 2SLS rührt daher, daß man diesen Schätzer auch erhalten kann, indem man in einer ersten Stufe die reduzierte Form für jede endogene Variable mit OLS schätzt, mittels dieser Ergebnisse $\hat{\mathbf{Y}}_l$ errechnet, und in einer zweiten Stufe mit OLS $\hat{\delta}_l$ schätzt (indem man \mathbf{y}_l auf $\hat{\mathbf{Y}}_l$ und \mathbf{X}_l regressiert). Dieses zweistufige Vorgehen ist mit jedem Programm, das eine OLS Prozedur bietet, durchführbar. Man beachte aber, daß bei dieser 2SLS-Regression „per Hand“ die Standardfehler falsch berechnet werden, weil sie auf der Basis von $\hat{\mathbf{Z}}_l$ errechnet werden. Um auch korrekte Standardfehler zu erhalten, muß man ein spezielles Programm für 2SLS verwenden.

Zum Schluß noch einige Worte über die *praktische Anwendung* dieser Verfahren. Alle Aussagen zur Konsistenz und Effizienz der IV-Schätzer sind asymptotischer Natur. Das bedeutet, daß deren Verhalten in finiten Stichproben nur schwer abzuschätzen ist. Insbesondere ist es durchaus möglich, daß in kleinen bis mittleren Stichproben der OLS-Schätzer auch nicht „schlechter“ ist als der IV-Schätzer. Deshalb verwundert es nicht, wenn trotz seiner Inkonsistenz der OLS-Schätzer der am häufigsten verwendete Schätzer für Simultangleichungssysteme ist (s. Johnston, 1984: 492). Weiterhin zeigt die Erfahrung, daß die Simultangleichungsschätzer sehr empfindlich auf die Modellspezifikation reagieren (insbesondere die noch komplexeren LISREL-Modelle scheinen in diesem Sinne nicht sehr robust zu sein). Diese Probleme rechtfertigen allerdings keinesfalls den bedenkenlosen Einsatz von OLS (etwa weil das verwendete Programm keine IV-Schätzer zur Verfügung stellt). Man sollte sich der Probleme von OLS bewußt sein und auch mit IV-Schätzern „experimentieren“.

2.3.2 Pfadanalyse

Hat man die Parameter eines Strukturmodells konsistent geschätzt, so kann man diese Schätzer zur genaueren Analyse der Kausalstruktur des Modells einsetzen. Man spricht von *Pfadanalyse* (s. Asher, 1983; Dillon und Goldstein, 1984: Kap. 12). Betrachten wir noch einmal das Modell von Abbildung (8). Die β 's und γ 's repräsentieren die linearen Effekte bei Erhöhung der unabhängigen Variable um eine Einheit. Da die Maßeinheiten der Variablen aber im Normalfall ganz

unterschiedlich sein werden, können die Effekte verschiedener Variablen nicht miteinander verglichen werden. Damit dies möglich ist, kann man die Regressionskoeffizienten jedoch standardisieren. Zum Beispiel errechnet man den standardisierten Effekt des Berufsprestiges des Vaters auf das Berufsprestige einer abhängig beschäftigten Person als

$$\beta_{22}^* = \beta_{22} \frac{\sigma_{x_2}}{\sigma_{y_2}}.$$

Ein standardisierter Regressionskoeffizient gibt mithin an, um wieviele Standard-einheiten sich Y verändert, wenn sich X um eine Standardeinheit erhöht. Damit sind die β^* vergleichbar und zeigen die relative Stärke des direkten Kausaleffektes an. Man trägt diese standardisierten Koeffizienten üblicherweise im Kausaldiagramm an den entsprechenden Pfeilen auf. Von daher rührt die Bezeichnung „Pfadkoeffizient“. Das Kausaldiagramm zusammen mit den Pfadkoeffizienten (Pfaddiagramm) ist ein äußerst hilfreiches Mittel, um die Kausalstruktur eines Modells näher zu beleuchten. Man kann an der Höhe der Pfadkoeffizienten sofort ablesen, welche Beziehungen bedeutsam und welche nur marginal sind. Weiterhin ist es möglich, den gesamten Kausaleffekt einer Variable zu ermitteln. Eine Variable in einem Simultangleichungssystem zeigt ja nicht nur einen direkten Effekt, der durch den Regressionskoeffizienten ausgedrückt wird, sondern auch noch indirekte Effekte, die über intervenierende Variablen laufen. Die Stärke eines indirekten Effektes ergibt sich aus dem Produkt der entsprechenden Pfadkoeffizienten. Z.B. ist der gesamte Kausaleffekt des Berufsprestiges des Vaters auf das Berufsprestige einer Person

$$\beta_{22}^* + \beta_{12}^* \gamma_{21}^*.$$

Der indirekte Effekt ergibt sich in diesem Beispiel aus der Multiplikation des direkten Effektes des Berufsprestiges des Vaters auf die Bildung und deren direkten Effekt auf das Berufsprestige. Sind alle Effekte positiv, so bedeutet dies, daß das Berufsprestige des Vaters nicht nur „direkt vererbt“ wird, sondern daß es auch vermittelt über das Bildungsniveau das Berufsprestige indirekt fördert. In Fällen, in denen ein Pfadkoeffizient negativ ist, kann es auch dazu kommen, daß sich direkter und indirekter Effekt „neutralisieren“. Trotz eines vorhandenen direkten Effektes kann der Gesamteffekt verschwindend sein. Die genaue Analyse des Pfaddiagramms ist ein einfaches Mittel, um zu solch aufschlußreichen Ergebnissen zu gelangen.

2.3.3 Anwendungsbeispiel: Scheidungsraten in 50 Ländern

Zur Demonstration dieser Verfahren führen wir das Scheidungsbeispiel aus dem vorigen Abschnitt fort. Oben haben wir einige weitere Überlegungen zu diesem Beispiel angeführt, die zu dem Modell von Abbildung (9) führten. Inhaltlich interessieren uns die Parameter der Scheidungsgleichung. Die OLS-Schätzer, die

wir im vorigen Abschnitt präsentiert haben, sind eventuell verzerrt, weil gemäß unseren Überlegungen die Erwerbsquote endogen ist. Deshalb liegt es nahe, einen IV-Schätzer einzusetzen. Die Scheidungsgleichung ist, wie wir oben bereits sahen, exakt identifiziert (eine Restriktion bei zwei Gleichungen im System). Deshalb können wir den ILS-Schätzer eindeutig bestimmen. Als Instrumente verwenden wir die Sex-Ratio, das durchschnittliche Heiratsalter von Frauen und den Modernisierungsindex. Das Heiratsalter ist die aus der Gleichung ausgeschlossene exogene Variable, die die endogene Erwerbsquote instrumentiert. Hier muß man sich natürlich noch die Frage stellen, ob das Heiratsalter ein gutes Instrument ist. Die Korrelation zwischen dem Heiratsalter und der Erwerbsquote ist mit +0,27 eher mäßig. Außerdem könnte man vermuten, daß auch das Heiratsalter endogen ist, denn eine höhere Scheidungsrate könnte einen Anstieg des Heiratsalters nach sich ziehen, weil die Frauen intensiver und länger nach Heiratspartnern suchen, um ihr individuelles Scheidungsrisiko möglichst niedrig zu halten. Diese Überlegungen deuten darauf hin, daß das Instrument Heiratsalter nicht unproblematisch ist. Dennoch müssen wir mit diesem Instrument Vorlieb nehmen, weil unser Datensatz kein Besseres enthält (ein häufiges Problem). Für Jamaika liegt das Heiratsalter nicht vor, weshalb sich die Fallzahl auf 43 Länder reduziert. Die Schätzer wurden mit LIMDEP (s. Greene, 1995: Kap. 19) errechnet.

In Spalte (4) von Tabelle (1) sind die Ergebnisse der ILS-Schätzung angeführt. Die Koeffizienten der Erwerbsquote und des Modernisierungsindex verändern sich gegenüber OLS (Spalte (2)) kaum. Der Koeffizient der Sex-Ratio ist nun negativ, was aber aufgrund des kleinen t-Wertes nicht überinterpretiert werden sollte.²⁵ Insgesamt zeigt sich somit für unser Anwendungsbeispiel, daß sich durch die Verwendung von ILS die Koeffizienten kaum verändern. Man kann auch das 2SLS-Verfahren einsetzen und in der ersten Stufe die reduzierte Form der Erwerbsquotengleichung schätzen (mit Sex-Ratio, Heiratsalter und Modernisierung als unabhängige Variablen), daraus für jedes Land die vorhergesagte Erwerbsquote errechnen, und schließlich die Scheidungsgleichung mit der vorhergesagten Erwerbsquote mittels OLS schätzen. Die 2SLS-Koeffizienten sind identisch mit den oben für ILS berichteten. Allerdings sind die so erhaltenen (nicht korrekten!) Standardfehler noch größer. Dies demonstriert, daß bei exakt identifizierten Gleichungen ILS und 2SLS äquivalent sind. Bei überidentifizierten Gleichungen trafe dies nicht zu, wie wir oben bemerkt haben.

Auffällig an den ILS-Schätzern sind die niedrigen t-Werte. Die Standardfehler der Erwerbsquoten- und Modernisierungsindex-Schätzer haben sich beinahe verdoppelt. Der Grund hierfür ist hohe Multikollinearität, wie anhand der \sqrt{VIF} -Werte von über zwei ablesbar ist. Dieses Problem hat man bei der Anwendung von ILS bzw. 2SLS häufig, weil die endogene Variable durch eine Linearkombination der

²⁵Die Vorzeichenumkehr ist nicht auf die Verwendung von ILS zurückzuführen, sondern auf den Ausschluß von Jamaika, denn ohne Jamaika liefert OLS auch einen Koeffizienten von -2,1.

exogenen Variablen ersetzt wird. Hat man (wie in unserem Fall) nur wenige exogene Variablen zur Verfügung, die nicht in der zu schätzenden Gleichung enthalten sind, so wird diese Linearkombination hoch mit den anderen unabhängigen Variablen der Gleichung korreliert sein. In diesem Fall gibt es nur zwei Möglichkeiten, die Schätzung trotz Multikollinearität präziser zu machen, d.h. den Standardfehler zu reduzieren: Man kann mehr Daten sammeln (mit höherer Fallzahl steigt die Präzision der Schätzer), oder man erweitert das Modell um weitere (gute!) Instrumentvariablen.

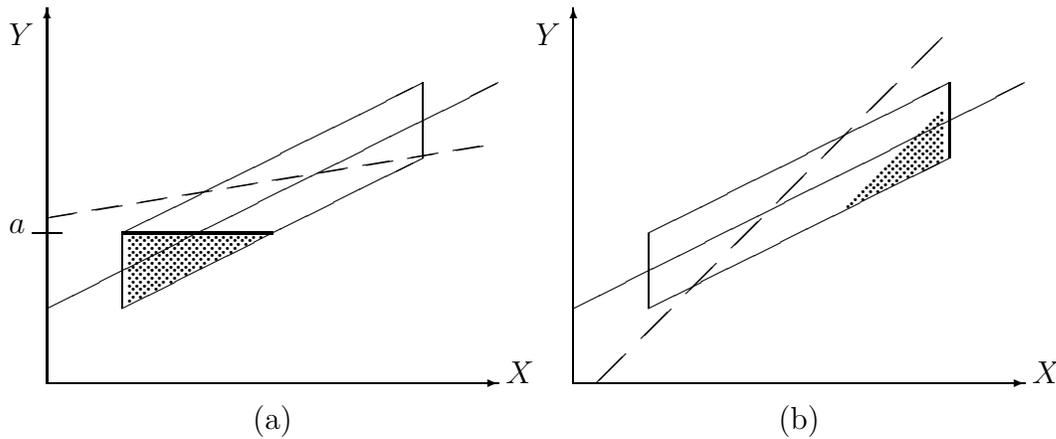
2.4 Begrenzte abhängige Variablen

In diesem Abschnitt betrachten wir Situationen, in denen die abhängige Variable „begrenzt“ ist. Von Begrenzung der abhängigen Variable spricht man, wenn sie „gestutzt“ oder „zensiert“ ist. Stutzung bedeutet, daß für Beobachtungen jenseits einer bestimmten Schwelle keine Information vorliegt (sie sind nicht in der Stichprobe enthalten, man hat somit auch keine Information über die unabhängigen Variablen). Zensierung ist gegeben, wenn die Beobachtungen jenseits einer Schwelle zwar in der Stichprobe enthalten sind, aber für die abhängige Variable nur der Schwellenwert bekannt ist. Weiterhin muß man zwischen direkter und indirekter Stutzung bzw. Zensierung unterscheiden. Bei einer direkten Begrenzung ist der Begrenzungsmechanismus eine Funktion der abhängigen Variable selbst, bei indirekter Begrenzung ist er eine Funktion einer unabhängigen Variable. Konsequenz einer Begrenzung ist, daß die OLS-Schätzer inkonsistent sind. Es gibt viele Gründe, weshalb abhängige Variablen gestutzt oder zensiert sein können. Wir betrachten hier nur zwei mögliche Situationen.

(1) *Die abhängige Variable ist direkt zensiert*, d.h. für Beobachtungen, die unter (oder über) eine bestimmte Schwelle fallen, ist nur der Wert der Schwelle bekannt. Ein Beispiel hierfür sind die Ausgaben von Haushalten für langlebige Gebrauchsgüter. Für viele Haushalte in einer Konsumstichprobe werden wir nur die Schwelle 0 DM beobachten. Tobin (1958) war der erste, der für diese Anwendung ein Regressionsmodell vorschlug, das deshalb *Tobit-Modell* heißt.

(2) *Die Stichprobe ist ausgewählt*, d.h. ein nicht-zufälliger Auswahlmechanismus sorgt dafür, daß nur für bestimmte Fälle eine Information zur abhängigen Variable vorliegt. Man spricht auch vom Problem der Stichprobenauswahl (sample selection bias). Stichprobenauswahl ist entweder eine indirekte Stutzung der abhängigen Variable (über die nicht-ausgewählten Beobachtungen liegt keinerlei Information vor), oder eine indirekte Zensierung (für die nicht-ausgewählten Beobachtungen hat man Information über unabhängige Variablen). Lohnregressionen waren der Anfang der Diskussion zu diesem Problem, denn nur für erwerbstätige Personen hat man eine Lohninformation vorliegen. Bei nicht erwerbstätigen Personen beobachten wir keinen Lohn, denn der für sie erreichbare Marktlohn liegt

Abbildung 10: Verzerrung des OLS-Schätzers durch Zensierung (a) und Stichprobenauswahl (b)



unter ihrem Reservationslohn. Heckman (1979) hat ein Verfahren vorgeschlagen, daß auch bei nicht-zufälliger Stichprobenauswahl konsistente Schätzer liefert (das *Heckit-Verfahren*).

Zur Verdeutlichung der *Inkonsistenz von OLS* in diesen beiden Situationen betrachte man die Abbildung (10) (vgl. auch Berk, 1983). Das Parallelogramm in beiden Abbildungen symbolisiert die Punktwolke der Datenpaare. Die durchgezogene Linie repräsentiert die „wahre“ Regressionsgerade, wie man sie mittels OLS erhält, wenn keine Probleme mit einer Begrenzung der abhängigen Variable vorliegen. In Abbildung (10a) ist die abhängige Variable beim Wert a zensiert, d.h. für $y \leq a$ gilt $y = a$ (die Beobachtungen der schattierten Fläche befinden sich auf der dicken Linie). Ist Y beispielsweise der Lohn und X die Bildung einer Person, so könnte a etwa ein Mindestlohn sein. Benutzt man die solchermaßen zensierten Daten, um mit OLS den Zusammenhang von Y und X zu schätzen, so erhält man die gestrichelte Regressionsgerade, die flacher ist als die „wahre“ Regressionsgerade (die Bildungsrendite wird unterschätzt). Bei Vorliegen von Zensierung ist OLS somit inkonsistent. Dies gilt auch im Fall der Stichprobenauswahl, der in Abbildung (10b) dargestellt ist. Die schattierte Fläche bedeutet hier, daß mit zunehmendem X für Beobachtungen mit niedrigen Y -Werten die Wahrscheinlichkeit in die Stichprobe zu gelangen sinkt (die Beobachtungen der schattierten Fläche sind in der Stichprobe nicht enthalten). In unserem Beispiel kann diese Situation auftreten, wenn mit steigender Bildung der Reservationslohn schneller ansteigt als der Marktlohn. Dann werden die Personen mit hoher Bildung, die nur einen unter dem Reservationslohn liegenden Marktlohn erzielen können, nicht erwerbstätig sein. Konsequenz wird sein, daß die OLS-Regressionsgerade der nicht-zufällig ausgewählten Stichprobe zu steil ist (die Bildungsrendite wird

überschätzt).

Insbesondere das *Problem der Stichprobenauswahl* ist in der empirischen Sozialforschung weit verbreitet. Angesichts der üblichen Ausfallquoten (20% bis 50%) bei Befragungen hat im Prinzip jede Befragungsstudie mit diesem Problem zu kämpfen. Manche Autoren (z.B. Berk, 1983) fordern deshalb, daß Sozialforscher routinemäßig von einer nicht-zufällig ausgewählten Stichprobe ausgehen und nur bei Beweis des Gegenteils die Standard-Verfahren eingesetzt werden sollten. Aufgrund solcher Forderungen fand das Problem der Stichprobenauswahl in den letzten Jahren auch in der empirischen Sozialforschung vermehrt Beachtung, weshalb es inzwischen eine Reihe guter Einführungsartikel hierzu gibt (z.B. Dubin und Rivers, 1989; Winship und Mare, 1992).

Wie erwähnt, diskutieren wir hier nur zwei relativ einfache Situationen, in denen begrenzte abhängige Variablen auftreten können. Weitere Möglichkeiten sind etwa: die abhängige Variable kann zweiseitig zensiert sein; sie liegt nur in gruppierter Form vor (alle Werte sind zensiert!); sie ist gestutzt, d.h. Beobachtungen jenseits einer bestimmten Schwelle fehlen ganz; die Stichprobenauswahl wird von mehreren Prozessen gesteuert. Unter dem Stichwort „limited dependent variable models“ findet man in der ökonometrischen Literatur inzwischen eine Vielzahl guter Darstellungen von Verfahren für diese Situationen (z.B. Maddala, 1983: Kap. 6; Amemiya, 1985: Kap. 10; Ronning, 1991: Kap. 3; Greene, 1993: Kap. 22; Breen, 1996). Die meisten dieser Situationen können auch als „missing data“ Problem gesehen werden, weshalb viele der Verfahren auch in der Literatur über fehlende Daten behandelt werden (z.B. Little und Rubin, 1987). LIMDEP (Greene, 1995: Kap. 27 und 28) enthält Schätzprogramme für eine große Zahl von Modellen mit begrenzten abhängigen Variablen.

2.4.1 Zensierte abhängige Variable: Das Tobit-Modell

Y^* sei die nicht-zensierte abhängige Variable, für die wir folgendes Regressionsmodell spezifizieren:

$$y_i^* = \beta' \mathbf{x}_i + \epsilon_i.$$

Es sollen alle Annahmen des klassischen Regressionsmodells gelten, insbesondere $\epsilon_i \sim N(0, \sigma^2)$. Nun ist Y^* aber eine latente Variable, denn wir können nur die zensierte Variable Y beobachten, für die gilt

$$\begin{aligned} y_i &= 0, & \text{wenn } y_i^* &\leq 0, \\ y_i &= y_i^*, & \text{wenn } y_i^* &> 0. \end{aligned}$$

Dies ist das klassische *Tobit-Modell* mit Zensierung bei null. Das Modell läßt sich analog für jede beliebige Schwelle formulieren. Der bedingte Erwartungswert für die latente Variable ist wie im linearen Regressionsmodell $E(y_i^* | \mathbf{x}_i) = \beta' \mathbf{x}_i$. Für

die zensierte Variable gilt dies allerdings nicht mehr. Dies wollen wir der Einfachheit halber hier nur für die nicht-zensierten Beobachtungen demonstrieren (die Berücksichtigung der zensierten Beobachtungen würde nichts am folgenden Argument verändern, aber die Formeln komplexer machen). Für die Beobachtungen mit $y_i > 0$ ist der bedingte Erwartungswert gegeben durch (dies folgt aus der Formel für den Erwartungswert einer „gestutzten“ Normalverteilung, s. Greene, 1993: 685):

$$E(y_i | y_i > 0, \mathbf{x}_i) = \boldsymbol{\beta}' \mathbf{x}_i + \sigma \lambda \left(-\frac{\boldsymbol{\beta}' \mathbf{x}_i}{\sigma} \right),$$

wobei $\lambda(z) = \phi(z)/(1-\Phi(z))$ ist und $\phi(\cdot)$ die Dichte und $\Phi(\cdot)$ die Verteilungsfunktion der Standardnormalverteilung repräsentieren. λ wird auch als Kehrwert von Mill's-Ratio bezeichnet und ist im Prinzip eine Hazardrate. Dieser Erwartungswert ist offensichtlich größer als ohne Zensierung. Man erkennt hier sofort, wieso OLS inkonsistent ist: der zweite Term des bedingten Erwartungswertes bleibt unberücksichtigt, wird dem Fehlerterm zugeschlagen. Da er aber eine Funktion von \mathbf{x}_i ist, erzeugt dies eine Korrelation zwischen den Kovariaten und dem Fehlerterm, weshalb A5 verletzt wird. Im Prinzip handelt es sich hier um ein Fehlspezifikationsproblem, denn die „Variable“ λ_i wird nicht berücksichtigt.

Dies legt eine Möglichkeit zur konsistenten *Schätzung* von $\boldsymbol{\beta}$ nahe: berechne die Variable λ_i und nimm sie in das Modell auf. Dieses Verfahren werden wir im nächsten Abschnitt im verwandten Fall der Stichprobenauswahl kennen lernen. Im hier zu besprechenden Zensierungs-Fall wird aber üblicherweise das ML-Verfahren eingesetzt. Der Likelihood-Beitrag einer nicht-zensierten Beobachtung ($y_i > 0$) ist wie gewöhnlich die Wahrscheinlichkeitsdichte für den beobachteten Wert. Wegen der hier getroffenen Normalverteilungsannahme ist dies

$$P(Y = y_i) = \frac{1}{\sigma} \phi \left(\frac{y_i - \boldsymbol{\beta}' \mathbf{x}_i}{\sigma} \right).$$

Der Likelihood-Beitrag einer zensierten Beobachtung ($y_i = 0$) dagegen ist die Wahrscheinlichkeit, daß $y_i^* \leq 0$, also

$$P(Y = 0) = P(y_i^* \leq 0) = \Phi \left(\frac{0 - \boldsymbol{\beta}' \mathbf{x}_i}{\sigma} \right).$$

Damit lautet die Log-Likelihood des Tobit-Modells:

$$\ln L = \sum_{y_i > 0} \ln \left[\frac{1}{\sigma} \phi \left(\frac{y_i - \boldsymbol{\beta}' \mathbf{x}_i}{\sigma} \right) \right] + \sum_{y_i = 0} \ln \left[\Phi \left(-\frac{\boldsymbol{\beta}' \mathbf{x}_i}{\sigma} \right) \right].$$

Die Schätzer erhält man mittels der üblichen iterativen Algorithmen zur Maximierung solcher nicht-linearer Funktionen.

Bei der *Interpretation der Schätzer* $\hat{\beta}_j$ muß man beachten, daß

$$\frac{\partial E(y^*|\mathbf{x})}{\partial x_j} = \beta_j,$$

was bedeutet, daß die Koeffizienten des Tobit-Modells Marginal- bzw. Einheits-
effekte auf die latente Variable wiedergeben. Will man Effekte für die zensierte
Population erhalten (etwa um Prognosen zu tätigen), so muß man die Formel für
den zensierten Erwartungswert ableiten (s. Greene, 1993: 694f) und erhält

$$\frac{\partial E(y|\mathbf{x})}{\partial x_j} = \beta_j \Phi\left(\frac{\beta' \mathbf{x}}{\sigma}\right).$$

Um den Marginaleffekt auf die beobachtete, zensierte abhängige Variable zu er-
halten, müssen die Koeffizienten β_j noch mit dem Skalierungsfaktor $\Phi(\beta' \mathbf{x}/\sigma)$
multipliziert werden.

2.4.2 Stichprobenauswahl: Das Heckit-Verfahren

Stichprobenauswahl ist im Prinzip eine indirekte Stützung bzw. Zensierung einer
Zufallsvariable Y durch eine zweite, mit Y korrelierte Zufallsvariable Z . Nehmen
wir an, daß Y nur dann beobachtet wird, wenn Z eine bestimmte Schwelle a
überschreitet. Sind Y und Z bivariat normalverteilt mit Korrelation ρ (y, z) \sim
 $N(\mu_y, \mu_z, \sigma_y^2, \sigma_z^2, \rho)$, so gilt für den indirekt gestützten Erwartungswert von Y

$$E(y|z > a) = \mu_y + \rho\sigma_y\lambda\left(\frac{a - \mu_z}{\sigma_z}\right).$$

Dies ist sehr ähnlich zu obigem Erwartungswert für eine direkt gestützte Zufalls-
variable. Allerdings muß hier die Korrelation berücksichtigt werden. Sind die bei-
den Zufallsvariablen nicht korreliert, so verändert sich der Erwartungswert nicht.
Die Stichprobenauswahl ist zufälliger Natur und stellt kein Problem bezüglich
der Schätzung der Parameter dar. Bei positiver Korrelation der beiden Zufallsva-
riablen gelangen aber eher die größeren Y -Werte in die Stichprobe, weshalb der
Erwartungswert nach oben verzerrt ist. Bei negativer Korrelation ist er dagegen
nach unten verzerrt.

Um Stichprobenauswahl zu modellieren, spezifiziert man eine *Auswahlgleichung*,
die festlegt, welche Beobachtungen in die Stichprobe gelangen und das eigentlich
interessierende Regressionsmodell für die ausgewählten Beobachtungen. Üblicher-
weise geht man von einer Probit-Auswahlgleichung aus (das Probit-Modell wird
genauer im nächsten Kapitel besprochen). Eine binäre Zufallsvariable Z indiziert,
ob eine Beobachtung in der Stichprobenauswahl ist ($z_i = 1$) oder nicht ($z_i = 0$).

Für den Auswahlprozeß spezifizieren wir folgendes Modell:

$$\begin{aligned} z_i^* &= \boldsymbol{\gamma}'\mathbf{w}_i + u_i, \\ z_i &= 1, \quad \text{wenn } z_i^* > 0, \\ z_i &= 0, \quad \text{wenn } z_i^* \leq 0. \end{aligned}$$

Für die zugrundeliegende latente Variable Z^* nehmen wir wieder ein klassisches Regressionsmodell an, womit diese Auswahlgleichung einem Probit-Modell entspricht. Man beachte, daß wir hier annehmen, daß auch für die nicht-ausgewählten Beobachtungen Information vorliegt: nämlich über \mathbf{w}_i (es handelt sich somit um eine indirekte Zensierung). In unserem obigen Beispiel hat man etwa auch die nicht erwerbstätigen Personen befragt. Oder bei Ausfällen in einem Survey hat man vielleicht Informationen aus der Stichprobenziehungliste. Hat man keinerlei Informationen über die nicht-ausgewählten Beobachtungen (indirekte Stützung), so kann man das hier vorgestellte Modell nicht anwenden.

Das eigentlich interessierende *Regressionsmodell*

$$y_i = \boldsymbol{\beta}'\mathbf{x}_i + \epsilon_i$$

hat man nur für $z_i = 1$ beobachtet. Nimmt man nun an, daß die Fehlerterme der beiden Gleichungen bivariat normalverteilt sind mit $(u_i, \epsilon_i) \sim N(0, 0, 1, \sigma_\epsilon^2, \rho)$,²⁶ so gilt für den bedingten Erwartungswert von Y (unter Verwendung des obigen Theorems):

$$\begin{aligned} E(y_i|z_i = 1, \mathbf{x}_i) &= \boldsymbol{\beta}'\mathbf{x}_i + E(\epsilon_i|u_i > -\boldsymbol{\gamma}'\mathbf{w}_i) \\ &= \boldsymbol{\beta}'\mathbf{x}_i + 0 + \rho\sigma_\epsilon\lambda(-\boldsymbol{\gamma}'\mathbf{w}_i) \\ &= \boldsymbol{\beta}'\mathbf{x}_i + \beta_\lambda\lambda(-\boldsymbol{\gamma}'\mathbf{w}_i). \end{aligned}$$

λ ist wie oben als Kehrwert von Mill's-Ratio definiert und ist die Nicht-Auswahl Rate einer Beobachtung mit Kovariatenvektor \mathbf{w}_i . λ ist also eine monoton fallende Funktion der Auswahlwahrscheinlichkeit. Wiederum erkennt man, daß OLS inkonsistent ist, weil die Variable λ_i im Fehlerterm enthalten ist. Einen konsistenten Schätzer kann man auch hier mit Maximum-Likelihood erhalten. Für dieses Modell hat sich jedoch ein zweistufiges Schätzverfahren durchgesetzt, das von Heckman (1979) vorgeschlagen wurde. Dieses sogenannte *Heckit-Verfahren* hat gegenüber Maximum-Likelihood den Vorteil, daß es auf schwächeren Annahmen beruht. In der ersten Stufe schätzt man mit allen Beobachtungen die Auswahlgleichung und berechnet anschließend für jede Beobachtung $\hat{\lambda}_i = \phi(\hat{\boldsymbol{\gamma}}'\mathbf{w}_i)/\Phi(\hat{\boldsymbol{\gamma}}'\mathbf{w}_i)$. In der zweiten Stufe schätzt man für die ausgewählten Beobachtungen die Regression mit der zusätzlichen Variable $\hat{\lambda}_i$ mittels OLS. Dieses Verfahren liefert konsistente Schätzer $\hat{\boldsymbol{\beta}}$. Allerdings ist es ineffizient, weil der Fehlerterm der Regression

²⁶ σ_u^2 ist nicht identifiziert und wird deshalb auf eins normiert.

der zweiten Stufe heteroskedastisch ist. Weiterhin sind die OLS-Standardfehler nicht konsistent, weil OLS nicht berücksichtigt, daß eine Variable (nämlich $\hat{\lambda}_i$) geschätzt ist (einem ähnlichen Problem sind wir bereits bei 2SLS begegnet). Damit sind die t-Tests der Koeffizientenschätzer nicht korrekt. Man kann aber mittels einer Korrektur konsistente Standardfehler berechnen, was aber nur in Spezialprogrammen (wie LIMDEP) implementiert ist.

Die *Interpretation der Koeffizientenschätzer* ist nicht einfach. Ist eine Variable X_j nicht in \mathbf{W} enthalten, so gilt

$$\frac{\partial E(y|\mathbf{x})}{\partial x_j} = \frac{\partial E(y|z = 1, \mathbf{x})}{\partial x_j} = \beta_j.$$

β_j gibt mithin sowohl den Effekt der Variablen X_j in der nicht-ausgewählten als auch den Effekt in der ausgewählten Stichprobe wider. Ist X_j dagegen in \mathbf{W} enthalten, so repräsentiert β_j nurmehr den Marginaleffekt in der nicht-ausgewählten Stichprobe. Der Marginaleffekt in der ausgewählten Stichprobe ist eine Kombination des Auswahleffektes und des Regressionseffektes. Will man also Prognosen für die ausgewählte Population machen, so muß man ähnlich wie im Tobit-Modell anhand obiger Formel für den bedingten Erwartungswert erst den entsprechenden Marginaleffekt errechnen. Schließlich liefert der Schätzer von β_λ eine nützliche Zusatzinformation: sein Vorzeichen ist identisch mit dem Vorzeichen von ρ . Mithin kann man daraus erschließen, ob die Fehlerterme der Auswahl- und der Regressionsgleichung positiv oder negativ korreliert sind.

Die *Präzision der Heckit-Schätzer* wird beeinflusst von der Varianz von λ und von der Kollinearität zwischen λ und \mathbf{x} (Winship und Mare, 1992). Die Varianz von λ hängt direkt von der Güte der Auswahlgleichung ab: Je besser die Prädiktoren die Auswahl vorhersagen können, desto größer wird die Varianz von λ sein und desto präziser werden die Schätzer der zweiten Stufe sein. Die Kollinearität zwischen λ und \mathbf{x} wird besonders hoch sein, wenn $\mathbf{w}_i = \mathbf{x}_i$. Das Modell ist in diesem Fall zwar dennoch schätzbar, weil λ eine hoch nicht-lineare Funktion von \mathbf{w}_i ist, aber die Schätzer sind recht unpräzise. Praktische Erfahrungen mit dem Heckit-Verfahren zeigen, daß \mathbf{w}_i mindestens eine Variable enthalten sollte, die nicht in \mathbf{x}_i enthalten ist.

Desweiteren ist die Konsistenz des Heckit-Verfahrens natürlich auch von der *Gültigkeit seiner Annahmen* abhängig. Insbesondere die Normalverteilungsannahme ist hier kritisch. Von ihr hängt nämlich die Probit-Form der Auswahlgleichung und die funktionale Form von λ ab. Gilt sie nicht, so sind sowohl die Auswahlgleichung als auch der Korrekturfaktor auf der zweiten Stufe fehlspezifiziert. Folge ist, daß die Heckit-Schätzer ebenfalls verzerrt sind. Für den Fall, daß die Normalverteilungsannahme nicht gilt, wurden semi-parametrische Korrekturverfahren vorgeschlagen (s. Winship und Mare, 1992). Doch selbst wenn die Normalverteilungsannahme gilt, produziert die Heckit-Korrektur in kleinen Stichproben häufig

eine Verzerrung (Stolzenberg und Relles, 1990). Vor dem routinemäßigen Einsatz des Heckit-Verfahrens muß somit gewarnt werden: „In short, there is considerable evidence that the method can easily do more harm than good, and that its careless or mechanical application runs much danger of producing vivid examples of the problems that Lieberson so aptly discussed when he cautioned against statistical ‚corrections‘ that in fact produce substantial distortions far worse than the problems they are designed to cure“ (Stolzenberg und Relles, 1990: 408f).

2.4.3 Exkurs: Fitmaße für Maximum-Likelihood Modelle

Das klassische Fitmaß für das lineare Regressionsmodell ist das Bestimmtheitsmaß R^2 . Dort waren wir in der günstigen Position, daß das Verfahren zur Schätzung der Parameter (OLS eben) gleichzeitig den Fit maximiert (den Abstand von beobachtetem und vom Modell vorhergesagtem Wert minimiert). Somit ergibt sich das Bestimmtheitsmaß unmittelbar aus der Schätzprozedur. Dem ist nicht mehr so bei nicht-linearen Regressionsmodellen, die mit der Maximum-Likelihood Methode geschätzt werden. Hier werden die Schätzer so bestimmt, daß die Likelihood der Stichprobe maximiert wird. Dies muß aber nicht notwendigerweise den Fit maximieren. Weiterhin ist es bei den meisten nicht-linearen Regressionsmodellen gar nicht sinnvoll, von einer erklärten Varianz zu sprechen, weil die abhängige Variable nicht metrisch oder weil sie beschränkt ist. Dies würde nur für die dem Modell zugrundeliegende latente Variable Sinn machen, welche man aber nicht kennt (sonst könnte man ja OLS einsetzen). Die Berechnung des Bestimmtheitsmaßes, wie wir es von der linearen Regression kennen, ist somit bei nicht-linearen Maximum-Likelihood Modellen nicht möglich. Dennoch wurden für solche Modelle eine Reihe von Pseudo- R^2 Maßen vorgeschlagen, von denen einige in diesem Exkurs vorgestellt werden sollen.

All diese Maße beruhen auf dem Vergleich der maximierten Likelihoods zweier Modelle: Die Likelihood des restringierten Modells (L_R , meist das Modell nur mit Konstante) wird mit der Likelihood des unrestringierten Modells verglichen (L_U , das interessierende Modell). Man berechnet die sogenannte *Likelihood-Ratio Testgröße* $LR = 2(\ln L_U - \ln L_R)$. Diese Größe ist (asymptotisch) χ^2 -verteilt mit der Zahl der Restriktionen als Freiheitsgrade. Ist LR größer als der kritische Wert, so können die Restriktionen zurückgewiesen werden. Am häufigsten wird dieser Test in Analogie zum F-Test eingesetzt: Man testet die Nullhypothese, daß alle Regressionskoeffizienten (bis auf die Konstante) gleich null sind.

Der Likelihood-Ratio Test kann aber auch zum Test der Signifikanz von Interaktionseffekten eingesetzt werden. Dabei hat man wie im linearen Regressionsmodell zwei Möglichkeiten. Im ersten Fall schätzt man ein unrestringiertes Modell mit allen Interaktionen und ein restringiertes Modell ohne die Interaktionen. Der LR-Test gibt dann Aufschluß über die Signifikanz der Interaktionseffekte. Die-

ses Vorgehen beinhaltet aber die Annahme, daß die übrigen Modellparameter in den verglichenen Gruppen identisch sind. Will man diese Annahme vermeiden, so kann man auch getrennte Modelle für die zu vergleichenden Gruppen schätzen. Die unrestringierte Likelihood ist dann die Summe der Gruppen-Likelihoods. Die restringierte Likelihood ist wie im ersten Fall die gemeinsame Regression ohne Interaktionen. Man beachte, daß bei diesem Vorgehen die Zahl der Restriktionen größer ist, weil die zusätzlichen Modellparameter im unrestringierten Modell frei sind.

Der Likelihood-Ratio Test ist allerdings kein *Fitmaß*. Aber auf seiner Basis haben verschiedene Autoren in Analogie zur linearen Regression sogenannte Pseudo- R^2 Maße vorgeschlagen. Dies sind Maße dafür, wieviel von der restringierten Likelihood (das Modell mit nur einer Konstanten) durch das unrestringierte Modell „erklärt“ werden kann. Sie sind null, wenn die zusätzlichen Koeffizienten des Modells die Likelihood nicht verbessern. Anders als im linearen Regressionsmodell ist ihre Obergrenze allerdings meist nicht eins, sondern liegt darunter. Einige dieser Maße sollen nun vorgestellt werden (weitere Maße werden bei Magee (1990) und Veall und Zimmermann (1994) diskutiert). Das *McFadden Pseudo- R^2* ist definiert als:

$$R_{MF}^2 = \frac{\ln L_R - \ln L_U}{\ln L_R}.$$

Es erfaßt die relative Log-Likelihood Verbesserung des unrestringierten Modells gegenüber dem Modell mit nur einer Konstanten. Ein weiteres Maß wurde von *Maddala* vorgeschlagen:

$$R_M^2 = 1 - e^{-LR/n}.$$

Es berücksichtigt zusätzlich die Fallzahl n . Schließlich lautet das Maß von *Aldrich und Nelson*:

$$R_{AN}^2 = \frac{LR}{n + LR}.$$

R_M^2 und R_{AN}^2 werden zwar öfter in der Literatur berichtet, sollten aber besser nicht verwendet werden, weil sie eine Obergrenze kleiner eins haben (s. Veall und Zimmermann, 1994). Erfahrungsgemäß fällt R_{MF}^2 am kleinsten aus, es ist am konservativsten. Unter anderem aufgrund dieser Eigenschaft wird es in vielen Lehrbüchern (z.B. Judge et al., 1985: 767) als *das* Fitmaß für nicht-lineare Regressionsmodelle empfohlen.²⁷

All diese Maße sind nur für den Vergleich von verschachtelten („nested“) Modellen geeignet. Will man Modelle vergleichen, die auf unterschiedlichen Wahrscheinlichkeitsverteilungen basieren, benötigt man andere Maße. Das Hauptproblem beim Vergleich nicht-verschachtelter Modelle ist die eventuell unterschiedliche Zahl von

²⁷Veall und Zimmermann (1994) empfehlen aufgrund ihrer Simulationsstudien allerdings ein anderes Maß: das McKelvey/Zavoina Pseudo- R^2 . Dieses Maß liegt laut ihren Ergebnissen am nächsten am Bestimmtheitsmaß der latenten Variable.

Parametern. Das Modell mit der größeren Zahl von Parametern wird im Normalfall die höhere Likelihood aufweisen. Es stellt sich dann die Frage, wie man die Zahl der Parameter gewichtet. Ein Vorschlag zur Behebung dieses Problems stammt von Raftery (1986). Er schlägt das *Bayesian Information Criterion* (BIC) zum Modellvergleich vor. Für jedes Modell berechnet man

$$\text{BIC} = -LR + \ln(n)r,$$

wobei LR die Likelihood-Ratio in Bezug auf ein Basismodell, n die Fallzahl und r die Zahl der zusätzlichen Parameter im Vergleich zu dem Basismodell ist. Jeder zusätzliche Parameter wird bei diesem Maß folglich mit $\ln(n)$ „bestraft“. Das Modell mit dem kleinsten BIC ist das „beste“ Modell. Eine tiefere Begründung dieses Maßes in der bayesianischen Teststatistik findet man bei Raftery (1995), der auch demonstriert, wie dieses Maß viele Probleme der herkömmlichen Teststatistik überwinden hilft.

2.4.4 Anwendungsbeispiel: Scheidungsraten in 50 Ländern

Wir demonstrieren das Heckit-Verfahren wiederum mit unseren Scheidungsdaten. Unser Datensatz enthält Angaben über 146 Länder, aber für nur 50 Länder verfügen wir über Angaben zur Scheidungsrate und zur Erwerbsquote. Für alle 146 Länder kennen wir aber das durchschnittliche Bruttonettoprodukt pro Kopf in US-Dollars (BSP). Die Länder ohne Angaben zu Scheidungsrate und Erwerbsquote haben ein mittleres BSP von 2078\$, während die Länder mit Angabe ein mittleres BSP von 9642\$ aufweisen. Unsere Stichprobe umfaßt also vor allem reiche Länder mit wohl eher höheren Scheidungsraten. Damit haben wir ein typisches Stichprobenauswahl-Problem und unsere bisher berichteten Schätzer sind eventuell verzerrt.

Zur Behebung dieser Verzerrung setzen wir das zweistufige Heckit-Verfahren ein (geschätzt mit LIMDEP, s. Greene, 1995: Kap. 28). Zuerst schätzen wir ein Auswahl-Probit (N=146), bei dem der Auswahlindikator eins ist, falls ein Land sowohl die Scheidungsrate als auch die Erwerbsquote berichtet hat. Unabhängige Variable ist das BSP. Die Schätzergebnisse bestätigen obige Vermutung: Länder mit höherem BSP berichten signifikant häufiger die endogenen Variablen „Scheidungsrate“ und „Erwerbsquote“. Anhand der Ergebnisse dieses Probit-Modells errechnen wir für jedes Land den Kehrwert von Mill's-Ratio λ . Im zweiten Schritt führen wir eine 2SLS-Schätzung durch, wie sie im vorigen Abschnitt beschrieben ist, nur daß jeweils λ als zusätzliche unabhängige Variable berücksichtigt wird. Dies bedeutet, daß wir mit unserer Schätzung sowohl der Endogenität der Erwerbsquote als auch dem Stichproben-Auswahlproblem Rechnung tragen! Das Ergebnis für die Scheidungsgleichung ist in Spalte (5) von Tabelle (1) angeführt.

Die Effekte der Erwerbsquote und des Modernisierungsindex verschwinden nun fast völlig. Dies ist auch verständlich, denn die reichen Länder weisen sowohl höhere Erwerbsquoten wie auch ein höheres Modernisierungsniveau auf. Deshalb waren in unseren früheren Modellen diese Koeffizienten besonders stark vom Auswahlfehler betroffen. Aber auch der Effekt der Sex-Ratio verändert sich deutlich: er wird noch negativer und man ist schon fast geneigt, diesen Effekt als eine Bestätigung der Guttentag-Secord Hypothese zu werten. Schließlich ist der Koeffizient von λ negativ, d.h. die Fehlerterme der Auswahl- und Scheidungsgleichung korrelieren negativ. Allerdings muß betont werden, daß die Anwendung des Heckit-Verfahrens bei unseren Daten problematisch ist. Zum einen ist die Auswahlgleichung mit nur einem Prädiktor (BSP) schlecht spezifiziert, zum anderen ist die Fallzahl unserer Stichprobe sehr klein. Beides kann - wie oben erwähnt - dazu führen, daß die Heckit-Schätzer verzerrt sind.

Damit schließen wir die Analyse der Scheidungsdaten ab. Dieses Anwendungsbeispiel hat deutlich gemacht, daß es in der Praxis nicht ganz einfach ist, das „beste“ Schätzverfahren zu finden. Die verschiedenen Verfahren zur Korrektur von Annahmeverletzungen von OLS beruhen ebenfalls auf Annahmen, die man eigentlich testen müßte. Wenn die Korrekturverfahren dann auch noch unterschiedliche Ergebnisse liefern, so ist oft nicht klar, ob die korrigierten Schätzer nun wirklich die „besseren“ sind. Deshalb wird manchmal der Vorschlag gemacht, bei OLS zu bleiben und Korrekturverfahren nur dann einzusetzen, wenn bewiesen ist, daß deren Annahmen zutreffen.

2.5 Modelle für Zähldaten

Die meisten demographischen Prozesse weisen nicht-absorbierende Zielzustände auf, d.h. das interessierende Ereignis kann mehrfach auftreten: Menschen können mehrmals heiraten, sie können mehrere Scheidungen haben, können mehrere Kinder bekommen und können öfters den Job wechseln. Die informativsten statistischen Verfahren zur Analyse solcher Prozesse sind sicherlich die Methoden der Ereignisdatenanalyse (s. den Beitrag von Diekmann in diesem Handbuch). Doch diese Methoden sind relativ komplex, insbesondere dann, wenn man Mehrfachepisoden (z.B. die Zeit bis zur ersten, zweiten, dritten, usw. Geburt) adäquat analysieren will. Zudem ist die Datenlage oft unbefriedigend und der exakte Zeitpunkt des Eintretens der Ereignisse unbekannt. In solch einer Situation bietet es sich an, das „Timing“ des Prozesses zu vernachlässigen und nur die Häufigkeit des Auftretens des Ereignisses zu untersuchen. Man spricht in diesem Fall von Zähldaten. Zähldaten können nur Werte aus der Menge der natürlichen Zahlen annehmen, d.h. $Y \in \{0, 1, 2, 3, \dots\}$. In diesem Abschnitt sollen einige Modelle zur Analyse von Zähldaten vorgestellt werden.

Man könnte auf den ersten Blick in Versuchung geraten, solche Zähldaten mittels

Verfahren wie OLS, Tobit oder ordinalem Probit (s. Abschnitt 3.2.3) zu analysieren. Diese Verfahren sind allerdings nicht speziell für Zähldaten entwickelt worden und deshalb werden bei deren Anwendung auf Zähldaten wichtige Annahmen verletzt. Bei OLS etwa wird die diskrete Natur und die Nicht-Negativität von Zähldaten nicht berücksichtigt. Deshalb empfiehlt es sich, die für solche Daten entwickelten Zähldatenmodelle einzusetzen, zumal dieselben auch nicht komplizierter sind als die (allerdings bekannteren) obigen Alternativen. Zudem erlauben einige Zähldatenmodelle die einfache Modellierung spezifischer Aspekte von Zähldaten. Aus diesen Gründen werden seit einigen Jahren vermehrt solche Modelle eingesetzt. Einige Anwendungen seien hier beispielhaft aufgelistet: Zahl der pro Jahr neugegründeten Betriebe einer Organisationspopulation (Hannan und Carroll, 1992), Zahl der Arbeitslosigkeitsepisoden von Arbeitnehmern (Andrefß, 1989) und Zahl der Kinder und der Scheidungen deutscher Frauen (Winkelmann und Zimmermann, 1994). Die zunehmende Bekanntheit der Zähldatenmodelle schlägt sich auch in der Zahl guter Darstellungen nieder (z.B. Ronning, 1991: Kap. 4.2; Barron, 1992; Winkelmann, 1994; Land et al., 1996). Im folgenden sollen die beiden am häufigsten verwendeten Zähldatenmodelle vorgestellt werden: das Poisson-Modell und das Negative-Binomial-Modell. Zusätzlich soll das Hürden-Poisson-Modell besprochen werden, weil es interessante demographische Anwendungsmöglichkeiten besitzt. Mit LIMDEP (Greene, 1995: Kap. 26) können viele weitere Varianten dieser Modelle geschätzt werden.

2.5.1 Drei Zähldatenmodelle

Im einfachsten Fall handelt es sich bei dem die Zähldaten generierenden stochastischen Prozeß um einen Poisson-Prozeß: die Rate λ , mit der ein Ereignis eintritt, ist konstant und unabhängig von der Vorgeschichte des Prozesses. Betrachtet man das Timing der Ereignisse, so verwendet man zur statistischen Analyse das Exponential-Ratenmodell. Komplementär hierzu ist jedoch die Analyse der Zahl der Ereignisse Y in einem Zeitintervall der Länge t . Die Dichte von Y ist dann durch die bekannte *Poisson-Verteilung* beschrieben:

$$P(Y = y) = \frac{e^{-\lambda t} (\lambda t)^y}{y!}, \quad y = 0, 1, 2, \dots$$

Die ersten beiden Momente sind $E(y) = V(y) = \lambda t$. Die „mittlere“ Zahl von Ereignissen ergibt sich somit aus dem Produkt von Rate pro Zeiteinheit und Zahl der Zeiteinheiten. Zugleich ist dieses Produkt gleich der Varianz! Man spricht deshalb davon, daß die Poisson-Verteilung Äqui-Dispersion ($E(y)/V(y) = 1$) impliziert. Zu einem Regressionsmodell gelangt man nun, indem man den Erwartungswert von Kovariaten abhängig macht, was meist in der Form $E(y_i | \mathbf{x}_i) = \exp(\boldsymbol{\beta}' \mathbf{x}_i + \ln t_i)$ geschieht. Die log-lineare Form stellt sicher, daß der Erwar-

tungswert (und die Varianz) positiv ist. Der „Offset“ $\ln t_i$ berücksichtigt, daß die erwartete Zahl der Ereignisse mit Zunahme der Risikozeit ansteigt.²⁸

Da in diesem Modell die Kovariateneffekte auf den Erwartungswert exponentiell spezifiziert sind, kann man ohne weitere Berechnungen nur die Vorzeichen der Koeffizienten interpretieren (vgl. Abschnitt 2.2.1). Ein positives β_j bedeutet, daß X_j die erwartete Zahl der Ereignisse (und die Rate) erhöht. Die *Marginal-effekte* erhält man aus

$$\frac{\partial E(y|\mathbf{x})}{\partial x_j} = \lambda \beta_j.$$

Der Marginal-effekt ist von X abhängig. Analog können die *Einheitseffekte* errechnet werden. Besonders anschaulich ist deren Interpretation als Prozenteffekt: $(e^{\beta_j} - 1) \cdot 100$ gibt an, um wieviel Prozent sich die erwartete Zahl der Ereignisse verändert, wenn X_j um eine Einheit steigt.

Der einfacheren Notation wegen nehmen wir im folgenden an, daß alle Individuen gleich lang dem Risiko ausgesetzt waren und setzen deshalb $t_i = 1$. Die *Likelihood* bei n unabhängigen Beobachtungen y_i lautet somit:

$$L = \prod_{i=1}^n \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!},$$

wobei $\lambda_i = \exp(\boldsymbol{\beta}' \mathbf{x}_i)$. Damit lautet die Log-Likelihood des Poisson-Regressionsmodells:

$$\ln L = \sum_{i=1}^n \{y_i \boldsymbol{\beta}' \mathbf{x}_i - \exp(\boldsymbol{\beta}' \mathbf{x}_i) - \ln(y_i!)\}.$$

Die ML-Parameterschätzungen erhält man mittels der Standard-Algorithmen.

Die Annahme der *Äqui-Dispersion* ist oft verletzt. Schätzt man in verschiedenen Subgruppen den Erwartungswert und die Varianz der Zählvariable, so sollte bei Gültigkeit der Äqui-Dispersions Annahme in einem Erwartungswert-Varianz Streudiagramm eine 45°-Linie resultieren. Meist ist die Varianz größer als der Erwartungswert ($V(y_i) > E(y_i)$, Über-Dispersion), aber insbesondere bei demographischen Daten zeigt sich auch Unter-Dispersion (s. Winkelmann und Zimmermann, 1994). Über-Dispersion kann etwa dadurch entstehen, daß die Rate nicht konstant ist sondern mit der Verweildauer ansteigt, oder dadurch, daß positive Ansteckung vorliegt (d.h. nach dem Eintritt eines Ereignisses ist die Rate für das nächste Ereignis höher, s. Barron, 1992). Unter-Dispersion kann auftreten, wenn die Rate mit der Verweildauer abnimmt, oder negative Ansteckung vorliegt.

²⁸In der Praxis wird allerdings $\ln t_i$ oft einfach als Kovariate eingeführt, ohne daß der zugehörige Parameter auf eins restringiert wird. Dieser Parameter, nennen wir ihn γ , ist als Elastizität interpretierbar: $\gamma-1$ gibt an, um wieviel Prozent sich die erwartete Zahl der Ereignisse verändert, wenn sich die Risikozeit um ein Prozent erhöht.

Ein weiterer Grund für Über-Dispersion kann unbeobachtete Heterogenität sein. Oben nahmen wir an, daß die beobachteten Kovariaten alle Variation der Rate erfassen. Ist dem nicht so, so liegt es nahe eine (zusätzliche) stochastische Komponente in unseren Regressionsansatz einzuführen: $\lambda_i = \exp(\boldsymbol{\beta}'\mathbf{x}_i)\epsilon_i$. Um mit diesem Ansatz zu einer handhabbaren Lösung für die Wahrscheinlichkeitsdichte von Y_i zu gelangen, wird unterstellt, daß ϵ_i gamma-verteilt ist mit den Parametern ϕ_i und ν_i . Damit gelingt es ϵ_i „herauszuintegrieren“ und man gelangt zum *Negativen-Binomial-Modell*:²⁹

$$P(Y_i = y_i) = \frac{\Gamma(y_i + \nu_i)}{\Gamma(\nu_i)y_i!} \left(\frac{\nu_i}{\nu_i + \phi_i} \right)^{\nu_i} \left(\frac{\phi_i}{\nu_i + \phi_i} \right)^{y_i}.$$

$\Gamma(\cdot)$ steht hier für die Gamma-Funktion. Es gilt nun:

$$E(y_i) = \phi_i \quad \text{und} \quad V(y_i) = \phi_i + \frac{1}{\nu_i}\phi_i^2.$$

Sinnvollerweise wird wieder die Parametrisierung $\phi_i = \exp(\boldsymbol{\beta}'\mathbf{x}_i)$ gewählt. Setzen wir $1/\nu_i = \gamma$, so folgt für das Verhältnis von Erwartungswert und Varianz:

$$\frac{V(y_i)}{E(y_i)} = 1 + \gamma E(y_i).$$

Für $\gamma > 0$ liegt also Über-Dispersion vor, bei $\gamma = 0$ Äqui-Dispersion (Unter-Dispersion kann dieses Modell nicht modellieren, da $\gamma < 0$ nicht zulässig ist). Die Schätzer für β und γ sind wiederum mittels ML erhältlich. Ein Test auf $\gamma = 0$ zeigt, ob Über-Dispersion vorliegt.

Im Falle von Unter-Dispersion muß man andere Modelle verwenden. Ein Modell, das sowohl bei Über- wie auch Unter-Dispersion eingesetzt werden kann, ist das sogenannte *Hürden-Poisson-Modell* (s. Winkelmann, 1994: 120ff). Es basiert auf der Grundidee, daß die Zählraten Resultat eines zweistufigen Entscheidungsprozesses sind. Beim Fertilitätsprozeß etwa ist es plausibel anzunehmen, daß sich Individuen erst entscheiden, ob sie Kinder haben wollen oder nicht. Dann müssen die, die auf der ersten Stufe positiv entschieden haben, die Zahl der Kinder festlegen. Um diesen Prozeß zu modellieren, benötigt man zwei Wahrscheinlichkeitsverteilungen f_1 und f_2 für positive, ganze Zahlen. f_1 beschreibt den Hürden-Prozeß und f_2 regiert den Prozeß für die, die die Hürde übersprungen haben. Damit lautet die Wahrscheinlichkeitsverteilung des Hürden-Modells:

$$\begin{aligned} P(Y_i = 0) &= f_1(0) \\ P(Y_i = y_i) &= (1 - f_1(0)) \frac{f_2(y_i)}{1 - f_2(0)}, \quad y_i = 1, 2, 3, \dots \end{aligned}$$

²⁹Das Negative-Binomial-Modell ist nicht nur über die Einführung unbeobachteter Heterogenität aus dem Poisson-Modell ableitbar. Auch positive Ansteckung kann zu diesem Modell führen. Somit folgt umgekehrt aus einem empirisch festgestellten besseren Fit dieses Modells gegenüber dem Poisson-Modell nicht notwendigerweise, daß unbeobachtete Heterogenität der Grund ist.

$1 - f_1(0)$ ist die Wahrscheinlichkeit, die Hürde zu überspringen, und der Bruch stellt eine Normalisierung von f_2 dar (damit sich die Wahrscheinlichkeiten wieder zu eins summieren).³⁰ Beim Hürden-Poisson-Modell unterstellt man, daß sowohl f_1 als auch f_2 Poisson-Verteilungen sind mit $\lambda_{1i} = \exp(\beta_1' \mathbf{x}_i)$ und $\lambda_{2i} = \exp(\beta_2' \mathbf{x}_i)$. Die β_2 sind interpretierbar wie im einfachen Poisson-Modell, während die β_1 die Wahrscheinlichkeit des Überspringens der Hürde verändern. Ist $\beta_1 = \beta_2$, so liegt wieder das einfache Poisson-Modell vor (zum Test kann ein Likelihood-Ratio-Test verwendet werden). Die Schätzung der Parameter erfolgt wiederum mit ML, wobei die Log-Likelihood lautet (die Terme $y_i!$ sind weggelassen):

$$\ln L = \sum_{y_i=0} \{ -\lambda_{1i} \} + \sum_{y_i>0} \{ \ln(1 - e^{-\lambda_{1i}}) + y_i \ln \lambda_{2i} - \ln(e^{\lambda_{2i}} - 1) \}.$$

Fitmaße für diese Modelle sind relativ einfach konstruierbar. Da es sich um ML-Schätzungen handelt, bietet sich die Berechnung des McFadden Pseudo- R^2 Maßes an. Darüberhinaus gibt es bei Zähldaten aber noch weitere Möglichkeiten (s. Winkelmann, 1994: 102). Man kennt ja die Zahl der Beobachtungen in jeder Kategorie j der Zählvariable (n_j). Für jede Beobachtung kann man außerdem mittels der geschätzten Parameter die $\hat{P}_i(Y_i = y_j)$ ermitteln. Summieren über alle i liefert die vom Modell vorhergesagten \hat{n}_j . Nun kann man die beobachteten und die vorhergesagten n_j vergleichen. Hierzu bietet sich etwa die bekannte Pearson χ^2 -Statistik an: $\chi^2 = \sum_{j=1}^J (n_j - \hat{n}_j)^2 / \hat{n}_j$. Dabei muß unter praktischen Gesichtspunkten die Summationsobergrenze J festgelegt werden (Faustregel: Wähle J so, daß $\sum_{j>J} \hat{n}_j < 5$).

2.5.2 Anwendungsbeispiel: Die Kinderzahl deutscher Frauen

Als Anwendungsbeispiel soll im folgenden die Kinderzahl deutscher Frauen untersucht werden. Die Daten stammen aus der ALLBUS-Basisumfrage von 1991, die sowohl West- als auch Ost-Deutsche erfaßt (Zentralarchiv, 1991). In dieser Studie wurde unter anderem nach der Zahl der Kinder gefragt. Bei der Kinderzahl handelt es sich um eine typische Zählvariable (ähnliche Analysen mit Daten des sozio-ökonomischen Panels findet man bei Winkelmann und Zimmermann, 1994). Wir beschränken die Analysen auf Frauen, die 1940 oder später geboren wurden, damit die Risikozeit in den Zeitraum der Existenz der BRD und der DDR fallen. Das Hauptproblem bei einer Analyse der Kinderzahl ist die Erfassung der Risikozeit. Eine Möglichkeit wäre es, nur Frauen zu betrachten, die älter als 45 sind. Das

³⁰Bei dieser Modellierung rechnet man alle Nullen dem ersten Prozeß zu, d.h. „null Kinder“ wird als Resultat der Entscheidung für Kinderlosigkeit betrachtet. Alternativ könnte man aber auch annehmen, daß einige Nullen bedeuten „habe noch keine Kinder“. Diese Modellierung führt zum sogenannten „zero inflated poisson model“ (Greene, 1995: 572ff).

würde in unserem Fall zu äußerst geringen Fallzahlen führen. Deshalb wird im folgenden die unterschiedliche Länge der Risikozeit dadurch berücksichtigt, daß wir die (logarithmierten) Jahre vom Zeitpunkt der Erstheirat bis zum Interview als Kovariate ins Modell aufnehmen (bei ledigen Frauen setzen wir diese Kovariate auf -5). Ungenauigkeiten ergeben sich bei diesem Vorgehen aus der Nicht-Berücksichtigung der Zeiten vor der ersten Heirat und nach einer eventuellen Scheidung. Zusätzlich nehmen wir in die Modelle die Religion, die Wohnortgröße, die Bildungsjahre der Befragten und die Bildungsjahre ihres gegenwärtigen Ehepartners (falls vorhanden) auf. Eine weitere Dummy gibt an, ob die Befragte ihre Jugend in West- oder Ost-Deutschland verbrachte. Inhaltlich interessiert uns insbesondere, ob Frauen in der DDR mehr Kinder hatten und welcher Art die Bildungseffekte sind. Meist wird vermutet, daß höher gebildete Frauen weniger Kinder haben. Das Hürden-Poisson-Modell gibt uns die Möglichkeit festzustellen, ob dies auf häufigerer Kinderlosigkeit höher gebildeter Frauen beruht, oder ob sie weniger Kinder haben (wenn sie mindestens eins haben).

Tabelle (2) enthält die Ergebnisse (alle Modelle wurden mit dem GAUSS Anwendermodul für Zähldaten geschätzt; s. King, 1992). Die mittlere Kinderzahl beträgt in unseren Daten 1,46 bei einer Varianz von 1,17. Dies deutet bereits auf Unter-Dispersion hin, was dann auch zur Folge hat, daß das Negative-Binomial-Modell nicht geschätzt werden kann ($\hat{\gamma}$ konvergiert nicht). Zum Vergleich sind in der ersten Spalte die Ergebnisse von OLS angeführt, die zweite Spalte gibt die Poisson-Schätzer wieder. Es zeigt sich, daß beide Modelle meist zu qualitativ recht ähnlichen Ergebnissen führen. Allerdings zeigen die unterschiedlichen Ergebnisse bei der Geburtskohorte und bei der Bildung des Ehemanns, daß OLS-Schätzungen unbedingt durch Zähldatenmodelle ergänzt werden sollten. Durchgängig weist die Risikozeit (Jahre seit Erstheirat) den erwarteten Effekt auf, auch wenn der Koeffizient nicht gleich eins ist, wie es bei korrekter Spezifikation des Modells und exakter Erfassung der Risikozeit zu erwarten wäre. Inhaltlich zeigt sich wie erwartet, daß westdeutsche Frauen weniger Kinder haben (-11%). Gleichfalls haben höher gebildete Frauen weniger Kinder (pro Bildungsjahr -1,4%). Interessant ist aber, daß mit steigender Bildung des Ehemanns die Kinderzahl ansteigt (+1%). Bildung der Frau und des Mannes zeigen entgegengesetzte Effekte, was eine Vorhersage der familienökonomischen Theorie bestätigt. Die anderen Kovariaten zeigen keine signifikanten Effekte. Die Spalten drei und vier schließlich enthalten die Ergebnisse des Hürden-Poisson-Modells (wobei die dritte Spalte sich auf den Hürden-Teil bezieht). Hier ergeben sich interessante Differenzierungen der bisherigen Ergebnisse. Frauen aus der BRD haben nicht deshalb weniger Kinder, weil sie kleinere Familien haben, sondern weil eine höhere Zahl von ihnen kinderlos bleibt. Die Wahrscheinlichkeit des Überspringens der Hürde ist für sie kleiner. Haben sie aber Kinder, so ist die Zahl der Kinder nicht niedriger als bei Frauen mit Kindern in der DDR. Demgegenüber wirkt die Bildung vorwiegend auf die Kinderzahl. Höher gebildete Frauen haben kleinere Familien. Die Fit-Maße zei-

Tabelle 2: Zähldatenmodelle für die Kinderzahl deutscher Frauen (N=946)

Variable	OLS	Poisson	Hürden-Poisson	
Konstante	1,760	0,052	0,335	0,032
Jahre seit Erstheirat (ln)	0,180* (12,10)	0,250* (10,12)	0,364* (12,82)	0,232* (6,11)
Geburtskohorte (1940=0)	-0,011* (2,43)	0,003 (0,72)	0,010 (1,39)	-0,003 (0,65)
Katholisch (1=Ja)	-0,016 (0,22)	-0,024 (0,49)	0,025 (0,62)	-0,046 (1,16)
Konfessionslos (1=Ja)	-0,106 (1,50)	-0,072 (1,54)	-0,306* (6,78)	-0,051 (1,36)
Stadt (1=Ja)	-0,026 (0,44)	-0,010 (0,28)	-0,122* (2,51)	0,020 (0,47)
Bildung (in Jahren)	-0,019 (1,92)	-0,014* (1,97)	0,026 (1,88)	-0,032* (2,83)
Bildung Ehemann (in Jahren)	0,008 (0,98)	0,010* (2,22)	-0,017 (1,23)	0,019* (2,46)
BRD (1=Ja)	-0,180* (2,38)	-0,115* (2,43)	-0,560* (12,33)	-0,028 (0,81)
(Pseudo)- R^2	0,38	0,23	0,30	
-Log-Likelihood	—	701,5	631,1	

* signifikant auf dem 5%-Niveau. Absolute t-Werte in Klammern (berechnet mit heteroskedastizitätskonsistenten Standardfehlern). McFadden Pseudo- R^2 Werte gegenüber dem Poisson-Modell ohne Kovariate. Bezugsgruppe ist protestantisch, Wohnort mit weniger als 50.000 Einwohner, Frau verbrachte Jugend in der DDR.

Quelle: Eigene Berechnungen mit Daten des ALLBUS 1991.

gen eine zufriedenstellende Anpassung der Modelle. Das Hürden-Poisson-Modell hat einen deutlich besseren Fit als das Poisson-Modell: ein Likelihood-Ratio-Test führt zur Ablehnung der Nullhypothese $\beta_1 = \beta_2$ ($\chi^2=140,8$ bei 9 Freiheitsgraden).

Dieses Anwendungsbeispiel konnte, so hoffe ich, die Nützlichkeit von Zähldatenmodellen für demographische Fragestellungen demonstrieren. Dennoch sollen noch einige Probleme dieser Modelle deutlich gemacht werden. Am schwersten

wiegt, daß demographische Prozesse im Querschnitt betrachtet werden. Längsschnittanalysen des Fertilitätsprozesses haben gezeigt, daß der Prozeß deutlich paritätsabhängig ist (Heckman und Walker, 1990). Zählmodellen beruhen aber auf der Annahme, daß die Teil-Prozesse identisch sind (Ausnahme: das Hürden-Poisson-Modell). Ebenso können zeitabhängige Veränderungen von Kovariaten nicht berücksichtigt werden (Beispiele: Erwerbstätigkeit, Familienstand). Dies führt unter anderem zu Problemen bei der Berechnung der Risikozeit, wenn etwa das Fertilitätsrisiko vor und nach einer Scheidung unterschiedlich ist. Ebenso können deshalb Endogenitäts-Probleme nicht gelöst werden. Bezüglich des Frauen-Bildungseffektes könnte man etwa argumentieren, daß der Fertilitätsprozeß auch das Bildungsniveau beeinflußt, wenn Frauen ihre Bildungslaufbahn abbrechen, weil sie ein Kind bekommen. Trotz all dieser Probleme sind Zählmodellen dann unentbehrlich, wenn nur Querschnittsinformationen über den demographischen Prozeß vorliegen. Hat man Längsschnittinformationen, so sollte man darüberhinaus auch Verfahren verwenden, die die Längsschnittnatur der Daten ausnutzen.

3 Regressionsverfahren für qualitative abhängige Variablen

In diesem Kapitel werden Regressionsverfahren für nominale und ordinale (qualitative) abhängige Variablen vorgestellt. Das klassische Verfahren zur Analyse qualitativer Daten ist die Tabellenanalyse. Doch seit etwa zwanzig Jahren werden immer öfter auch Regressionsverfahren für qualitative Daten eingesetzt. Es ist sinnvoll, diese Verfahren in binäre und multinomiale Modelle einzuteilen. Binäre Modelle eignen sich für dichotome abhängige Variablen mit nur zwei Ausprägungen. Sie werden in Abschnitt 3.1 vorgestellt. Hat die abhängige Variable mehrere Ausprägungen, so muß man multinomiale Modelle einsetzen, die in Abschnitt 3.2 besprochen werden. An Lehrbüchern über Regressionsverfahren für qualitative Daten besteht mittlerweile kein Mangel mehr (z.B. Maddala, 1983; Amemiya, 1985: Kap. 9; Ronning, 1991: Kap. 2; Greene, 1993: Kap. 21). Ausführliche, anwendungsorientierte Einführungen geben Hosmer und Lemeshow (1989) und Menard (1995). Das umfassende Lehrbuch von Agresti (1990) geht auch auf die klassischen und modernen (log-linearen) Verfahren der Tabellenanalyse ein (s.a. Clogg und Shihadeh, 1994). Alle in diesem Kapitel vorzustellenden Modelle (und noch viele weitere) können mit LIMDEP (Greene, 1995: Kap. 21 bis 24) einfach geschätzt werden. Dieses Programm gibt neben den Koeffizientenschätzern auch automatisch die Marginalwirkungen aus.

3.1 Binäre Modelle

Binäre abhängige Variablen treten in der demographischen Forschung häufig auf. Einige willkürlich ausgewählte Beispiele: Untersuchungen über die Verwendung von Verhütungsmitteln (Njogu, 1991), über die Einnahme von Drogen (Flewelling und Baumann, 1990), über vorehelichen Geschlechtsverkehr und Scheidung (Kahn und London, 1991) und über Familienstruktur und Geburt eines Kindes im Teenager-Alter (Powers, 1993). Für die Analyse von binären Daten haben sich in der Forschungsliteratur insbesondere zwei Modelle durchgesetzt: Das Logit- und das Probit-Modell. Im ersten Abschnitt werden diese beiden Modelle vorgestellt. Beide Modelle sind jedoch Einzelgleichungsmodelle. Eine Erweiterung des Probit-Modells auf den bivariaten Fall (zwei Gleichungen) werden wir im zweiten Abschnitt kennenlernen. Ein Anwendungsbeispiel über nichteheliche Lebensgemeinschaften und Scheidung beschließt den Abschnitt über binäre Modelle.

3.1.1 Logit- und Probit-Modelle

Nehmen wir an, wir wollen anhand einer Stichprobe von Ehen untersuchen, von welchen Faktoren es abhängt, ob während der ersten zehn Ehejahre eine Scheidung erfolgt. Es gibt mithin zwei mögliche Ereignisse: keine Scheidung ($y = 0$) oder Scheidung ($y = 1$). Die abhängige Variable Y nimmt nur zwei Werte an. Um zu Regressionsverfahren für solche Variablen zu gelangen, wendet man den „Trick“ an, ein Wahrscheinlichkeitsmodell zu spezifizieren. Die Wahrscheinlichkeiten für die beiden möglichen Ereignisse formulieren wir in Abhängigkeit von Kovariaten \mathbf{x}_i als (der Personenindex i wird im folgenden meist nicht mehr berücksichtigt)

$$\begin{aligned}P(Y_i = 1) &= F(\boldsymbol{\beta}'\mathbf{x}_i), \\P(Y_i = 0) &= 1 - F(\boldsymbol{\beta}'\mathbf{x}_i).\end{aligned}$$

$F(\cdot)$ steht hier für eine Verteilungsfunktion. Verteilungsfunktionen sind bekanntlich monoton steigende Funktionen ihres Arguments. In unserem Beispiel würde damit ein positiver Koeffizient β_j bedeuten, daß die Wahrscheinlichkeit einer Scheidung mit X_j ansteigt. Das Problem ist nun, $F(\cdot)$ so zu spezifizieren, daß ein leicht handhabbares Modell resultiert.³¹ Die einfachste Möglichkeit ist die Spezifikation $F(\boldsymbol{\beta}'\mathbf{x}) = \boldsymbol{\beta}'\mathbf{x}$. Dies ist das sogenannte *lineare Wahrscheinlichkeitsmodell*.

³¹Man beachte, daß dieses Modell (ebenso wie die Zähldatenmodelle) im Unterschied zum normalen Regressionsmodell keinen Fehlerterm hat. Dies führt manchmal zu der irrigen Meinung, daß es sich um ein „deterministisches“ Modell handle, weil keine stochastischen Abweichungen zugelassen seien. Dem ist natürlich nicht so, da das Modell selbst stochastisch formuliert ist. Wie wir unten sehen werden, kann dieses Modell anhand der latenten Variable Y^* wie ein normales Regressionsmodell mit Fehlerterm formuliert werden. Umgekehrt kann eine gewöhnliche Regression stochastisch, aber ohne Fehlerterm formuliert werden als $E(y|\mathbf{x}) = \boldsymbol{\beta}'\mathbf{x}$.

Wendet man trotz dichotomer abhängiger Variable OLS an, so schätzt man dieses Modell. Die Fehlerterme der Regression sind allerdings heteroskedastisch und die vom linearen Wahrscheinlichkeitsmodell prognostizierten Wahrscheinlichkeiten können ohne weiteres außerhalb des Intervalls (0,1) liegen. Vor dem Aufkommen leistungsfähiger Software zur Maximierung von Likelihoods wurde dieses Modell dennoch häufig verwendet. Um beide Probleme zu vermeiden, sollte man allerdings Modelle spezifizieren, für die folgendes gilt:

$$\begin{aligned}\lim_{\boldsymbol{\beta}'\mathbf{x} \rightarrow +\infty} P(Y = 1) &= 1, \\ \lim_{\boldsymbol{\beta}'\mathbf{x} \rightarrow -\infty} P(Y = 1) &= 0.\end{aligned}$$

$F(\cdot)$ sollte also eine Funktion sein, die sich den „extremen“ Wahrscheinlichkeiten null und eins asymptotisch annähert. Eine naheliegende Wahl ist die Normalverteilung $F(\boldsymbol{\beta}'\mathbf{x}) = \Phi(\boldsymbol{\beta}'\mathbf{x})$, woraus das *Probit-Modell* resultiert (s. auch Abbildung (11)):

$$P(Y = 1) = \Phi(\boldsymbol{\beta}'\mathbf{x}) = \int_{-\infty}^{\boldsymbol{\beta}'\mathbf{x}} \phi(t) dt.$$

Ähnlich häufig wird die logistische Verteilung verwendet $F(\boldsymbol{\beta}'\mathbf{x}) = \Lambda(\boldsymbol{\beta}'\mathbf{x})$, wobei $\Lambda(\cdot)$ die Verteilungsfunktion der logistischen Verteilung notiert. Für das *Logit-Modell* gilt

$$P(Y = 1) = \frac{e^{\boldsymbol{\beta}'\mathbf{x}}}{1 + e^{\boldsymbol{\beta}'\mathbf{x}}}.$$

Weitere Modelle können analog formuliert werden, allerdings werden fast ausschließlich diese beiden verwendet. Die logistische Verteilung und die Normalverteilung sind sich recht ähnlich, nur daß erstere dickere Ränder hat. Die Schätzergebnisse beider Modelle sind in den meisten Situationen (außer wenn die Stichprobe fast nur Nullen oder Einsen enthält) bis auf einen Skalierungsfaktor annähernd identisch: $\hat{\beta}_{\text{logit}} \approx 1,6 \cdot \hat{\beta}_{\text{probit}}$ (Greene (1993: 640) gibt eine schöne Herleitung dieses Skalierungsfaktors).

Alternativ kann man die binären Modelle auch als *Schwellenwertmodell* mittels einer stetigen, latenten Variable Y^* formulieren. In unserem Scheidungsbeispiel könnte Y^* für einen Ehepartner die (unbeobachtete) Differenz des Nutzens aus einer Scheidung und des Nutzens aus der Ehe sein. Wir spezifizieren nun folgendes Regressionsmodell für Y^* :

$$y^* = \boldsymbol{\beta}'\mathbf{x} + \epsilon,$$

wobei wir allerdings Y^* nicht kennen. Wir kennen nur die resultierende binäre Variable Y , die sich aus dem folgenden Schwellenwertmodell ergibt:

$$\begin{aligned}y &= 1, & \text{wenn } y^* > 0, \\ y &= 0, & \text{wenn } y^* \leq 0.\end{aligned}$$

Nun müssen wir für ϵ eine Verteilungsannahme treffen. Bei Annahme einer logistischen Verteilung erhalten wir das Logit-Modell. Ist $\epsilon \sim N(0, 1)$, so erhalten wir das Probit-Modell, denn

$$\begin{aligned}
 P(Y = 1) &= P(y^* > 0) \\
 &= P(\boldsymbol{\beta}'\mathbf{x} + \epsilon > 0) \\
 &= P(\epsilon > -\boldsymbol{\beta}'\mathbf{x}) \\
 &= P(\epsilon < \boldsymbol{\beta}'\mathbf{x}) \quad (\text{wegen Symmetrie der Verteilung}) \\
 &= \Phi(\boldsymbol{\beta}'\mathbf{x}).
 \end{aligned}$$

Zwei Annahmen dieses Modells bedürfen der Erläuterung. Erstens muß σ^2 auf eins normiert werden, weil es nicht identifizierbar ist. Dies ist in der Struktur unserer Beobachtungsgleichung für Y begründet: Egal welche Varianz Y^* aufweist, wir werden immer dieselben Daten Y beobachten. Weiterhin ist die Annahme einer Schwelle von null unerheblich, solange das Modell eine Konstante enthält.

Die *Interpretation der Koeffizienten* dieser Modelle (s. hierzu Liao, 1994; Kaufman, 1996) ist nicht einfach und man findet häufig Forschungsarbeiten, in denen Logit- oder Probit-Effekte fehlerhaft interpretiert werden (davon zeugen die vielen Kommentare über die korrekte Interpretation der Effekt, s. z.B. Petersen, 1985; Roncek, 1991). Im Prinzip können die Koeffizienten der binären Modelle als Effekte auf die latente Variable interpretiert werden. Dies ist jedoch bei den meisten Anwendungen wenig hilfreich, da die latente Variable nicht beobachtbar ist und auch beliebig skaliert werden kann. Bei der Interpretation der Koeffizienten als Effekte auf die beobachteten Daten ist aber zu beachten, daß die Koeffizienten selbst weder Einheits- noch Marginaleffekte sind. Im Prinzip ist nur ihr Vorzeichen sinnvoll zu interpretieren (eine Ausnahme ist das Logit-Modell, s.u.). Die Koeffizienten müssen erst transformiert werden, um zu den *Marginaleffekten* zu gelangen. Für den bedingten Erwartungswert von Y gilt allgemein

$$\begin{aligned}
 E(y|\mathbf{x}) &= 0(1 - F(\boldsymbol{\beta}'\mathbf{x})) + 1(F(\boldsymbol{\beta}'\mathbf{x})) \\
 &= F(\boldsymbol{\beta}'\mathbf{x}),
 \end{aligned}$$

was identisch ist mit $P(Y = 1)$. Daraus ergibt sich für die Marginaleffekte (β_j^*)

$$\frac{\partial E(y|\mathbf{x})}{\partial x_j} = \frac{\partial P(Y = 1)}{\partial x_j} = f(\boldsymbol{\beta}'\mathbf{x})\beta_j.$$

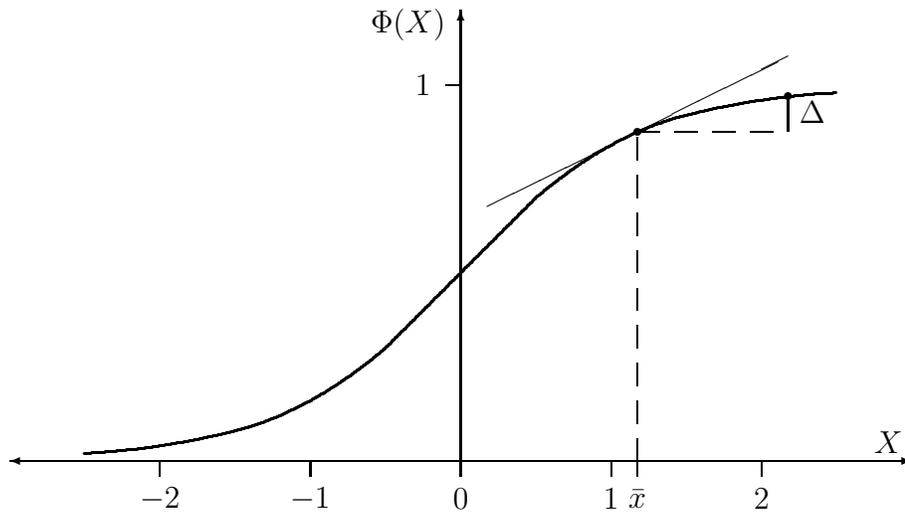
$f(\cdot)$ ist die zu $F(\cdot)$ gehörige Dichtefunktion. Für das Probit-Modell ergibt sich

$$\frac{\partial E(y|\mathbf{x})}{\partial x_j} = \phi(\boldsymbol{\beta}'\mathbf{x})\beta_j$$

und für das Logit-Modell

$$\frac{\partial E(y|\mathbf{x})}{\partial x_j} = \Lambda(\boldsymbol{\beta}'\mathbf{x})(1 - \Lambda(\boldsymbol{\beta}'\mathbf{x}))\beta_j.$$

Abbildung 11: Marginal- und Einheitseffekt im Probit-Modell



Die Koeffizienten müssen also jeweils mit einem Skalierungsfaktor multipliziert werden, um die Marginaleffekte zu erhalten. Man beachte, daß die Marginaleffekte eine Funktion von \mathbf{X} sind, was bedeutet, daß sie bei verschiedenen Kovariatenkombinationen unterschiedlich hoch ausfallen werden. Am informativsten ist es, wenn man jeden Marginaleffekt in Abhängigkeit von \mathbf{X} graphisch aufträgt. Will man jedoch eine Tabelle erstellen, so muß man sich für eine Kovariatenkombination entscheiden. Üblicherweise wählt man hierzu den Mittelwert aller Kovariaten und benutzt $\beta' \bar{\mathbf{x}}$ zur Errechnung des Skalierungsfaktors. $\beta_j^* \cdot 100$ gibt an, um wieviele Prozentpunkte sich $E(y|\bar{\mathbf{x}})$ bzw. $P(Y = 1)$ in etwa verändert, wenn sich X_j um eine Einheit erhöht. Angenommen in unserem Scheidungsbeispiel sei X_j das Alter bei Eheschließung und $\beta_j^* = -0,01$ am Mittelwert aller Kovariaten, so senkt eine um ein Jahr verschobene Heirat die (erwartete) Scheidungswahrscheinlichkeit um einen Prozentpunkt.

Diese Interpretation der Marginaleffekte als Einheitseffekte ist allerdings nur näherungsweise gültig, denn in einem nicht-linearen Modell sind die Marginal-effekte nicht identisch mit den Einheitseffekten, weil ein Marginal-effekt bei einer nicht-linearen Funktion nur für infinitesimale Änderungen der X -Variablen die Veränderung der Y -Variable angibt. Dies ist in Abbildung (11) verdeutlicht. Gehen wir von einem Probit-Modell mit nur einer Kovariaten X aus. Der Effekt dieser Kovariaten sei gleich eins und die Konstante sei gleich null. Dann ist $P(Y = 1) = \Phi(x)$. Diese Funktion ist in der Abbildung aufgetragen. Angenommen $\bar{x} = 1,175$, dann ist der Marginal-effekt $\phi(1,175) = 0,20$, was der Steigung der eingezeichneten Tangente entspricht. Um den *exakten Einheitseffekt* zu erhalten, berechnen wir $\Phi(\bar{x}) = 0,88$ und $\Phi(\bar{x} + 1) = 0,99$. Der exakte Einheitseffekt

ergibt sich dann aus der Differenz der beiden Wahrscheinlichkeiten und beträgt $\Delta = 0,11$. In diesem Beispiel ist der Marginaleffekt also deutlich größer als der Einheitseffekt. Der gegenteilige Fall ergäbe sich für $\bar{x} = -1,175$. Wir erkennen anhand der Abbildung aber auch, daß die Normalverteilungsfunktion für Wahrscheinlichkeiten von 0,2 bis 0,8 annähernd linear ist, weshalb die Marginaleffekte in diesem Intervall recht gut mit den Einheitseffekten übereinstimmen werden (dies gilt auch für die logistische Verteilung). Ein weiteres Problem der Marginaleffekte ist, daß sie für Dummy-Variablen im Prinzip nicht definiert sind. Dennoch sind die mit den obigen Formeln für Dummy-Variablen errechneten Marginaleffekte ebenso wie bei einer stetigen Kovariate Näherungen des Einheitseffektes. Hat man viele Dummies im Modell, so empfiehlt es sich allerdings nicht die Marginaleffekte zu berichten, sondern die Wahrscheinlichkeiten explizit auszuweisen. Dazu wählt man vorteilhaft als Referenzgruppe den Fall, bei dem alle Dummies gleich null sind. Dann berechnet man die Wahrscheinlichkeit des Ereignisses für diese Referenzgruppe und all die Fälle, bei denen jeweils eine Kovariate gleich eins ist. Die entsprechenden Wahrscheinlichkeitsdifferenzen ergeben dann die Einheitseffekte der Dummies, welche anschaulich etwa in Form eines Balkendiagramms präsentiert werden können (s. Brüderl et al., 1996).

Die Koeffizienten des Logit-Modells können auch direkt als Einheitseffekte interpretiert werden. Um dies zu erkennen, muß das Modell anders niedergeschrieben werden. Aus obigem Ausdruck für das Logit-Modell folgt (P_1 steht hier für $P(Y = 1)$)

$$\ln\left(\frac{P_1}{1 - P_1}\right) = \beta' \mathbf{x}.$$

Der Term auf der linken Seite heißt „Logit“ (oder „Log-Odds“). Er repräsentiert in unserem Beispiel das (logarithmierte) Verhältnis der Scheidungswahrscheinlichkeit zur Wahrscheinlichkeit des Fortbestandes einer Ehe. Ein β_j erfaßt mithin den Einheitseffekt auf das Logit. Das Logit ist aber keine „natürliche“ Größe, weshalb diese Interpretation noch nicht allzu anschaulich ist. Faßbarer sind allerdings die antilogarithmierten Koeffizienten e^{β_j} . Sie geben den Einheitseffekt auf die *Odds-Ratio* (das Verhältnis der Odds) an. Betrachten wir eine Dummy-Variable X_j . Für die Odds-Ratio der beiden Gruppen ergibt sich

$$\frac{P_1(X_j = 1)/P_0(X_j = 1)}{P_1(X_j = 0)/P_0(X_j = 0)} = e^{\beta_j}.$$

Nehmen wir an, der Effekt einer Land-Stadt Dummy in einer logistischen Scheidungsregression sei 0,693. Dann bedeutet dies, daß die Scheidungs-Odds in der Stadt doppelt so hoch sind ($\exp(0,693) = 2$). Häufig liest man auch die Interpretation, die Scheidungswahrscheinlichkeit sei in der Stadt doppelt so hoch wie auf dem Land. Diese Interpretation ist allerdings falsch und allenfalls für sehr kleine Wahrscheinlichkeiten näherungsweise gültig. Eine ausführliche Darstellung der

Odds-Ratio Interpretation der Koeffizienten des Logit-Modells findet man bei Morgan und Teachman (1988).

Die Parameter der binären Modelle können mit der Maximum-Likelihood Methode geschätzt werden. Die *Log-Likelihood* lautet:

$$\ln L = \sum_{i=1}^n \{y_i \ln[F(\boldsymbol{\beta}'\mathbf{x}_i)] + (1 - y_i) \ln[1 - F(\boldsymbol{\beta}'\mathbf{x}_i)]\}.$$

Die ersten und zweiten Ableitungen sind relativ einfach zu errechnen (s. Greene, 1993: 643ff), so daß die üblichen Maximierungsverfahren eingesetzt werden können. Es kann bei diesen Modellen allerdings leicht vorkommen, daß der Algorithmus nicht konvergiert bzw. ungewöhnlich große Parameterschätzer und Standardfehler ausgibt. Dies deutet darauf hin, daß Y „monoton“ in einer Kovariaten ist. Bei einer Dummy können etwa alle Beobachtungen einer Gruppe $y = 1$ aufweisen. Auch mit stetigen Variablen kann dieses Problem auftreten, wenn z.B. für alle Werte über einer bestimmten Schwelle $y = 1$ gilt. Erstellt man eine Kreuztabelle zwischen den betroffenen Kovariaten und der abhängigen Variable, so sind einzelne Zellen nicht besetzt. Folge wird sein, daß $\hat{\beta}_j$ während der Iterationen gegen unendlich läuft. In diesem Fall muß die betreffende Variable anders verkodet werden.

Man kann für Logit- und Probit-Modelle die oben vorgestellten *Fitmaße* für Maximum-Likelihood Modelle verwenden. Eine Vereinfachung bei binären Modellen ist, daß die Log-Likelihood für das Modell mit nur der Konstanten einfach zu berechnen ist als

$$\ln L_R = n[P \ln(P) + (1 - P) \ln(1 - P)],$$

wobei P der Anteil der Einsen in der Stichprobe ist. Ein weiteres Fitmaß kann für binäre Modelle über den Anteil der korrekt vorhergesagten Beobachtungen konstruiert werden, indem man jede Beobachtung aufgrund ihres \hat{F}_i nach folgender Regel klassifiziert:

$$\hat{y}_i = 1, \quad \text{wenn } \hat{F}_i > 0,5, \quad \text{sonst } 0.$$

Der Anteil der korrekt vorhergesagten Beobachtungen ergibt sich dann aus dem Vergleich von \hat{y}_i und y . Dieses Fitmaß hat allerdings einen gravierenden Mangel. Mit der naiven Vorhersageregeln $\hat{y}_i = 1$ erzielt man $P \cdot 100\%$ korrekte Vorhersagen. Ist nun P sehr groß (viele Einsen in der Stichprobe), so ist die naive Regel leicht besser als das Modell. Deshalb ist dieses Fitmaß mit Vorsicht zu genießen, falls P deutlich von 0,5 abweicht. Es sind allerdings verfeinerte Fitmaße, die auf dem Anteil der korrekt vorhergesagten Beobachtungen beruhen, vorgeschlagen worden. Wie Veall und Zimmermann (1992) zeigen, sind einige dieser Maße durchaus brauchbar.

Ebenso wie im Fall der linearen Regression empfiehlt es sich bei diesen binären Modellen *diagnostische Verfahren* einzusetzen. In SPSS für Windows sind beispielsweise für das Logit-Modell die Analoga zu Cook's D und DFBETAS implementiert. Wie im Anwendungsbeispiel zur linearen Regression demonstriert, können diese Statistiken zur Identifikation einflußreicher Datenpunkte eingesetzt werden. Diese und weitere diagnostische Verfahren für binäre Modelle werden bei Pregibon (1981), Hosmer und Lemeshow (1989: Kap. 5), Schnell (1994: Kap. 8.5) und bei Menard (1995: Kap. 4) beschrieben.

3.1.2 Das bivariate Probit-Modell

Ähnlich wie in Abschnitt 2.3 für die lineare Regression erörtert, wird man auch bei binären Modellen häufig mehrere Regressionsgleichungen spezifizieren. Besteht dann eine Korrelation zwischen den Fehlertermen, so kann die Verwendung getrennter Modelle einen Effizienzverlust oder gar Inkonsistenz der Parameterschätzer zur Folge haben. Deshalb wurden auch für binäre Modelle multivariate Ansätze formuliert.³² Allerdings gilt dies nur für das Probit-Modell und hierfür wiederum nur für den bivariaten Fall. Bei mehr als zwei Gleichungen kann zwar ein multivariates Probit-Modell formuliert werden, welches aber bis dato wegen der dafür notwendigen Berechnung mehrdimensionaler Integrale praktisch nicht schätzbar ist. Die sich gegenwärtig vollziehende Entwicklung von Simulationsschätzern für Wahrscheinlichkeitsfunktionen mit mehrdimensionalen Integralen wird hier aber demnächst wohl Abhilfe schaffen (Überblick bei Monfort und Dijk, 1995). Im folgenden können diese Simulationsschätzer allerdings nicht dargestellt werden, weshalb nur das *bivariate Probit-Modell* vorgestellt wird.

Wir formulieren das Modell anhand der latenten Variablen Y_1^* und Y_2^* wie folgt:

$$\begin{aligned} y_1^* &= \beta_1' \mathbf{x}_1 + \epsilon_1 \\ y_2^* &= \beta_2' \mathbf{x}_2 + \epsilon_2 \\ y_1 &= 1, \quad \text{wenn } y_1^* > 0, \quad \text{sonst } 0 \\ y_2 &= 1, \quad \text{wenn } y_2^* > 0, \quad \text{sonst } 0. \end{aligned}$$

Dies ist ein bivariates Probit-Modell, wenn man annimmt, daß die Fehlerterme bivariat normalverteilt sind: $(\epsilon_1, \epsilon_2) \sim N(0, 0, 1, 1, \rho)$. Der entscheidende Unterschied zu zwei getrennten Probits ist in diesem Modell, daß eine Korrelation der Fehlerterme (ρ) zugelassen wird. Ein Test auf die $H_0 : \rho = 0$ gibt Aufschluß darüber, ob getrennte Modelle statthaft sind. Man beachte, daß dies *kein* Simultanmodell ist. Simultane Logit- und Probit-Modelle können zwar ebenfalls formuliert werden, erfordern jedoch zusätzlich zu den üblichen Identifikationsbedingungen komplexe Konsistenzbedingungen (s. Ronning, 1991: Kap. 2.8). Da

³²Auch für Simultanmodelle mit einer quantitativen und einer binären abhängigen Variable gibt es inzwischen geeignete Verfahren (einen Überblick geben Bollen et al., 1995).

obiges Modell als reduzierte Form niedergeschrieben ist, sind hier keine weiteren Restriktionen nötig. Es kann sogar $\mathbf{x}_1 = \mathbf{x}_2$ sein.

Weil wir zwei abhängige Variablen mit je zwei Ausprägungen haben, läßt sich jede Beobachtung einem von vier Fällen zuordnen. Für jeden der vier Fälle läßt sich anhand der Formel für die bivariate Normalverteilung ein Ausdruck für die Wahrscheinlichkeitsdichte angeben. Zum Beispiel lautet dieser Ausdruck für eine Beobachtung ($y_1 = 1, y_2 = 1$):

$$P(Y_1 = 1, Y_2 = 1) = \int_{-\infty}^{\beta_2' \mathbf{x}_2} \int_{-\infty}^{\beta_1' \mathbf{x}_1} \phi_2(z_1, z_2, \rho) dz_1 dz_2,$$

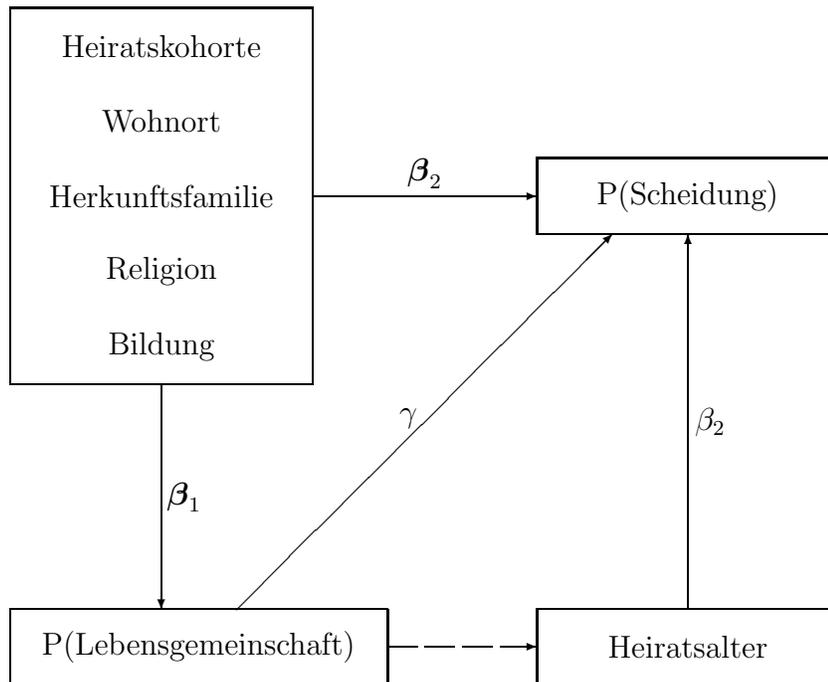
wobei $\phi_2(\cdot)$ die Dichte der bivariaten Normalverteilung ist. Multiplikation über alle Beobachtungen ergibt die Likelihood. Die ersten und zweiten Ableitungen dieser Likelihood können analytisch bestimmt werden (s. Greene, 1993: 660ff), weshalb Maximum-Likelihood Schätzer mit den üblichen Algorithmen erhältlich sind.

Das bivariate Probit-Modell ermöglicht in einfacher Art und Weise die Berücksichtigung von *Selektivität*. Wie im Stichprobenauswahl-Modell aus Abschnitt 2.4.2 ist die Auswahlgleichung ein Probit-Modell, aber die interessierende Regression ist nun ebenfalls ein Probit. Nimmt man eine bivariate Normalverteilung für die Fehlerterme der beiden Gleichungen an, so hat man ein bivariates Probit-Modell vorliegen (s. Dubin und Rivers, 1989). Der einzige Unterschied zu oben ist, daß die Fallzahl der beiden Probits unterschiedlich ist, denn für die nicht-ausgewählten Beobachtungen liegt keine Information über die abhängige Variable des zweiten Probits vor. Deshalb sind im bivariaten Probit-Modell mit Selektion auch nur drei Fälle möglich: nicht-ausgewählt ($y_1 = 0$), ausgewählt und null ($y_1 = 1, y_2 = 0$), ausgewählt und eins ($y_1 = 1, y_2 = 1$). In der Likelihood können also nur drei verschiedene Ausdrücke auftauchen. Der Schätzer von ρ gibt dann wie im Heckit-Modell die Korrelation der Fehlerterme von Auswahlgleichung und interessierender Regressionsgleichung wieder. Anwendungen dieses Modells findet man u.a. bei Haveman und Wolfe (1994) und Brüderl et al. (1996).

3.1.3 Anwendungsbeispiel: Nichteheliche Lebensgemeinschaften und Scheidung

Mit der Zunahme von nichtehelichen Lebensgemeinschaften seit den 60er Jahren begann auch die intensive Beschäftigung der Bevölkerungswissenschaft mit dieser Lebensform. Eine interessante Frage in diesem Zusammenhang ist die nach den Auswirkungen dieser Lebensform auf die Stabilität einer eventuellen späteren Ehe. Mit der bekannten Umschreibung „Ehe auf Probe“ ist die am weitesten verbreitete Hypothese diesbezüglich ausgedrückt: Eine nichteheliche Lebensgemeinschaft ist eine Art „Testphase“ vor der Ehe. Paare, die nicht zueinander

Abbildung 12: Ein Kausalmodell der Beziehung von nichtehelicher Lebensgemeinschaft und Scheidung



passen, trennen sich wieder und heiraten nicht. Damit sollten die Lebensgemeinschaften, die schließlich doch heiraten, eine höhere Ehestabilität aufweisen. Dem widersprechend fanden aber viele empirische Studien ein deutlich höheres Scheidungsrisiko von Ehepaaren, die bereits vorher zusammen lebten. Zwei mögliche Erklärungen hierfür wurden angeboten (s. Axinn und Thornton, 1992; Thomson und Colella, 1992): i) Personen, die vor einer Ehe eine nichteheliche Lebensgemeinschaft eingehen, sammeln Erfahrungen mit einer Alternative zur Ehe und es fällt ihnen deshalb bei einer Ehekrise leichter, sich auf die bereits bekannte Alternative zurückzuziehen. Diese Hypothese postuliert einen direkten Kausaleffekt einer nichtehelichen Lebensgemeinschaft auf die Scheidungswahrscheinlichkeit. ii) Personen, die vor der Ehe eine nichteheliche Lebensgemeinschaft eingehen, unterscheiden sich in ihren grundlegenden Werthaltungen und Präferenzen von denen, die sofort heiraten: Sie messen der Institution Ehe geringeren Wert bei und sind eher hedonistisch orientiert. Deshalb sind diese Personen auch schneller bereit, sich im Krisenfall scheiden zu lassen. Diese Hypothese postuliert keinen Kausaleffekt, sondern Selbst-Selektion als Ursache für die höhere Scheidungswahrscheinlichkeit.

Um herauszufinden, welche der beiden Erklärungen die richtige ist, verwenden

wir das Modell in Abbildung (12).³³ Die Wahrscheinlichkeit eine nichteheliche Lebensgemeinschaft einzugehen, hängt von verschiedenen Faktoren ab, von denen einige in der Abbildung angeführt sind. Die Effekte β_1 kann man z.B. mit dem Probit-Modell schätzen. Dasselbe gilt für die Effekte β_2 des Scheidungsmodells. Die Scheidungswahrscheinlichkeit hängt von einigen weiteren Faktoren ab, von denen stellvertretend das Heiratsalter angeführt ist. Ein eventueller, direkter Kausaleffekt einer nichtehelichen Lebensgemeinschaft wird von γ erfaßt (zusätzlich könnte noch ein indirekter Effekt über das Heiratsalter auftreten). Schreibt man dieses Modell mit zwei latenten abhängigen Variablen nieder, so erkennt man, daß es sich um ein rekursives Simultangleichungssystem handelt.

Das obige Selektivitätsargument besagt nun, daß es unbeobachtete Faktoren gibt, die beide Wahrscheinlichkeiten gleichzeitig beeinflussen. Z.B. ist die Variable „traditionelle Einstellung zur Ehe“ nicht in unserem Modell enthalten, aber es ist zu vermuten, daß Personen mit „moderner“ Einstellung sowohl häufiger eine nichteheliche Lebensgemeinschaft eingehen als sich auch häufiger scheiden lassen. Diese unbeobachteten Faktoren sind in den beiden Fehlertermen enthalten, welche miteinander korreliert sein sollten. Damit haben wir strenggenommen ein Identifikationsproblem vorliegen, weil die Scheidungsgleichung nicht identifiziert ist.³⁴ Deshalb berücksichtigen wir im folgenden die Bildung nicht im Scheidungsmodell. Die Korrelation der Fehlerterme der beiden Probit-Modelle hat noch ein zweites Problem zur Folge: Die Wahrscheinlichkeit einer nichtehelichen Lebensgemeinschaft und der Fehlerterm der Scheidungsgleichung sind korreliert, weshalb der Schätzer von γ verzerrt ist und eventuell der falsche Eindruck eines direkten Kausaleffektes entstehen kann. Dieses Problem bekommt man mit dem bivariaten Probit-Modell in den Griff. Es berücksichtigt die Korrelation der Fehlerterme und beseitigt damit auch die Verzerrung des Schätzers von γ . Trifft das Selektivitätsargument zu, so sollte im bivariaten Probit-Modell eine deutliche Korrelation der beiden Gleichungen feststellbar sein und γ gegen null gehen.³⁵

Wir schätzen dieses Modell anhand der Daten des DJI-Familiensurveys West (Infratest, 1989). Im Auftrag des Deutschen Jugendinstitutes wurde 1988 eine

³³Dieses Modell und auch das im folgenden angewandte analytische Vorgehen wurden von Kahn und London (1991) zur Analyse einer ähnlichen Fragestellung eingesetzt: Sie untersuchten, warum vorehelicher Geschlechtsverkehr die Scheidungswahrscheinlichkeit erhöht. Es zeigte sich, daß dieser Zusammenhang ein Resultat von Selbst-Selektion ist.

³⁴Die Scheidungsgleichung wäre wegen der Nicht-Linearität des Probit-Modells allerdings dennoch schätzbar. Genauer behandeln die Identifikationsbedingungen solcher Modelle Winship und Mare (1983).

³⁵Wir haben es hier mit dem in der Sozialforschung häufig auftretenden Problem der Isolierung eines „Treatment-Effektes“ mittels Surveydaten zu tun. Bei fehlender Randomisierungsmöglichkeit kann es immer zu Selbst-Selektion kommen, was einen Kausaleffekt vortäuschen kann. Könnten wir Versuchspaare per Randomisierung dem „Treatment“ nichteheliche Lebensgemeinschaft zuweisen, so könnte der „wahre“ Kausaleffekt ohne Schwierigkeiten bestimmt werden (s. Heckman und Robb, 1985).

Zufallsstichprobe von 10.043 Personen im Alter von 18 bis 55 Jahren zu ihrer Familienbiographie befragt. Für unsere Analysen beschränken wir uns auf Personen, die zwischen 1949 und 1977 heirateten und vollständige Angaben zu ihrer Partnerbiographie machten ($N=4467$). Es wird jeweils die erste Heirat einer Person betrachtet. Eine nichteheliche Lebensgemeinschaft liegt vor, wenn der Befragte angab, vor dem Eheschließungsjahr bereits mit dem Ehepartner zusammengezogen zu sein. Dies ist bei 12% der Ersten der Fall. Um die Scheidungsvariable zu konstruieren, muß man sich auf einen Zeitraum festlegen. Wir wählen zehn Jahre und sprechen von einer Scheidung, wenn die Ehepartner spätestens nach zehn Jahren getrennt lebten oder geschieden waren. Bei 10% der Ehen trat eine so definierte Scheidung auf. Ein Problem dieser Querschnittbetrachtung sind Ehen, bei denen ein Ehepartner vor dem zehnten Ehejahr starb. Hier kann man keine Aussage darüber machen, ob die Ehe zehn Jahre Bestand hatte oder nicht. Deshalb müssen diese (zensierten) Ehen aus der Analyse ausgeschlossen werden (deshalb beschränken wir uns auch auf bis 1977 geschlossene Ehen). Als unabhängige Variablen werden die in Abbildung (12) angeführten Variablen berücksichtigt.

Mit der hier vorgenommenen Dichotomisierung der kontinuierlichen Variable „Ehedauer“ verschenkt man natürlich Information über das „Timing“ der Scheidung und muß die zensierten Ehen ausschließen. Insofern sind Logit- bzw. Probit-Modelle bei unserer Anwendung nicht effizient. Verfahren der Ereignisdatenanalyse wären effizienter. Brüderl et al. (1997) untersuchen auch mit diesen Verfahren den Scheidungsprozeß im DJI-Familiensurvey. Ihre Ergebnisse stimmen qualitativ mit den hier berichteten Ergebnissen überein, was zeigt, daß Ereignisdatenmodelle und binäre Regressionsmodelle in diesem Fall zu ähnlichen Ergebnissen führen. Ein Nachteil der Ereignisdatenverfahren ist jedoch, daß Modelle zur Analyse bivariater Prozesse noch nicht allgemein verfügbar sind (s. aber Lillard et al., 1995).

In Tabelle (3) sind die Ergebnisse der Probit-Schätzungen aufgeführt (alle Modelle wurden mit LIMDEP geschätzt, s. Greene, 1995: Kap. 21 und 22). Es sind die Marginaleffekte am Mittelwert aller Kovariaten angegeben. In den ersten beiden Spalten sind die Schätzer zweier getrennter Probit-Modelle angegeben. Pseudo- R^2 fällt für beide Modelle relativ bescheiden aus. Betrachten wir nun den uns besonders interessierenden Effekt: Personen, die vor der Ehe eine nichteheliche Lebensgemeinschaft eingingen, haben eine um fünf Prozentpunkte höhere Scheidungswahrscheinlichkeit. Berücksichtigt man, daß die vom Modell vorhergesagte Scheidungswahrscheinlichkeit am Mittelwert aller Kovariaten 7,8% ist, so entspricht dies einer Erhöhung der Scheidungswahrscheinlichkeit um etwa 60%. Die getrennten Probit-Modelle legen also den Schluß nahe, daß ein scheidungsfördernder Kausaleffekt einer Lebensgemeinschaft vorhanden ist. Verwendet man aber ein bivariates Probit-Modell (s. Spalten 3 und 4), so fällt der Lebensgemeinschafts-Effekt kleiner aus und ist nicht signifikant. Dies läßt den Schluß zu, daß der Schätzer aus der univariaten Probit-Gleichung verzerrt war, daß also kein Kausaleffekt vorliegt sondern nur Selbst-Selektion. Der Schätzer

Tabelle 3: Probit Marginaleffekte auf die Wahrscheinlichkeit einer nichtehelichen Lebensgemeinschaft und einer Scheidung (N=4432)

Variable	Probit		Bivariates Probit	
	Lebensgem.	Scheidung	Lebensgem.	Scheidung
Konstante	-0,33*	-0,08*	-0,33*	-0,08*
Lebensgem. (1=Ja)	—	0,05* (4,34)	—	0,03 (0,24)
Heiratskohorte (1=1961-70)	0,04* (2,56)	0,08* (5,99)	0,04* (2,55)	0,08* (5,49)
Heiratskohorte (1=1971-77)	0,13* (9,41)	0,11* (8,22)	0,13* (9,45)	0,11* (4,65)
Kleinstadt (1=bis 4999 Einw.)	-0,06* (4,04)	-0,05* (3,47)	-0,06* (4,01)	-0,05* (3,16)
Mittelstadt (1=5000-99999)	-0,03* (3,21)	-0,03* (3,27)	-0,04* (3,22)	-0,03* (2,77)
Intakte Familie (1=Ja)	-0,01 (0,54)	-0,04* (4,40)	-0,01 (0,52)	-0,04* (4,33)
Beide katholisch (1=Ja)	-0,04* (3,67)	-0,05* (4,69)	-0,04* (3,69)	-0,05* (4,02)
Aktiver Gläubiger (1=Ja)	-0,05* (3,68)	-0,05* (4,03)	-0,05* (3,72)	-0,05* (3,73)
Schulbildung (in Jahren)	0,01* (2,83)	—	0,01* (2,75)	—
Alter bei Heirat (in Jahren)	—	-0,01* (4,87)	—	-0,01* (5,41)
ρ	—	—	0,08 (0,19)	
R_{MF}^2	0,07	0,08	0,07	
-Log-Likelihood	1483,4	1296,6	2780,0	

* signifikant auf dem 5%-Niveau. Absolute t-Werte in Klammern. McFadden Pseudo- R^2 Werte in bezug auf das jeweilige Modell mit nur einer Konstanten. Abhängige Variablen sind „nichteheliche Lebensgemeinschaft vor der Heirat“ und „Scheidung nach zehn Ehejahren“. Bezugsgruppe ist Heiratskohorte 1949-60, Großstadt mit über 100.000 Einwohnern, Kindheit nicht mit beiden Eltern verbracht, andere Religion oder Atheist, weniger als einmal pro Monat Kirchenbesuch.

Quelle: Eigene Berechnungen mit Daten des DJI-Familiensurvey West.

für ρ ist positiv, allerdings nicht signifikant. Ein Likelihood-Ratio Test weist das bivariate Probit auch nicht als signifikant besseres Modell aus: Die Summe der Log-Likelihoods der getrennten Modelle ist 2780,02, was einen χ^2 -Wert von 0,05 ergibt, der bei einem Freiheitsgrad nicht signifikant ist. Trotzdem sind aufgrund der Schätzergebnisse deutliche Zweifel an der Existenz eines Kausaleffektes einer nichtehelichen Lebensgemeinschaft angebracht.

Was besonders auffällt ist, daß alle anderen Effekte im bivariaten Probit-Modell fast identisch zu denen aus den getrennten Modellen sind. Dies deutet darauf hin, daß die anderen Effekte nicht vom Selektivitäts-Problem betroffen sind. Je jünger die Heiratskohorte ist, desto häufiger gehen die Paare nichteheliche Lebensgemeinschaften ein und desto höher ist die Scheidungswahrscheinlichkeit (die Kohorteneffekte sind die stärksten des gesamten Modells). Dasselbe gilt für die Größe des Wohnortes. Dagegen gehen Personen, die aus einer Familie mit beiden Elternteilen kommen, seltener Lebensgemeinschaften ein (nicht signifikant) und lassen sich auch seltener scheiden. Dasselbe gilt auch für Katholiken und aktive Kirchgänger aller Religionen. Mit zunehmender Schulbildung steigt die Wahrscheinlichkeit einer nichtehelichen Lebensgemeinschaft und je älter man bei der Heirat ist, desto geringer ist die Scheidungswahrscheinlichkeit.

Wie oben ausführlich diskutiert, sind die Marginaleffekte nur näherungsweise mit den Einheitseffekten identisch. Dies sei am Effekt einer Lebensgemeinschaft auf die Scheidungswahrscheinlichkeit demonstriert (anhand des univariaten Probits, Spalte 2). $\beta' \bar{x}$ beträgt -1,42, woraus sich eine Scheidungswahrscheinlichkeit von 7,8% ergibt. Der Probit-Effekt einer Lebensgemeinschaft ist 0,32, woraus eine Scheidungswahrscheinlichkeit von 13,7% resultiert. Mithin ist der Einheitseffekt einer Lebensgemeinschaft 5,9 Prozentpunkte, während der Marginaleffekt 4,7 Prozentpunkte ist (in der Tabelle sind gerundet fünf Prozentpunkte angeführt). Der Einheitseffekt ist größer wie der Marginaleffekt, weil wir uns im unteren Bereich der kumulierten Normalverteilung befinden, wo sich die Kurve nach oben wölbt. Weil der Probit-Effekt mit 0,32 relativ gering ist, fällt der Unterschied zwischen Marginal- und Einheitseffekt nicht allzu groß aus. Bei größeren Effekten kann der Unterschied aber beträchtlich werden: Der univariate Effekt der Heiratskohorte 1971-77 auf die Wahrscheinlichkeit einer Lebensgemeinschaft ist 0,73. $\beta' \bar{x} = -1,28$, woraus eine Wahrscheinlichkeit für die wilde Ehe von 10,0% resultiert. Addieren wir den Effekt der Heiratskohorte, so ist die Wahrscheinlichkeit 29,2%. Der Einheitseffekt ist mithin 19,2 Prozentpunkte, während in Tabelle (3) ein Marginaleffekt von 12,8 Prozentpunkten ausgewiesen ist.

3.2 Multinomiale Modelle

Verfahren für multinomiale abhängige Variablen sind komplexer wie die soeben vorgestellten für binäre Daten. Dies hat zur Folge, daß multinomiale Daten häufig

dichotomisiert werden. Dabei geht allerdings Information verloren. Angesichts der heute verfügbaren Software ist dies jedoch nicht mehr nötig. In den letzten Jahren findet man in der Forschungsliteratur auch immer häufiger Artikel, die multinomiale Verfahren einsetzen. Einige Beispiele hierfür sind Untersuchungen über die Art der Verhütung bei Jugendlichen (Kahn et al., 1990) und über die Todesursachen bei Neugeborenen (Eberstein et al., 1990). Zuerst wird das multinomiale Logit-Modell vorgestellt. Im Exkurs des zweiten Abschnitts wird eine entscheidungstheoretische Begründung für die Logit- und Probit-Modelle gegeben. Das Probit-Modell für ordinale Daten wird im dritten Abschnitt vorgestellt. Ein Anwendungsbeispiel über die Einstellung zur Ehe beschließt den Beitrag.

3.2.1 Das multinomiale Logit-Modell

Angenommen wir wollen untersuchen, von welchen Faktoren die Wahl des Verhütungsmittels abhängt. Sinnvollerweise müssen wir mehrere Alternativen unterscheiden, z.B.: keine Verhütung ($y = 0$), Pille ($y = 1$), Kondom ($y = 2$), andere Verhütungsmittel ($y = 3$). Das *multinomiale Logit-Modell* ist ein Regressionsmodell für abhängige Variablen mit mehreren Kategorien. Wie beim binären Logit-Modell formuliert man ein Wahrscheinlichkeitsmodell für die Alternativen $j = 0, 1, \dots, J$:

$$P(Y = j) = \frac{e^{\beta_j' \mathbf{x}}}{\sum_{k=0}^J e^{\beta_k' \mathbf{x}}}.$$

Es handelt sich um ein Wahrscheinlichkeitsmodell, weil $\sum_{j=0}^J P(Y = j) = 1$. Man beachte, daß jede Alternative einen eigenen Parametervektor β_j hat, weshalb eine Kovariate X_j unterschiedliche Effekte auf die Auswahlwahrscheinlichkeiten der Alternativen haben kann. Das so formulierte Modell ist allerdings nicht identifizierbar, denn man kann zu den Parametervektoren beliebige Zahlen addieren, ohne daß sich die Wahrscheinlichkeiten ändern. Deshalb wird meist die Normalisierung $\beta_0 = \mathbf{0}$ gewählt. Das Modell lautet dann

$$P(Y = j) = \frac{e^{\beta_j' \mathbf{x}}}{1 + \sum_{k=1}^J e^{\beta_k' \mathbf{x}}}, \quad \text{für } j = 1, 2, \dots, J$$

$$P(Y = 0) = \frac{1}{1 + \sum_{k=1}^J e^{\beta_k' \mathbf{x}}}.$$

Das binäre Modell ist offensichtlich ein Spezialfall dieses multinomialen Modells. Das multinomiale Modell kann auch in der Form eines Regressionsmodells auf latente Variablen formuliert werden. Dazu muß für jede Alternative eine latente Variable postuliert werden. Das Modell ist also ein Mehrgleichungsmodell (s. den nächsten Abschnitt).

Die Koeffizienten dieses Modells sind nicht leicht zu *interpretieren*. Kürzen wir $P(Y = j)$ mit P_j ab, so impliziert das Modell

$$\ln\left(\frac{P_j}{P_0}\right) = \beta_j' \mathbf{x}.$$

Ebenso wie im binären Modell sind die Parameter als Effekte auf die Log-Odds im Vergleich zum Null-Ereignis interpretierbar. Auch die Odds-Ratio Interpretation ist im multinomialen Modell möglich. Man beachte, daß jeder der J Parametervektoren auf das Null-Ereignis bezogen ist. Will man eine andere Bezugsbasis, so muß man die jeweiligen Parameterdifferenzen ausrechnen. Wegen der Vielzahl der Parameter dieses Modells empfiehlt Long (1987), die Odds-Effekte in graphischer Form zu präsentieren. Im Unterschied zum binären ist im multinomialen Fall allerdings das Vorzeichen der Odds-Effekte *nicht* unbedingt identisch mit den Vorzeichen der Marginaleffekte. Ein positiver Odds-Effekt einer Kovariaten kann mithin nicht als ein positiver Effekt auf die Wahrscheinlichkeit der betreffenden Alternative interpretiert werden (ein Punkt, der häufig übersehen wird!). Er kann nur als Odds-Effekt interpretiert werden und besagt, daß P_j gegenüber P_0 zunimmt, wenn die betreffende Kovariate ansteigt. Die Wahrscheinlichkeit P_j selbst muß aber nicht notwendigerweise ansteigen. Es kann nämlich passieren, daß ein Ansteigen der Kovariate die Wahrscheinlichkeit eines dritten Ereignisses P_k erhöht, P_j und P_0 aber sinken (nur daß P_j eben relativ weniger zurückgeht als P_0). Die Formel für die Marginaleffekte β_j^* auf die Wahrscheinlichkeiten im multinomialen Logit-Modell lautet

$$\frac{\partial P_j}{\partial \mathbf{x}} = P_j \left(\beta_j - \sum_{k=1}^J P_k \beta_k \right).$$

In β_j^* können durchaus einige Elemente ein anderes Vorzeichen haben als in β_j . Die Interpretation der Parameter des multinomialen Logit-Modells erfordert also besondere Sorgfalt.

Auch die üblichen *Tests* auf die Signifikanz der Odds-Effekte, wie sie die meisten Programme automatisch erstellen, sind in diesem Modell problematisch. Da jede Variable mehrere Effekte liefert, können die Einzeltests widersprüchlich ausfallen. Deshalb ist es besser, mittels eines LR-Tests zu überprüfen, ob die Variable insgesamt einen signifikanten Einfluß zeigt.

Die *Log-Likelihood* des multinomialen Logit-Modells ist eine Generalisierung derjenigen des binären Modells. Für jede Beobachtung benötigen wir $J + 1$ Indikatorvariablen d_{ij} , wobei $d_{ij} = 1$, falls die Alternative j gewählt wurde, und alle anderen Indikatorvariablen sind null. Dann kann die Log-Likelihood geschrieben werden als

$$\ln L = \sum_{i=1}^n \sum_{j=0}^J d_{ij} \ln P(Y_i = j).$$

Das multinomiale Logit-Modell beruht auf einer restriktiven Annahme. Betrachtet man obigen Ausdruck für die Log-Odds, so erkennt man, daß P_j/P_0 unabhängig von den anderen Alternativen ist. Man nennt dies die Annahme der *Unabhängigkeit von irrelevanten Alternativen* (englisch IIA). Diese Annahme impliziert, daß sich die Parameterschätzer nicht ändern, wenn wir die Alternativenmenge ändern. In unserem Beispiel mit den Verhütungsmitteln sollten sich mithin die Parameterschätzer nicht verändern, wenn wir die Alternative drei (andere Verhütungsmittel) aufgliedern und differenzierter im Modell berücksichtigen. In den meisten praktischen Anwendungen des multinomialen Logit-Modells wird dies nicht zutreffen. Ein Test der Annahme der Unabhängigkeit von irrelevanten Alternativen ist anhand dieser Implikation aber leicht möglich: Schätze zuerst das Modell mit allen Alternativen, dann ein zweites Modell, bei dem eine Alternative fehlt. Ein Hausman-Test gibt dann Auskunft darüber, ob sich die Parameterschätzer signifikant unterscheiden (ausführlich beschreiben solche Tests Zhang und Hoffman, 1993). Ist die Annahme der Unabhängigkeit von irrelevanten Alternativen verletzt, so kann man eventuell das multinomiale Probit-Modell verwenden, das diese Annahme nicht beinhaltet (s.u.). Eine weitere Möglichkeit ist das *verschachtelte Logit-Modell* (s. Ronning, 1991: 77ff). Dieses Modell geht nicht von einer Entscheidung zwischen gleichberechtigten Alternativen aus, sondern modelliert den Entscheidungsprozeß mehrstufig. In unserem Verhütungsmittel-Beispiel hieße das zum Beispiel, daß die Individuen auf der ersten Stufe eine Entscheidung treffen, ob sie ein Verhütungsmittel anwenden oder nicht und auf der zweiten Stufe gegebenenfalls eine Wahl zwischen den Verhütungsmitteln treffen. In diesem Modell müssen nur noch die Alternativen jeder Stufe unabhängig sein.

3.2.2 Exkurs: Entscheidungstheoretische Begründung

In diesem Abschnitt soll gezeigt werden, wie Logit- und Probit-Modelle entscheidungstheoretisch begründet werden können (s. ausführlicher Maier und Weiss, 1990). Eine solche Begründung hat den Vorteil, daß die Schätzgleichungen eventuell direkt aus einem theoretischen Modell folgen und nicht wie so häufig mehr oder weniger ad-hoc spezifiziert werden müssen.

Die Ableitung beginnt bei einem *Modell diskreter Wahl* („discrete choice model“). Von „diskreter“ Wahl spricht man, weil nicht Entscheidungen über eine stetige Gütermenge sondern über diskrete Alternativen getroffen werden. Jede dieser Alternativen hat bestimmte Charakteristika, die dem Individuum Nutzen stiften. Gemäß dem Rational-Choice Ansatz wird sich ein Akteur für die Alternative mit dem höchsten Nutzen entscheiden. Formalisiert wird die diskrete Wahl mittels des sogenannten Zufallsnutzenmodells. Der Nutzen einer Alternative j lautet:

$$U_j = V(\mathbf{c}_j, \mathbf{s}) + \epsilon_j,$$

wobei $V(\cdot)$ für die indirekte Nutzenfunktion steht (maximal erreichbarer Nutzen bei gegebenem \mathbf{c}_j und \mathbf{s}). \mathbf{c}_j ist ein Vektor alternativenspezifischer Charakteristika (der Preis, die Qualität, usw.) und \mathbf{s} ein Vektor sozio-ökonomischer Merkmale der Person. ϵ_j repräsentiert eine stochastische Nutzenkomponente, in der unbeobachtete Charakteristika und sozio-ökonomische Merkmale zusammengefaßt sind. Spezifiziert man nun eine Dichtefunktion $f(\boldsymbol{\epsilon})$, so ist damit auch $P(Y = j)$ berechenbar. Man beachte, daß es sich hier um ein stochastisches Modell handelt, denn es wird nicht mit Sicherheit die Alternative mit dem maximalen V gewählt.

Um von diesem theoretischen Modell zum Schätzmodell zu gelangen, sind Annahmen über V und $f(\boldsymbol{\epsilon})$ nötig. V wird gewöhnlich als eine lineare Nutzenfunktion spezifiziert: $V = \boldsymbol{\alpha}'\mathbf{c}_j + \boldsymbol{\beta}'_j\mathbf{s}$. $\boldsymbol{\alpha}$ ist der Vektor der „Nutzwengewichte“ der Charakteristika. Die Nutzwengewichte sind nicht absolut interpretierbar, weil sie von der Skalierung der Charakteristika abhängen. Ihre relativen Größen geben allerdings bei identischer Skalierung der Charakteristika Auskunft über deren Bedeutung im Entscheidungsprozeß. $\boldsymbol{\beta}_j$ repräsentiert den Einfluß der sozio-ökonomischen Merkmale auf den Nutzen der Alternative j . Dieser Einfluß variiert über die Alternativen. Ein identischer Einfluß auf jede Alternative wäre nicht identifizierbar. Man beachte die Symmetrie: Die Charakteristika variieren über die Alternativen, ihre Koeffizienten sind konstant; die individuellen Merkmale sind konstant, aber ihre Koeffizienten variieren über die Alternativen.

Zum Logit-Modell gelangt man, wenn man annimmt, daß die Fehlerterme unabhängig, identisch standard-extremwertverteilt sind. Dann ergibt sich nämlich der schon bekannte Ausdruck

$$P(Y = j) = \frac{e^{V(\mathbf{c}_j, \mathbf{s})}}{\sum_k e^{V(\mathbf{c}_k, \mathbf{s})}}.$$

Dieses Modell unterscheidet sich vom normalen Logit, indem es alternativenspezifische Charakteristika enthält. Man nennt es das *konditionale Logit-Modell*.³⁶ Ist $\boldsymbol{\alpha} = \mathbf{0}$, so hat man ein normales (multinomiales) Logit-Modell vorliegen.

Ein *multinomiales Probit-Modell* erhält man, wenn man annimmt, daß die Fehlerterme multivariat normalverteilt sind: $(\epsilon_0, \epsilon_1, \dots, \epsilon_J) \sim N(\mathbf{0}, \boldsymbol{\Sigma})$. Im Unterschied zu den Annahmen, die zum Logit-Modell führten, können in diesem Modell die Zufallskomponenten unterschiedliche Varianz haben und auch über die Alternativen hinweg korreliert sein. Dies impliziert insbesondere, daß die Annahme der Unabhängigkeit von irrelevanten Alternativen nicht getroffen werden muß. Das multinomiale Probit-Modell ist somit eine Lösung, wenn Abhängigkeit zwischen den Alternativen besteht. Allerdings ist in diesem Modell die Bestimmung von $P(Y = j)$ mit der Berechnung von mehrdimensionalen Integralen verbunden,

³⁶Anwendungen des konditionalen Logit-Modells findet man z.B. bei Hoffman und Duncan (1988) sowie bei Brüderl und Preisendörfer (1995).

weshalb das multinomiale Probit-Modell aus den selben Gründen wie das multivariate Probit-Modell nur mit den in den letzten Jahren entwickelten Simulations-Verfahren schätzbar ist.

3.2.3 Das ordinale Probit-Modell

Ordinale Daten fallen häufig an, wenn man versucht Einstellungen zu messen. Dazu legt man dem Befragten eine Aussage vor, von der man vermutet, daß sie ein Indikator für die latente Einstellung ist. Der Befragte soll auf einer z.B. fünf-stufigen Skala seine Zustimmung zu dieser Aussage angeben. $y = 0$ bedeutet etwa „stimme überhaupt nicht zu“, während mit $y = 4$ das Gegenteil „stimme voll zu“ verkodet wird. Das implizite Meßmodell dieses Vorgehens ist, daß die Befragten sich in die Kategorie einordnen, die ihrer latenten Einstellung am nächsten ist. Man kann dieses Modell ebenfalls als *Schwellenwertmodell* auffassen:

$$y^* = \beta' \mathbf{x} + \epsilon,$$

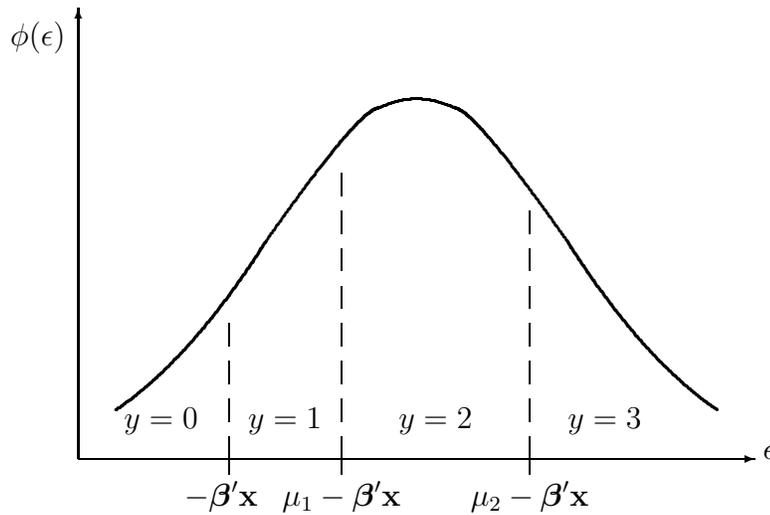
wobei Y^* die latente Einstellung ist. Was wir beobachten ist jedoch

$$\begin{aligned} y &= 0, & \text{wenn } y^* \leq \mu_0, \\ y &= 1, & \text{wenn } \mu_0 < y^* \leq \mu_1, \\ y &= 2, & \text{wenn } \mu_1 < y^* \leq \mu_2, \\ & \vdots \\ y &= J, & \text{wenn } \mu_{J-1} < y^*. \end{aligned}$$

Ordinale Daten sind also nach diesem Modell gruppierte metrische Daten, wobei allerdings die Schwellen μ_j unbekannt sind. Sie müssen zusammen mit β anhand der Daten geschätzt werden.³⁷ Man beachte, daß in diesem Modell im Gegensatz zum multinomialen Modell nur *ein* Parametervektor β spezifiziert ist. Insofern ist ein ordinale Modell eine restringierte Version eines multinomialen Modells, weshalb mittels eines Likelihood-Ratio Tests die Signifikanz dieser Restriktion getestet werden kann. Die Modellkonstante ist nur identifiziert, wenn eine Schwelle normiert ist, weshalb man meist $\mu_0 = 0$ setzt. Häufig verwenden Forscher zur Schätzung eines Modells mit ordinaler abhängiger Variable OLS. Damit handelt man sich die selben Probleme ein wie im linearen Wahrscheinlichkeitsmodell (s. Winship und Mare, 1984). Zusätzlich trifft man bei der Verwendung von OLS die restriktive Annahme, daß die Abstände zwischen den Kategorien identisch gleich eins sind. Weitere Probleme der OLS-Schätzung bei ordinaler abhängiger Variable führen Clogg und Shihadeh (1994: 142f) an.

³⁷Hat man metrische Daten gruppiert erhoben (wie das häufig bei Einkommensdaten der Fall ist), so paßt ebenfalls dieses Modell, nur daß dann die Schwellen μ_j bekannt sind.

Abbildung 13: Das ordinale Probit-Modell



Nehmen wir an, daß $\epsilon \sim N(0, 1)$, so erhalten wir das *ordinale Probit-Modell* (ordinale Logit-Modelle lassen sich ebenfalls formulieren; s. Ludwig-Mayerhofer, 1990; Clogg und Shihadeh, 1994: Kap. 7):

$$\begin{aligned}
 P(Y = 0) &= \Phi(-\beta' \mathbf{x}), \\
 P(Y = 1) &= \Phi(\mu_1 - \beta' \mathbf{x}) - \Phi(-\beta' \mathbf{x}), \\
 P(Y = 2) &= \Phi(\mu_2 - \beta' \mathbf{x}) - \Phi(\mu_1 - \beta' \mathbf{x}), \\
 &\vdots \\
 P(Y = J) &= 1 - \Phi(\mu_{J-1} - \beta' \mathbf{x}).
 \end{aligned}$$

Für $J = 3$ ist dieses Modell in Abbildung (13) dargestellt. Die Wahrscheinlichkeitsdichten entsprechen jeweils den Flächen zwischen den Gruppengrenzen. Für $J = 1$ ergibt sich als Spezialfall das binäre Probit-Modell. Die Schätzung der Parameter erfolgt mittels Maximum-Likelihood (die Likelihood kann analog zum multinomialen Logit-Modell notiert werden).³⁸

Die β sind als Effekte auf die latente Variable interpretierbar. Allerdings ist diese Variable wiederum beliebig skalierbar, weshalb nur die Vorzeichen interpretierbar sind. Die *Interpretation* anhand der beobachteten ordinalen Variable ist jedoch recht komplex. Angenommen β_j sei positiv. Wie verändern sich dann die Wahrscheinlichkeiten der einzelnen Kategorien, wenn X_j zunimmt? Ein Blick auf Abbildung (13) macht deutlich, daß sich dann alle Schwellen nach links verschieben.

³⁸Auch das ordinale Probit kann mit einer Probit-Auswahlgleichung kombiniert werden. Dies erlaubt die einfache Modellierung von Selektivität im Fall ordinaler abhängiger Variablen. Ein solches Modell ist in LIMDEP implementiert (s. Greene, 1995: Kap. 23). Eine Anwendung findet man bei Liao (1995).

Mithin sinkt $P(Y = 0)$ und $P(Y = J)$ steigt an. Wie sich aber die dazwischenliegenden Wahrscheinlichkeiten verändern, ist unklar. Dies hat zur Folge, daß die Vorzeichen der Marginaleffekte auf die Wahrscheinlichkeiten nur für die beiden Randkategorien sofort gegeben sind, während sie für die anderen Kategorien nur durch zusätzliche Rechnungen bestimmbar sind. Will man beim ordinalen Probit die Marginaleffekte für alle Kategorien ausweisen, so müssen diese jeweils explizit errechnet werden. Die Formel für die Marginaleffekte lautet:

$$\frac{\partial P(Y = j)}{\partial x_j} = (\phi(\mu_{j-1} - \beta' \mathbf{x}) - \phi(\mu_j - \beta' \mathbf{x}))\beta_j.$$

Bei Dummy-Variablen empfiehlt es sich, die Wahrscheinlichkeiten P_j für beide Gruppen direkt auszurechnen und gegenüberzustellen.

3.2.4 Anwendungsbeispiel: Die Einstellung zur Ehe

Im obigen Anwendungsbeispiel haben wir argumentiert, daß unbeobachtete Faktoren wie die Einstellung zur Institution Ehe sowohl die Wahrscheinlichkeit einer nichtehelichen Lebensgemeinschaft wie auch die einer Scheidung beeinflussen. Der DJI-Familiensurvey enthält auch einige Items, die die Einstellung zur Ehe messen sollen. Wir hätten diese Items somit in unseren obigen Analysen als weitere Kontrollvariablen einsetzen können. Dies wurde jedoch unterlassen, weil bei solchen Einstellungsmessungen mit Querschnittsdaten immer das Problem umgekehrter Kausalität auftritt. So ist es in unserem Beispiel wahrscheinlich, daß Personen, die eine nichteheliche Lebensgemeinschaft oder eine Scheidung hinter sich haben, aufgrund dieser Erfahrung eine geringere Wertschätzung der Institution Ehe zeigen.

Ungeachtet dieser Probleme soll in diesem Unterabschnitt eines dieser Items zur Illustration des ordinalen Probit-Modells eingesetzt werden. Im Familiensurvey wurde nach der Bedeutung der Ehe gefragt. Die Befragten konnten ihre Zustimmung zu der Aussage „Eine Ehe bedeutet Sicherheit und Geborgenheit“ auf einer vierstufigen Skala ausdrücken. Wir rekodierten diese Skala so, daß null bedeutet „stimme überhaupt nicht zu“ und drei „stimme voll und ganz zu“. Die Stichprobenabgrenzung und die unabhängigen Variablen sind wie oben definiert („Katholik“ und „Intakte Familie“ zeigten keine signifikanten Effekte und wurden deshalb der Übersichtlichkeit wegen unterdrückt).

Die Schätzergebnisse findet man in Tabelle (4). Das Modell wurde mit LIMDEP geschätzt (s. Greene, 1995: Kap. 23). Das Pseudo- R^2 fällt sehr niedrig aus, was bei Einstellungsregressionen häufig der Fall ist. In der ersten Spalte sind die Probit-Effekte angeführt. Wie oben erwähnt, ist nur deren Vorzeichen sinnvoll zu interpretieren. Wir sehen, daß eine Lebensgemeinschaft die Wertschätzung der Ehe vermindert. In welche Richtung hier die Kausalität geht, ist allerdings

Tabelle 4: Ein ordinales Probit-Modell für die Einstellung zur Ehe (N=4444)

Variable	<u>Probit-Effekte</u>	<u>Marginaleffekte am Mittelwert</u>			
	Zust. zur Ehe	P(Y=0)	P(Y=1)	P(Y=2)	P(Y=3)
Konstante	3,13*	0,03	0,08	0,32	0,57
Lebensgem. (1=Ja)	-0,16* (2,89)	0,01	0,02	0,03	-0,06
Heiratskohorte (1=1961-70)	-0,20* (4,19)	0,01	0,02	0,04	-0,08
Heiratskohorte (1=1971-77)	-0,23* (4,49)	0,01	0,03	0,05	-0,09
Kleinstadt (1=bis 4999)	0,26* (4,57)	-0,02	-0,03	-0,06	0,11
Mittelstadt (1=5000-99999)	0,19* (4,54)	-0,01	-0,02	-0,04	0,07
Aktiver Gläubiger (1=Ja)	0,26* (5,68)	-0,02	-0,03	-0,06	0,11
Schulbildung (in Jahren)	-0,12* (8,63)	0,01	0,01	0,03	-0,05
μ_1	0,66				
μ_2	1,74				
R_{MF}^2	0,03				
-Log-Likelihood	4287,8				

* signifikant auf dem 5%-Niveau. Absolute t-Werte in Klammern. McFadden Pseudo- R^2 Werte in bezug auf das Modell mit nur einer Konstanten. Abhängige Variable ist die Zustimmung zu dem Item „Ehe bedeutet Sicherheit und Geborgenheit“, wobei null bedeutet „stimme überhaupt nicht zu“ und drei „stimme voll und ganz zu“. In der ersten Zeile unter „Marginaleffekte“ sind die am Mittelwert der Kovariaten geschätzten relativen Häufigkeiten der Kategorien angeführt. Bezugsgruppe ist Heiratskohorte 1949-60, Großstadt mit über 100.000 Einwohnern, weniger als einmal pro Monat Kirchenbesuch.

Quelle: Eigene Berechnungen mit Daten des DJI-Familiensurvey West.

unklar. Ansonsten erkennen wir, daß die Wertschätzung der Ehe mit der Kohorte und mit der Größe des Wohnortes sinkt. Aktive Gläubige schätzen die Ehe mehr und höher Gebildete weniger. Man beachte, daß die Richtung der Effekte genau entgegengesetzt zu den weiter oben berichteten Scheidungseffekten ist.

Die anderen vier Spalten der Tabelle enthalten die Marginaleffekte auf die vier

Wahrscheinlichkeiten am Mittelwert der Kovariaten. In der ersten Zeile sind die am Mittelwert der Kovariaten vom Modell vorhergesagten relativen Häufigkeiten der Kategorien berichtet, damit man die Größenordnung der Marginaleffekte besser einschätzen kann. Die Marginaleffekte jeder Variable summieren sich zu null, denn die Summe der Wahrscheinlichkeiten muß immer eins sein. Man erkennt, daß negative Probit-Effekte die Wahrscheinlichkeit von „stimme voll zu“ reduzieren, die anderen Wahrscheinlichkeiten aber erhöhen. Umgekehrt verhält es sich mit positiven Probit-Effekten. Betrachten wir die Marginaleffekte einer Lebensgemeinschaft, so sehen wir, daß Personen, die eine nichteheliche Lebensgemeinschaft eingingen, um sechs Prozentpunkte seltener die Kategorie „stimme voll zu“ wählen, während sie entsprechend häufiger die anderen Kategorien angeben. Die Einheitseffekte muß man sich ausrechnen: Gehen wir vom Mittelwert der Kovariaten aus, so errechnen sich folgende Wahrscheinlichkeiten für die vier Kategorien: 0,03, 0,08, 0,32 und 0,57. Hätte diese Person aber zusätzlich eine nichteheliche Lebensgemeinschaft hinter sich, so würden die Wahrscheinlichkeiten lauten: 0,04, 0,09, 0,36 und 0,51. Die Differenzen der Wahrscheinlichkeiten sind bei diesem Beispiel beinahe identisch mit den Marginaleffekten.

Da unsere abhängige Variable nur vier Ausprägungen hat, kann man auch ein multinomiales Logit-Modell verwenden. Dabei schätzt man 21 Regressionskoeffizienten anstatt von sieben. Das ordinale Modell vereinfacht also, indem es davon ausgeht, daß die Effekte der Kovariaten für jede Kategorie der abhängigen Variable identisch sind. Diese Annahme kann man mittels eines Likelihood-Ratio Tests überprüfen (Vergleich von multinomialen und ordinalem Logit). Die χ^2 -Testgröße beträgt in unserem Anwendungsfall 27,9, was bei 14 Freiheitsgraden auf dem 5%-Niveau signifikant ist. Mithin ist das multinomiale Logit-Modell vorzuziehen. Betrachtet man allerdings die Marginaleffekte des multinomialen Logit-Modells, so ergeben sich nur geringfügige Abweichungen zu den in Tabelle (4) berichteten Effekten.³⁹

³⁹In manchen Arbeiten wird argumentiert, daß ordinale Modelle „sparsamer“ seien, weil weniger Koeffizienten geschätzt werden. Dies ist allerdings nur vordergründig so. Berichtet man zwecks anschaulicherer Interpretation die Marginaleffekte, so ist deren Zahl bei multinomialen und ordinalen Modellen gleich.

Literatur

- Achen, C.H. und W.P. Shively (1995) *Cross-Level Inference*. Chicago: University of Chicago Press.
- Agresti, A. (1990) *Categorical Data Analysis*. New York: Wiley.
- Andreas, H.J. (1989) Recurrent Unemployment: The West German Experience. *European Sociological Review* 5: 275-297.
- Amemiya, T. (1985) *Advanced Econometrics*. Cambridge: Harvard University Press.
- Aptech (1992) *The GAUSS System Version 3.0*. Maple Valley: Aptech Systems.
- Arminger, G. (1995) Specification and Estimation on Mean Structures: Regression Models. S. 77-183, in: Arminger, G.; Clogg, C. und M.E. Sobel (Hrsg.) *Handbook of Statistical Modeling for the Social and Behavioral Sciences*. New York: Plenum.
- Asher, H. (1983²) *Causal Modeling*. Newbury Park: Sage.
- Axinn, W. und A. Thornton (1992) The Relationship Between Cohabitation and Divorce: Selectivity or Causal Influence? *Demography* 29: 357-374.
- Backhaus, K.; Erichson, B.; Plinke, W. und R. Weiber (1994) *Multivariate Analysemethoden: Eine anwendungsorientierte Einführung*. Berlin: Springer.
- Barron, D.N. (1992) The Analysis of Count Data: Overdispersion and Autocorrelation. *Sociological Methodology* 22: 179-220.
- Berk, R. (1983) An Introduction to Sample Selection Bias in Sociological Data. *American Sociological Review* 48: 386-398.
- Berk, R. (1990) A Primer on Robust Regression. S. 292-324, in: Fox, J. und J.S. Long (Hrsg.) *Modern Methods of Data Analysis*. Newbury Park: Sage.
- Berry, W.D. (1984) *Nonrecursive Causal Models*. Newbury Park: Sage.
- Blau, P.M. und O.D. Duncan (1967) *The American Occupational Structure*. New York: Free Press.
- Bollen, K.A. (1989) *Structural Equations With Latent Variables*. New York: Wiley.
- Bollen, K.A. und R. Jackman (1990) Regression Diagnostics: An Expository Treatment of Outliers and Influential Cases. S. 257-291, in: Fox, J. und J.S. Long (Hrsg.) *Modern Methods of Data Analysis*. Newbury Park: Sage.
- Bollen, K.A.; Guilkey, D.K. und T.A. Mroz (1995) Binary Outcomes and Endogenous Explanatory Variables. *Demography* 32: 111-131.
- Breen, R. (1996) *Regression Models: Censored, Sample Selected, or Truncated Data*. Thousand Oaks: Sage.
- Brüderl, J. und P. Preisendörfer (1995) Der Weg zum Arbeitsplatz: Eine empirische Untersuchung zur Verkehrsmittelwahl. S. 69-88, in: Diekmann, A. und A. Franzen (Hrsg.) *Kooperatives Umwelthandeln*. Zürich: Rüegger.
- Brüderl, J.; Preisendörfer, P. und R. Ziegler (1996) *Der Erfolg neugegründeter Betriebe*. Berlin: Duncker & Humblot.

- Brüderl, J.; Diekmann, A. und H. Engelhardt (1997) Drum prüfe, wer sich ewig bindet! Eine empirische Untersuchung zum Zusammenhang zwischen vorehelicher Lebensgemeinschaft und Scheidungsrisiko. *Kölner Zeitschrift für Soziologie und Sozialpsychologie* (im Erscheinen).
- Chatterjee, S. und A.S. Hadi (1988) *Sensitivity Analysis in Linear Regression*. New York: Wiley.
- Chow, G. (1960) Tests of Equality Between Sets of Coefficients in Two Linear Regressions. *Econometrica* 28: 591-605.
- Clogg, C.C.; Petkova, E. und A. Hariton (1995) Statistical Methods for Comparing Regression Coefficients Between Models. *American Journal of Sociology* 100: 1261-1293.
- Clogg, C.C. und E.S. Shihadeh (1994) *Statistical Models for Ordinal Data*. Thousand Oaks: Sage.
- Cook, D.R. und S. Weisberg (1994) *An Introduction to Regression Graphics*. New York: Wiley.
- Deal, J.E. und E.R. Anderson (1995) Reporting and Interpreting Results in Family Research. *Journal of Marriage and the Family* 57: 1040-1048.
- DeVeaux, R. (1990) Finding Transformations for Regression: Using the ACE Algorithm. S. 177-208, in: Fox, J. und J.S. Long (Hrsg.) *Modern Methods of Data Analysis*. Newbury Park: Sage.
- Diekmann, A. (1992) Sex-Ratio, Divorce, and Labor Force Participation: An Analysis of International Aggregate Data. S. 283-293, in: Haag, G.; Müller, U. und K. Troitzsch (Hrsg.) *Economic Evolution and Demographic Change*. Berlin: Springer.
- Dillon, W. und M. Goldstein (1984) *Multivariate Analysis*. New York: Wiley.
- DiPrete, T.A. und J.D. Forristal (1994) Multilevel Models: Methods and Substance. *Annual Review of Sociology* 20: 331-357.
- Dubin, J. und D. Rivers (1989) Selection Bias in Linear Regression, Logit, and Probit Models. *Sociological Methods & Research* 18: 360-390.
- Eberstein, I.; Nam, C. und R. Hummer (1990) Infant Mortality by Cause of Death. *Demography* 27: 413-430.
- Fahrmeir, L. und A. Hamerle (1984) (Hrsg.) *Multivariate statistische Verfahren*. Berlin: De Gruyter.
- Fahrmeir, L. und G. Tutz (1994) *Multivariate Statistical Modelling Based on Generalized Linear Models*. New York: Springer.
- Flewelling, R. und K. Baumann (1990) Family Structure as a Predictor of Initial Substance Use and Sexual Intercourse in Early Adolescence. *Journal of Marriage and the Family* 52: 171-181.
- Fox, J. (1991) *Regression Diagnostics*. Newbury Park: Sage.
- Goodall, C. (1990) A Survey of Smoothing Techniques. S. 126-176, in: Fox, J. und J.S. Long (Hrsg.) *Modern Methods of Data Analysis*. Newbury Park: Sage.
- Greene, W. (1993²) *Econometric Analysis*. New York: MacMillan.
- Greene, W. (1995) *LIMDEP: Version 7.0 User's Manual*. Bellport: Econometric Software.
- Guttentag, M. und P. Secord (1983) *Too Many Women: The Sex-Ratio Question*. Beverly Hills: Sage.

- Hannan, M. und G. Carroll (1992) *Dynamics of Organizational Populations*. New York: Oxford University Press.
- Härdle, W. (1990) *Applied Nonparametric Regression*. Cambridge: Cambridge University Press.
- Hardy, M. (1993) *Regression With Dummy Variables*. Newbury Park: Sage.
- Haveman, R. und B. Wolfe (1994) *Succeeding Generations: On the Effects of Investments in Children*. New York: Russell Sage.
- Heckman, J. (1979) Sample Selection Bias as a Specification Error. *Econometrica* 47: 153-161.
- Heckman, J. und R. Robb (1985) Alternative Methods for Evaluating the Impact of Interventions. S. 156-245, in: Heckman, J. und B. Singer (Hrsg.) *Longitudinal Analysis of Labor Market Data*. Cambridge: Cambridge University Press.
- Heckman, J. und J. Walker (1990) The Relationship Between Wages and Income and the Timing and Spacing of Births. *Econometrica* 58: 1411-1441.
- Hoffman, S. und G. Duncan (1988) Multinomial and Conditional Logit Discrete-Choice Models in Demography. *Demography* 25: 415-427.
- Hosmer, D. und S. Lemeshow (1989) *Applied Logistic Regression*. New York: Wiley.
- Hox, J.J. und I.G. Kreft (1994) Multilevel Analysis Methods. *Sociological Methods & Research* 22: 283-299.
- Hoyle, R.H. (1995)(Hrsg.) *Structural Equation Modeling: Concepts, Issues, and Applications*. Thousand Oaks: Sage.
- Infratest (1989) *DJI-Familiensurvey: Methodenbericht*. München: Infratest Sozialforschung.
- Jaccard, J.; Turrisi, R. und C. Wan (1990) Interaction Effects in Multiple Regression. Newbury Park: Sage.
- Johnston, J. (1984) *Econometric Methods*. New York: McGraw-Hill.
- Judge, G.; Griffiths, W.E.; Carter-Hill, R.; Lütkepohl, H. und T. Lee (1985²) *The Theory and Practice of Econometrics*. New York: Wiley.
- Kahn, J.R. und K. London (1991) Premarital Sex and the Risk of Divorce. *Journal of Marriage and the Family* 53: 845-855.
- Kahn, J.R.; Rindfuss, R. und D. Guilkey (1990) Adolescent Contraceptive Method Choices. *Demography* 27: 323-335.
- Kahn, J.R. und J.R. Udry (1986) Marital Coital Frequency: Unnoticed Outliers and Unspecified Interactions Lead to Erroneous Conclusions. *American Sociological Review* 51: 734-737.
- Kaufman, R.L. (1996) Comparing Effects in Dichotomous Logistic Regression: A Variety of Standardized Coefficients. *Social Science Quarterly* 77: 90-109.
- King, G. (1992) *Event Count and Duration Regression*. Maple Valley: Aptech Systems.
- Krämer, W. und H. Sonnberger (1986) *The Linear Regression Model under Test*. Heidelberg: Physica.
- Land, K.C.; McCall, P.L. und D.S. Nagin (1996) A Comparison of Poisson, Negative Binomial, and Semiparametric Mixed Poisson Regression Models. *Sociological Methods & Research* 24: 387-442.

- Liao, T.F. (1994) *Interpreting Probability Models: Logit, Probit, and Other Generalized Linear Models*. Thousand Oaks: Sage.
- Liao, T.F. (1995) The Nonrandom Selection of Don't Knows in Binary and Ordinal Responses: Corrections With the Bivariate Probit Model With Sample Selection. *Quality & Quantity* 29: 87-110.
- Lillard, L.A.; Brien, M.J. und L.J. Waite (1995) Premarital Cohabitation and Subsequent Marital Dissolution: A Matter of Self-Selection? *Demography* 32: 437-457.
- Little, R. und D. Rubin (1987) *Statistical Analysis with Missing Data*. New York: Wiley.
- Long, J. (1987) A Graphical Method for the Interpretation of Multinomial Logit Analysis. *Sociological Methods & Research* 15: 420-446.
- Long, J. und P. Trivedi (1992) Some Specification Tests for the Linear Regression Model. *Sociological Methods & Research*, 21: 161-204.
- Ludwig-Mayerhofer, W. (1990) Multivariate Logit-Modelle für ordinalskalierte abhängige Variablen. *ZA-Information* 27: 62-88.
- Maddala, G.S. (1983) *Limited-Dependent and Qualitative Variables in Econometrics*. Cambridge: Cambridge University Press.
- Magee, L. (1990) R^2 Measures Based on Wald and Likelihood Ratio Joint Significance Tests. *American Statistician* 44: 250-253.
- Maier, G. und P. Weiss (1990) *Modelle diskreter Entscheidungen*. Wien: Springer.
- McCloskey, D.N. und S.T. Ziliak (1996) The Standard Error of Regressions. *Journal of Economic Literature* 34: 97-114.
- McCullagh, P. und J.A. Nelder (1989) *Generalized Linear Models*. London: Chapman und Hall.
- Menard, S. (1995) *Applied Logistic Regression Analysis*. Thousand Oaks: Sage.
- Monfort, A. und H. van Dijk (1995) Simulation-Based Econometrics. S. 1-20, in: H. van Dijk, A. Monfort und B. Brown (Hrsg.) *Econometric Inference Using Simulation Techniques*. Chichester: Wiley.
- Morgan, S. und J. Teachman (1988) Logistic Regression: Description, Examples, and Comparisons. *Journal of Marriage and the Family* 50: 929-936.
- Morrison, D.F. (1990) *Multivariate Statistical Methods*. New York: McGraw-Hill.
- Njogu, W. (1991) Trends and Determinants of Contraceptive Use in Kenya. *Demography* 28: 83-99.
- Norusis, M. (1993) *SPSS for Windows: Base System User's Guide, Release 6.0*. Chicago: SPSS Inc.
- Petersen, T. (1985) A Comment on Presenting Results from Logit and Probit Models. *American Sociological Review* 50: 130-131.
- Powers, D. (1993) Alternative Models of the Effects of Family Structure on Early Family Formation. *Social Science Research* 22: 283-299.
- Pregibon, D. (1981) Logistic Regression Diagnostics. *The Annals of Statistics* 9: 705-724.
- Raftery, A. (1986) A Note on Bayes Factors for Log-Linear Contingency Table Models with Vague Prior Information. *Journal of the Royal Statistical Society, B* 48: 249-250.

- Raftery, A. (1995) Bayesian Model Selection in Social Research. *Sociological Methodology* 25: 111-163.
- Roncek, D. (1991) Using Logit Coefficients to Obtain the Effects of Independent Variables on Changes in Probabilities. *Social Forces* 70: 509-518.
- Ronning, G. (1991) *Mikroökonomie*. Berlin: Springer.
- Rousseeuw, P. und A. Leroy (1987) *Robust Regression and Outlier Detection*. New York: Wiley.
- Schnell, R. (1994) *Graphisch gestützte Datenanalyse*. München: Oldenbourg.
- Seber, G. und C. Wild (1989) *Nonlinear Regression*. New York: Wiley.
- Sobel, M.E. (1995) Causal Inference in the Social and Behavioral Sciences. S. 1-38, in: Armingier, G.; Clogg, C. und M.E. Sobel (Hrsg.) *Handbook of Statistical Modeling for the Social and Behavioral Sciences*. New York: Plenum.
- Stolzenberg, R. und D. Relles (1990) Theory Testing in a World of Constrained Research Design. *Sociological Methods & Research* 18: 395-415.
- Tabachnick, B. und L. Fidell (1989) *Using Multivariate Statistics*. New York: Harper Collins.
- Thomson, E. und U. Colella (1992) Cohabitation and Marital Stability: Quality or Commitment? *Journal of Marriage and the Family* 54: 259-267.
- Tobin, J. (1958) Estimation of Relationships for Limited Dependent Variables. *Econometrica* 26: 24-36.
- Veall, M. und K.F. Zimmermann (1992) Performance Measures From Prediction-Realization Tables. *Economics Letters* 39: 129-134.
- Veall, M. und K.F. Zimmermann (1994) Evaluating Pseudo-R²'s For Binary Probit Models. *Quality & Quantity* 28: 151-164.
- Weisberg, S. (1985) *Applied Linear Regression*. New York: Wiley.
- White, H. (1980) A Heteroscedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroscedasticity. *Econometrica* 48: 817-838.
- Winkelmann, R. (1994) *Count Data Models*. Berlin: Springer.
- Winkelmann, R. und K.F. Zimmermann (1994) Count Data Models for Demographic Data. *Mathematical Population Studies* 4: 205-221.
- Winship, C. und R.D. Mare (1983) Structural Equations and Path Analysis for Discrete Data. *American Journal of Sociology* 89: 54-110.
- Winship, C. und R.D. Mare (1984) Regression Models with Ordinal Variables. *American Sociological Review* 49: 512-525.
- Winship, C. und R.D. Mare (1992) Models for Sample Selection Bias. *Annual Review of Sociology* 18: 327-350.
- Zentralarchiv (1991) ALLBUS: Basisumfrage 1991 in Gesamtdeutschland. Köln: Zentralarchiv für empirische Sozialforschung.
- Zhang, J. und S. Hoffman (1993) Discrete-Choice Logit Models: Testing the IIA Property. *Sociological Methods & Research* 22: 193-213.