Geman Microdata Lab

gesis

LEM-Beispielprogramme für Analysen zur Stichprobenselektivität des Mikrozensuspanels

Mikrodaten-Tools

Mai 2009

Dieses Papier ist ein Auszug aus der Veröffentlichung

> Schimpl-Neimanns, Bernhard: Bildungsverläufe und Stichprobenselektivität – Analysen zur Stichprobenselektivität des Mikrozensuspanels 1996 - 1999 am Beispiel bildungsstatistischer Fragestellungen. GESIS-Forschungsberichte: Reihe Sozialwissenschaftliche Datenanalyse; Bd. 1. Bonn: GESIS, 2008.

Das Kapitel beschreibt die methodischen Grundlagen für die Aufdeckung und Behandlung selektiver Ausfälle. Im ersten Abschnitt wird die Klassifikation verschiedener Ausfalltypen vorgestellt. Nach einem Überblick zum möglichen Umgang mit Ausfällen bei kategorialen Daten im zweiten Abschnitt konzentrieren sich die Abschnitte drei und vier auf statistische Modelle für diesen Datentyp und erläutern die Modelle anhand einfacher Beispiele. Die mit dem Programm LEM geschätzten Modelle sind im Anhang dokumentiert und im WWW zugänglich. Das Handbuch zum Mikrozensuspanel 1996-1999 enthält weitere LEM-Programme.

Die Konstruktion der in den Beispielen verwendeten Daten des Mikrozensuspanels 1996-1999 und der Beschäftigtenstichprobe (IABS-R01) wird in der o. g. Veröffentlichung in Abschnitt 6.1 (S. 113-120) und im ZUMA-Arbeitsbericht 2006/02 beschrieben.*

Bernhard Schimpl-Neimanns, Mai 2009

---

* Die Datenbasis des Mikrozensuspanels bildet das Arbeitsfile des Methodenverbundprojektes und ist eine faktisch anonymisierte 60-Prozent-Substichprobe von Auswahlbezirken des 1996 bis 1999 befragten Rotationsviertels. Das nach Abschluss des Projektes verfügbare Scientific Use File unterscheidet sich vom Arbeitsfile u. a. durch den höheren Auswahlsatz von 70 Prozent.

Die Grundlage der hier kurz als Beschäftigtenstichprobe bezeichneten Daten ist die faktisch anonymisierte IAB-Regionalstichprobe 1975-2001 (IABS-R01). Dabei handelt es sich um eine Zwei-Prozent-Stichprobe aller sozialversicherungspflichtig Beschäftigten ergänzt um Zeiten des Leistungsbezugs. Grundlage der Stichprobenziehung ist die Beschäftigten-Leistungsempfänger-Historik (BLH) des Instituts für Arbeitsmarkt- und Berufsforschung (IAB) der Bundesagentur für Arbeit. Die IABS-R01 kann über das GESIS Datenarchiv bezogen werden. Für die Verwendung der Daten in diesem Beitrag trägt das IAB keine Verantwortung.

Auszug aus: Schimpl-Neimanns, Bernhard: Bildungsverläufe und Stichprobenselektivität - Analysen zur Stichprobenselektivität des Mikrozensuspanels 1996 - 1999 am Beispiel bildungsstatistischer Fragestellungen. GESIS-Forschungsberichte: Reihe Sozialwissenschaftliche Datenanalyse; Bd. 1. Bonn: GESIS, 2008: Kapitel 3 (S. 25-42) und Anhang (S. 177-180).

3 Methodische Grundlagen

Ziel dieses Kapitels ist es, die methodischen Grundlagen für den Umgang mit dem Kernproblem der durch die aus dem Haushalt des Auswahlbezirks wegziehenden Personen entstehenden Ausfälle im Mikrozensuspanel darzustellen. Das einfachste Vorgehen könnte darin bestehen, bei Längsschnittanalysen nur die Daten von räumlich Immobilen zu verwenden. Wegen des geringeren Stichprobenumfangs führt dies bei Populationsschätzungen, wie z. B. der Zahl oder dem Anteil von Ausbildungsabsolventen, zu einem höheren Stichprobenfehler bzw. bei der Schätzung von Regressionsmodellen zu einem höheren Standardfehler der Koeffizienten. Gegebenenfalls unterscheidet sich aber die Umzugs- bzw. Ausfallwahrscheinlichkeit für bestimmte Teilgruppen. Wenn beispielsweise die Wahrscheinlichkeit eines Umzugs für Personen nach Erreichen des Ausbildungsabschlusses höher ist als für Personen, die sich noch in der Ausbildung befinden, oder wenn die Umzugswahrscheinlichkeit mit der zeitlichen Nähe zum Abschlusszeitpunkt korreliert, hängen die Ausfälle systematisch mit dem interessierenden Ereignis zusammen. Würde man in diesem Fall räumlich mobile Personen vernachlässigen, wäre die resultierende Stichprobe der räumlich Immobilen nicht mehr als einfache Zufallsauswahl der interessierenden Gesamtheit zu betrachten. Die daraus entstehenden Verzerrungen können gravierender sein als der eingangs genannte Präzisionsverlust, der durch die bloße Verringerung der Stichprobe entsteht.

Das Ausmaß der möglichen Verzerrung hängt allgemein vom Anteil der ausgefallenen Personen innerhalb der betrachteten Gruppe und der Kovariation der interessierenden Variablen mit dem Ausfall ab. Es ist deshalb vordringlich notwendig, diese Zusammenhänge zwischen dem Ausfall und den Analysevariablen genauer auf ihre Implikationen hinsichtlich möglicher Verzerrungen zu untersuchen. Dies geschieht im folgenden ersten Abschnitt. Die darauf aufbauende Darstellung im zweiten Abschnitt konzentriert sich auf die Schätzung statistischer Modelle. Anschließend wird anhand einer einfachen Beispieltabelle beschrieben, mit welchen Verfahren für die im Mikrozensus hauptsächlich vorliegenden qualitativen Variablen bzw. kategorialen Daten Ausfalluntersuchungen und modell-

basierte Ausfallkorrekturen vorgenommen werden können. Abschließend wird gezeigt, wie mit diesen Modellen auch Referenzdaten zur Validierung genutzt werden können.

3.1 Klassifikation der Ausfallprozesse

Die Klassifikation der Ausfallprozesse folgt der von Rubin (1976) entwickelten Systematik, mit der die Implikationen fehlender Werte bei inferenzstatistischen Schlüssen dargestellt wurden.

Bei fehlenden Werten in einer oder mehreren Variablen liegen Datenkonstellationen vor, die mittels Indikatorvariablen unterschieden werden können. Für Analyseeinheiten ohne fehlende Werte stehen vollständige Angaben zur interessierenden Variablen Y (z. B. Schulbesuch oder Ausbildungsabschluss) sowie zu einer oder mehreren Kovariaten X zur Verfügung. Die Indikatorvariable R (Response) zeigt für diese Gruppe den Wert Eins ($R = 1$) an.[6] Für die Gruppe ausgefallener Personen mit fehlenden Angaben zu Y ($R = 0$) liegt dagegen nur X vor. Y setzt sich somit aus beobachteten („observed“) und fehlenden („missing“) Daten zusammen: $\left(Y_{obs}, Y_{mis}\right)$. Im Folgenden wird angenommen, dass X in jedem Fall beobachtet werden kann. Dies entspricht auch der Datenlage in den späteren Analysen (siehe Kapitel 5 und 6).

Die gemeinsame Verteilung der Analysevariablen (X, Y) und des Ausfallindikators R kann für ein statistisches Modell, in dem θ und φ zu schätzende Parameter des unter inhaltlichen Aspekten interessierenden Strukturmodells der Analysevariablen bzw. des Ausfallmodells bezeichnen, wie folgt faktorisiert werden:

$$f\left(X, Y, R \mid \theta, \varphi\right) = f\left(X, Y \mid \theta\right) f\left(R \mid X, Y, \varphi\right) \tag{3.1}$$

Die Ausfallfunktion auf der rechten Seite im zweiten Teil der Gleichung beschreibt die bedingte Wahrscheinlichkeit von Ausfällen, die abhängig von X, Y und φ sind.

[6] Der Ausfallindikator wird bei Little und Rubin (2002) M (missing) genannt. Die nachstehende Darstellung folgt auch Copeland (2004: 14f.).

Betrachtet man das Entstehen fehlender Daten als Zufallsexperiment und den Indikator *R* als Zufallsvariable, können hinsichtlich der Zusammenhänge zwischen dem Ausfall und anderen Variablen drei Situationen unterschieden werden (Little und Rubin 2002: 11f.): vollkommen zufällige, bedingt zufällige und nicht ignorierbare Ausfälle.

Falls die Wahrscheinlichkeit eines Ausfalls bzw. fehlenden *Y*-Wertes weder mit *Y* noch mit *X* zusammenhängt, gilt für alle *Y*-, *X*- und φ-Werte:

$$f\left(R \mid X, Y, \varphi\right) = f\left(R \mid \varphi\right) \tag{3.2}$$

Diese Situation wird als vollkommen zufälliger Ausfall („missing completely at random“, MCAR) bezeichnet. Die ausgefallenen bzw. fortgezogenen Personen können dann als einfache Zufallsstichprobe aus der insgesamt zufällig ausgewählten Stichprobe angesehen werden. Im Fall von MCAR entsprechen die lediglich auf Basis der Daten ohne Ausfälle ermittelten Analysen zum Zusammenhang zwischen *Y* und *X* den Ergebnissen, die man bei Berücksichtigung aller Befragten erhalten würde.

Trifft die MCAR-Annahme (3.2) zu, vereinfacht sich die in Gleichung (3.1) formulierte Schätzung der interessierenden Zusammenhänge zwischen *Y* und *X* zu

$$f\left(Y, R \mid X, \theta, \varphi\right) = f\left(Y \mid X \mid \theta\right) f\left(R \mid \varphi\right).$$

Hängt der Ausfall von *X*, aber nicht von *Y* ab, gilt für die Gruppe mit fehlenden Werten in $Y\left(Y_{mis}\right)$:

$$f\left(R \mid X, Y, \varphi\right) = f\left(R \mid X, Y_{obs}, \varphi\right) \tag{3.3}$$

Dies entspricht dem Typ eines bedingt zufälligen Ausfalls („missing at random“, MAR). MAR trifft beispielsweise zu, wenn Wegzüge mit dem Alter korrelieren, aber in jeder Altersgruppe kein Zusammenhang zwischen Schulbesuch oder Ausbildungsabschluss und Ausfall besteht. In diesem Fall können die ausgefallenen Personen wie eine einfache Zufallsstichprobe aus der jeweiligen Unterstichprobe von Personen mit denselben *X*-Werten behandelt werden.

Unter dieser Bedingung (3.3) lässt sich die Schätzung der interessierenden Zusammenhänge in (3.1) als $f\left(Y,R \mid X,\theta,\varphi\right)=f\left(Y \mid X,\theta\right)f\left(R \mid X,\varphi\right)$ darstellen. Wenn X beobachtet wurde, können bei vollständigem und bedingt zufälligem Ausfall die Antwortwahrscheinlichkeiten auf Grundlage der vorliegenden Daten leicht berechnet und für Ausfallkorrekturen, wie z. B. Gewichtung mit dem Kehrwert der geschätzten Antwortwahrscheinlichkeit (Horvitz-Thompson-Schätzfunktion), verwendet werden. In der Regel werden bei Paneldaten hierfür Angaben zum Zeitpunkt vor dem Ausfall herangezogen.

Treffen die MCAR- oder MAR-Annahmen zu, kann die gemeinsame Verteilung von Y, X und R in ein Strukturmodell und in ein Ausfallmodell faktorisiert werden. Sind die Parameter des Struktur- und Ausfallmodells (θ, φ) unabhängig voneinander schätzbar (Separierbarkeit), können somit alle Informationen zum Zusammenhang zwischen Y und X in der Gesamtheit (inkl. Ausfällen) aus den beobachteten Daten geschätzt werden. Diese beiden Ausfalltypen werden deshalb als ignorierbar bezeichnet.

Ist jedoch die Wahrscheinlichkeit eines Ausfalls mit der abhängigen Variablen verbunden, d. h. wenn

$$\begin{aligned} &f\left(R \mid X,Y,\varphi\right)=f\left(R \mid X,Y_{obs},Y_{mis},\varphi\right) \text{ oder} \\ &f\left(R \mid X,Y,\varphi\right)=f\left(R \mid X,Y_{mis},\varphi\right) \end{aligned} \tag{3.4}$$

zutrifft, spricht man von nicht ignorierbaren bzw. nicht zufälligen Ausfällen („non-ignorable nonresponse“ (NINR), bzw. „missing not at random“ (MNAR)). Der Zusammenhang zwischen dem Ausfall und der abhängigen Variablen muss dann wie im Unterabschnitt 3.3.3 beschrieben explizit modelliert werden. Bei Verlaufsanalysen bedeutet dies, dass der Beobachtungszeitraum und das Ereignis nicht unabhängig sind.

3.2 Selektionsmodelle für kategoriale Daten

Für die Modellierung der gemeinsamen Verteilung der Analysevariablen (*X*, *Y*) und des Ausfallindikators *R* werden hauptsächlich Selektionsmodelle und Pattern-Mixture Modelle herangezogen (Little und Rubin 2002).

Selektionsmodelle[7] entsprechen direkt der obigen Klassifikation der Ausfallprozesse und der Faktorisierung der gemeinsamen Wahrscheinlichkeitsverteilung für *Y*, *X* und *R*: $P(Y,X)P(R\,|\,Y,X)$; siehe Gleichung (3.1). Es werden zunächst Zusammenhänge und Verteilungen für die vollständigen Daten einschließlich der Ausfälle spezifiziert. Des Weiteren werden Annahmen über das Zustandekommen der Ausfälle formuliert. Hierbei können die Ausfälle sowohl von den beobachteten *X*-Werten als auch von den unbeobachteten *Y*-Werten abhängen.

Selektionsmodelle für metrische abhängige Variablen haben sich in der Praxis insbesondere aufgrund nicht überprüfbarer Verteilungsannahmen teilweise als problematisch herausgestellt (Little und Rubin 2002: 321-324; Stolzenberg und Relles 1997). Bei dem als Alternative dazu entwickelten Pattern-Mixture Modell wird die gemeinsame Wahrscheinlichkeit für *Y, X* und *R* als $P(Y,X\,|\,R)P(R)$ faktorisiert (Little und Rubin 2002: 312ff.; Molenberghs und Verbeke 2005: 555ff.). Aber nicht nur beim Selektionsmodell, sondern auch beim Pattern-Mixture Modell gehen nicht überprüfbare Annahmen und Restriktionen in die Schätzung ein, beispielsweise in der Weise, dass die Zusammenhänge zwischen abhängiger und erklärenden Variablen in der Gruppe mit Ausfall denen der Gruppe ohne Ausfall entsprechen (Little 1993: 129). Als Vorteil von Pattern-Mixture Modellen wird gesehen, dass die Restriktionen offenkundiger sind als bei Selektionsmodellen. Ein weiterer Vorteil von Pattern-Mixture Modellen liegt darin, dass damit Unterschiede der Ausfallprozesse verschiedener Gruppen direkt untersucht werden können. Ist man jedoch an Aussagen über Zusammenhänge in der Gesamtheit (einschließlich Ausfällen) interessiert, ist das Selektionsmodell zu

[7] Der Begriff Selektionsmodell geht auf das Verfahren von Heckman (1976) zurück, das für fehlende oder unvollständig beobachtete metrische Variablen entwickelt wurde (Glynn et al. 1986; Molenberghs und Verbeke 2005: 484).

präferieren, da für die Modellierung der strukturellen Zusammenhänge das gleiche Modell wie im Falle ohne Ausfälle verwendet wird (Allison 2002: 79; Molenberghs und Verbeke 2005: 573).[8]

Im Folgenden wird deshalb das der Klasse der Selektionsmodelle zugehörige log-lineare Modell bzw. die log-lineare Pfadanalyse verwendet,[9] mit der die Ausfallprozesse und strukturellen Zusammenhänge bei kategorialen Daten modellierbar sind. Die gemeinsame Wahrscheinlichkeit von Y und R in Abhängigkeit von X ist $p_{yr|x}$. Die Likelihood ist (Baker und Laird 1988: 63):

$$L = \left[\prod_x \prod_y \left(p_{y1|x} \right)^{n_{xy1}} \right] \left[\prod_x \left(p_{+0|x} \right)^{n_{x+0}} \right] \tag{3.5}$$

mit n_{xy1} Fallzahl der Merkmalskombination X und Y für die Gruppe ohne Ausfall ($R = 1$)

n_{x+0} Fallzahl der Merkmalskombination X für die Ausgefallenen ($R = 0$) mit unbekannten Y-Werten

Für ignorierbare Ausfälle (MCAR, MAR) entspricht der erste Teil der Likelihood einem Strukturmodell, in dem die substanzwissenschaftlich relevanten Variablen modelliert werden. Der zweite Teil der Likelihood kann als Ausfallmodell betrachtet werden. Die Voraussetzung für diese Faktorisierung der Likelihood besteht darin, dass der Ausfall R bedingt unabhängig von Y und X sein muss. Bei Paneldaten mit monotonem Ausfallmuster trifft dies zu (Baker und Laird 1988: 66). Als monotones Ausfallmuster wird bezeichnet, wenn Einheiten im Zeitverlauf sukzessive ausfallen, ohne später wieder befragt zu werden und für sie vollständige Angaben zu den X-Variablen vorliegen. Auf Basis der vollständig

[8] Mit Selektionsmodellen können ebenfalls Zusammenhänge für die Gruppen mit vollständigen Antworten vs. Ausfällen dargestellt werden, sodass sich in dieser Hinsicht Selektionsmodelle und Pattern-Mixture Modelle ähnlich sind (siehe Molenberghs et al. 1999: 113). Selektionsmodelle und Pattern-Mixture Modelle – obwohl sie mathematisch ineinander überführbar sind – mit Ausnahme vollständig zufälliger Ausfälle (MCAR) zu unterschiedlichen Modellen (Little 1993: 126; Toutenburg et al. 2004: 34-37). Während z. B. bei MAR-Annahme das Selektionsmodell $f(X, Y, R) = f(X, Y) f(R \mid X)$ lautet, wird mit dem Pattern-Mixture Modell $f(X, Y \mid R) = f(X, Y) f(R \mid X) / f(R)$ geschätzt (Toutenburg et al. 2004: 37).

[9] Die Bezeichnung des Modells als log-lineares Pfadmodell bzw. Pfadanalyse folgt Vermunt (1997a). Sie geht auf das von Goodman (1973) entwickelte Modell zurück, in dem die gemeinsame Wahrscheinlichkeit von Variablen entsprechend von Überlegungen über ihre kausale Anordnung in ein Produkt von marginalen und bedingten Wahrscheinlichkeiten zerlegt wird.

beobachteten Daten können dann für ein gegebenes Muster von Ausfällen und Kovariaten die fehlenden *Y*-Werte geschätzt werden. Bei vollständig kategorialen Daten ist die Unterscheidung zwischen MAR- und NINR-Modellen allerdings nicht mehr bzw. nur eingeschränkt möglich, da unter MAR-Annahme die Koeffizienten des Ausfallmodells nicht mehr unabhängig von den Koeffizienten des Strukturmodells geschätzt werden können (Vermunt 1997a: 77).

Da *Y* für die Ausgefallenen ($R = 0$) nicht beobachtbar ist, können für die Schätzung keine Standardverfahren der log-linearen Analyse benutzt werden. In den hier verwendeten speziellen log-linearen Modellen, bei denen die fehlenden Daten latente Klassen bzw. Zellenbesetzungen repräsentieren,[10] werden Schätzungen zumeist mithilfe des EM-Algorithmus (EM: Expectation, Maximisation) vorgenommen. Die Grundidee besteht darin, die Schätzung für unvollständige Daten durch eine einfachere iterative Schätzung für vollständige Daten zu ersetzen. In einem ersten E-Schritt werden die unter einem gegebenen Modell erwarteten Zellenbesetzungen für die Ausfallgruppe geschätzt. In einem zweiten M-Schritt erfolgt die Maximierung der Likelihood der vollständigen Daten (inkl. der Ausfälle). Der M-Schritt ergibt neue erwartete Zellenbesetzungen. Dies wird so lange wiederholt, bis sich die Schätzungen nicht mehr unterscheiden (Dempster et al. 1977; Little und Rubin 2002: 164ff.; Schafer 1997: 37ff.); siehe dazu auch das Beispiel auf Seite 21 [hier: S. 11]. Im E-Schritt der Schätzung des obigen Modells (Gl. 3.5) wird die folgende Reparametrisierung verwendet (Baker und Laird 1988):

$$\begin{aligned} p_{yr|x} &= p_{y|x}\, p_{r|yx} \\ p_{y|x} &= m_{xy} / m_{x+} \\ p_{r|yx} &= M_{xyr} / M_{xy+} \end{aligned} \tag{3.6}$$

Mit *m* werden die erwarteten Zellenbesetzungen für die Teiltabelle $R = 0$ und mit *M* die erwarteten Zellenbesetzungen der Gesamttabelle bezeichnet.

[10] Während in der üblichen Analyse latenter Klassen (LCA) jede Beobachtung zu einer potenziellen latenten Dimension gehört, stellen hier die fehlenden Werte latente Zellenbesetzungen dar (Winship et al. 2002: 413).

Zur Lösung der Schätzprobleme bei unvollständigen Daten kann neben dem EM-Algorithmus und bayesianischen Methoden (siehe Little und Rubin 2002; Park und Brown 1994; Forster und Smith 1998) auch ein sogenanntes „Composite Link-Model“ eingesetzt werden (siehe Baker 1994; Chambers und Welsh 1993; Rindskopf 1992, Molenberghs und Goetghebeur 1997). Mit diesem log-linearen Modell werden in einer speziellen Matrix Restriktionen gesetzt, mit denen die linearen Beziehungen zwischen den vollständigen und den unvollständigen Daten abgebildet werden (siehe als einfaches Beispiel für die Behandlung der linearen Beziehungen Seite 34 [hier: S. 15]). Die Konstruktion dieser Restriktionsmatrizen kann jedoch sehr aufwendig sein.

3.3 Beispiele

Das gewählte log-lineare Modell wird im Folgenden anhand eines einfachen Beispiels von Daten des Mikrozensuspanels 1996-1999 aus Kapitel 6, Abschnitt 6.1, beschrieben. Tabelle 3.1 berichtet für Auszubildende des dualen Systems im April 1996 deren Ausbildungsstatus (Y) im April 1997 in Abhängigkeit von der Beschäftigungsdauer (X). Von insgesamt 1.381 Auszubildenden liegen für 1.123 (81,3 %) räumlich immobile Personen ohne Ausfall ($R = 1$) Angaben zu ihrem Ausbildungsstatus im April 1997 vor. Für 258 Auszubildende (18,7 %), die zwischen den Mikrozensuserhebungen 1996 und 1997 fortgezogen sind ($R = 0$), fehlen diese Informationen.

Tabelle 3.1: Ausbildungsstatus im April 1997 nach Beschäftigungsdauer für Auszubildende im April 1996 im Mikrozensuspanel 1996-1999 – Fallzahlen und Zeilenprozentwerte

Beschäftigungsdauer 1996 (X)	mit Angaben zum Ausbildungsstatus (Y)				ohne Ang.	Insgesamt
	1 Auszubildender	2 Abschluss	3 Abbruch	Insges. (R = 1)	Insges. (R = 0)	(R = 0, 1)
1 ≤12 Mon.	420	19	54	493	92	585
	(85,2)	(3,9)	(11,0)	(100)		
2 13-24 Mon.	282	94	43	419	85	504
	(67,3)	(22,4)	(10,3)	(100)		
3 25+ Mon.	25	167	19	211	81	292
	(11,8)	(79,1)	(9,0)	(100)		
Insgesamt	727	280	116	1.123	258	1.381
	(64,7)	(24,9)	(10,3)	(100)		

Quelle: Mikrozensuspanel 1996-1999; eigene Berechnungen (siehe Abschnitt 6.1).

Mit der Ausnahme von MCAR-Modellen können die Schätzungen für diese quadratische (3 x 3 -) Tabelle der Gesamtheit in geschlossener Form dargestellt werden (Baker et al. 1992; Little und Rubin 2002: 268f., 344).[11] Damit sind die Schätzungen, die mit dem verwendeten Programm LEM (Vermunt 1997b) mittels EM-Algorithmus durchgeführt werden, anschaulich nachvollziehbar. Die LEM-Programme zu den in diesem Kapitel geschätzten Modellen sind im Anhang wiedergegeben.

Es wird folgende Notation verwendet. Die entsprechenden Laufindizes der Variablen X, Y und R sind i, j und k. Die beobachteten bzw. geschätzten Zellenbesetzungen sind dann n_{ijk} bzw. m_{ijk}. Das Zeichen „+“ steht für die Summierung über dem entsprechenden Index.

[11] Dass gerade bei dem einfachsten Ausfallmodell MCAR keine Schätzung in geschlossener Form durchführbar ist, erscheint überraschend, liegt jedoch an den Restriktionen log-linearer Modelle der Übereinstimmung von beobachteten und geschätzten Randverteilungen der Gesamttabelle. Z. B. entsprechen aufgrund der unterschiedlichen Randverteilungen der X-Variablen der Gruppen ohne und mit Ausfall bei einfacher Berechnung der Zellenbesetzungen der Ausfallgruppe durch $m_{ij0} = (m_{ij1}/m_{++1})*m_{++0}$ die geschätzten Randverteilungen der X-Variablen für die Gesamttabelle nicht den beobachteten Werten. Berücksichtigt man die unterschiedliche Verteilung der X-Variablen in den Teiltabellen z. B. durch die Berechnung mit $m_{ij0} = (m_{ij1}/m_{i+1})*m_{i+0}$, unterscheiden sich die Randverteilungen der Y-Variablen in den Gruppen ohne und mit Ausfall, was der MCAR-Annahme $P(Y=j \mid R=0) = P(Y=j \mid R=1)$ widerspricht.

Wie die Teiltabelle (3.1) der Befragten mit vollständigen Angaben ($R = 1$) zeigt, hängen die Beschäftigungsdauer und der Ausbildungsstatus eng zusammen. Beträgt die Beschäftigungsdauer im Jahre 1996 bis zu zwölf Monate, befinden sich 1997 noch rund 85 Prozent (420 / 493 = 0,85) der Auszubildenden in der Ausbildung. Dagegen haben 1997 rund 80 Prozent (167 / 211 = 0,79) der Personen mit einer bisherigen Beschäftigungsdauer von wenigstens 25 Monaten den Abschluss erreicht. Zwischen 1996 und 1997 brechen in jeder Klasse der Beschäftigungsdauer rund zehn Prozent die Ausbildung ab.

Wird angenommen, dass der erreichte Ausbildungsstatus (Y) von der Beschäftigungsdauer (X) abhängt, entspricht das Strukturmodell dem log-linearen Modell $\log(m_{ij}) = \alpha + \beta_i^X + \beta_j^Y + \beta_{ij}^{XY}$, kurz $\{XY\}$, bzw. dem äquivalenten Logit-Modell für $P(Y=j|X) = m_{ij} / \sum_j m_{ij}$.

Das Ausfallmodell für $P(R=0)$ ist ebenfalls ein log-lineares bzw. Logit-Modell, das je nach Annahmen zum Ausfallzusammenhang Effekte für X bzw. Y enthalten kann. Wegen der Konzentration auf eine explizite abhängige Variable wird das log-lineare Modell als Logit-Modell spezifiziert und die Notation des LEM-Programms verwendet: $Y|X\ \{XY\}\ R|XY\ \{R|XY\}$. Der erste Ausdruck $Y|X\ \{XY\}$ beschreibt das Strukturmodell und der zweite Ausdruck $R|XY\ \{R|XY\}$ das Ausfallmodell.

3.3.1 Vollständig zufälliger Ausfall (MCAR)

Unter vollkommen zufälligen Ausfällen (MCAR), d. h. wenn in diesem Beispiel Wegzüge weder mit der Beschäftigungsdauer noch mit dem Ausbildungsstatus zusammenhängen, gilt $P(R|X,Y) = P(R)$. Das entsprechende Logit-Modell lautet: $Y|X\ \{XY\}\ R|XY\ \{R\}$. Somit wird für alle Merkmalskombinationen ein konstantes Verhältnis von Ausfällen ($R = 0$) und vollständigen Angaben ($R = 1$) angenommen: $b = P(R=0) / P(R=1) = m_{ij0} / m_{ij1}$. In diesem Fall beträgt $b = 0{,}2297$ (= 258 / 1.123; siehe Tab. 3.1).

Die Zellenbesetzungen m_{ij} und Anteile θ_{ij} werden mit einem einfachen EM-Verfahren geschätzt. Im E-Schritt werden die unter dem MCAR-Modell erwarteten Zellenbesetzungen m_{ij+} der Gesamttabelle wie folgt geschätzt:

$$\hat{E}\left(m_{ij+} \mid Y_{obs}, \theta\right) = \hat{E}\left(m_{ij1} + m_{ij0} \mid Y_{obs}, \theta\right) = n_{ij1} + n_{i+0}\, \hat{\theta}_{ij} \big/ \hat{\theta}_{i+}$$

Im M-Schritt, mit dem die Likelihood der vervollständigten Daten maximiert wird, erfolgt die Schätzung der Anteile θ_{ij} durch:

$$\hat{\theta}_{ij} = \hat{E}\left(m_{ij+} \mid Y_{obs}, \theta\right) \big/ n_{+++}$$

Kombiniert man den E- und M-Schritt ergibt sich für eine Iteration t (Schafer 1997: 44):

$$\hat{\theta}_{ij}^{(t+1)} = \left[n_{ij1} + n_{i+0} \left(\hat{\theta}_{ij}^{t} \big/ \hat{\theta}_{i+}^{t} \right) \right] \big/ n_{+++}$$

Mit den Zellenbesetzungen bzw. Anteilen der räumlich Immobilen ($R = 1$) als Startwerten erhält man nach zehn Iterationen das Endergebnis.

Alternativ kann in diesem Fall der quadratischen Tabelle auch das Verfahren von Baker et al. (1992: 647 [Model (a)]) angewendet werden. Der kombinierte E- und M-Schritt für die Schätzung der Zellenbesetzungen der räumlich Immobilen m_{ij1} mit

$$\hat{m}_{ij1}^{(t+1)} = \left(\left(n_{ij1} + n_{i+0}\, \hat{m}_{ij1}^{t} \big/ \hat{m}_{i+1}^{t} \right) n_{++1} \right) \big/ n_{+++}$$

und $m_{ij1}^{(t=1)} = n_{ij1}$ als Startwerten ergibt bereits bei der zweiten Iterationen das Endergebnis. Durch Multiplikation $m_{ij0} = b * m_{ij1}$ erhält man die Fallzahlen für die Ausfallgruppe ($R = 0$).

Tabelle 3.2 zeigt hierfür die nach der Beschäftigungsdauer (X) bedingten Anteile des Ausbildungsstatus (Y), die modellgemäß für beide Teilgruppen identisch sind. Wie aber der Vergleich mit Tabelle 3.1 zeigt, weichen bei diesem MCAR-Modell die Randverteilungen der geschätzten von denen der beobachteten Daten in den Teilgruppen ab. Aus Platzgründen werden keine Regressionskoeffizienten, sondern nur die geschätzten Anteilswerte und Randverteilungen sowie Kennziffern der Modellschätzung berichtet.

Als Maß der Modellanpassung wird die Devianz $G^2 = -2 * (l_M - l_{beob})$ verwendet, mit der die Log-Likelihood eines gegebenen Modells *M* mit der Log-Likelihood eines saturierten Modells verglichen wird, das Gleichheit von erwarteten und beobachteten Zellenbesetzungen (*beob*) unterstellt. Da sich die statistische Inferenz auf die beobachteten Verteilungen beziehen soll, werden bei der Berechnung der Devianz die einzelnen geschätzten *Y*-Werte der Ausfallgruppe zu einem Wert zusammengefasst. Bei kleinen Zellenbesetzungen kann die übliche Annahme einer Chi-Quadrat-Verteilung verletzt sein. Weniger trifft dies auf die Likelihood-Ratio-Teststatistik $G^2 = -2 * (l_{M1} - l_{M2})$ zu, mit der die Differenz der Devianzen verschiedener Modelle berechnet wird. Die Datentabelle 3.1 umfasst zwölf Zellen. Mit dem MCAR-Modell *Y|X {XY} R|XY {R}* werden zehn log-lineare Parameter geschätzt (α, β_2^X, β_3^X, β_2^Y, β_3^Y, β_{22}^{XY}, β_{23}^{XY}, β_{32}^{XY}, β_{33}^{XY}, β_0^R). Somit verbleiben zwei Freiheitsgrade (*d.f.*; siehe Tabelle 3.2).

Tabelle 3.2: Unter der Annahme vollständig zufälliger Ausfälle (MCAR) geschätzter Ausbildungsstatus – Zeilenprozentwerte

Beschäftigungsdauer 1996 (*X*)	Ausbildungsstatus im April 1997 (*Y*) 1 Auszub.	2 Abschluss	3 Abbruch	Insgesamt
	mit Angaben zum Ausbildungsstatus (*R* = 1)			
1 ≤12 Mon.	85,2	3,9	11,0	475,7
2 13-24 Mon.	67,3	22,4	10,3	409,8
3 25+ Mon.	11,8	79,1	9,0	237,4
Insgesamt	63,2	26,6	10,3	1.123
	ohne Angaben zum Ausbildungsstatus (*R* = 0)			
1 ≤12 Mon.	85,2	3,9	11,0	109,3
2 13-24 Mon.	67,3	22,4	10,3	94,2
3 25+ Mon.	11,8	79,1	9,0	54,6
Insgesamt	63,2	26,6	10,3	258
Insgesamt (*R*=1,0)	63,2	26,6	10,3	1.381

Datenbasis: Tabelle 3.1. Modellspezifikation (Logit-Modell): Y|X {XY} R|XY {R};
Devianz G² = 18,84; d.f. = 2; Log-Likelihood = -2.200,99; Anzahl log-linearer Parameter = 10.

3.3.2 Bedingt zufälliger Ausfall (MAR)

Es ist zu vermuten, dass die Ausfälle infolge räumlicher Mobilität nach erfolgreichem Abschluss, d. h. bei längerer Beschäftigungsdauer, größer sind als für Auszubildende, die noch am Anfang der betrieblichen Ausbildung stehen. Im nächsten Modell soll deshalb ein bedingt zufälliger Ausfall (MAR) angenommen werden, bei dem $P(Y=j \mid X=i, R=0) = P(Y=j \mid X=i,\ R=1)$ gilt. Das Logit-Modell $Y|X$ $\{XY\}$ $R|XY$ $\{RX\}$ enthält für die Schätzung dieser Zusammenhänge die Interaktion zwischen R und X.

Für diese Beispieldaten liefert bereits die einfache getrennte Schätzung von Struktur- und Ausfallmodell (Gl. 3.5) alle Informationen, um die bedingten Wahrscheinlichkeiten der Teiltabelle ohne Ausfälle $P(Y \mid X,\ R=1)$ auf die Teiltabelle mit Ausfällen ($R = 0$) entsprechend der pro X-Kategorie geschätzten Ausfallwahrscheinlichkeit $P(R=0|X)$ und der Verteilung von X (m_{i+0}) in dieser Teiltabelle zu übertragen (Little und Rubin 2002: 266ff.). Die Zellenbesetzungen für die Ausfallgruppe sind einfach wie folgt zu schätzen: $m_{ij0} = \hat{p}_{j \mid i,R=1}\ m_{i+0}$, z. B. $m_{110} = (420/493) * 92 = 78{,}4$.

Die Ausfallwahrscheinlichkeit beträgt für eine Beschäftigungsdauer von bis zu zwei Jahren 16 bzw. 17 Prozent, bei 25 und mehr Monaten jedoch rund 28 Prozent (81 / (81 + 211) = 0,277; siehe Tab. 3.1). Gegenüber dem MCAR-Modell verbraucht das MAR-Modell zwei Parameter mehr, die Likelihood-Ratio-Statistik zeigt aber mit $G^2 = 18{,}84$ (= 2*(|−2.191,57 – −2.200,99|)) eine signifikant bessere Modellanpassung an, sodass die MCAR-Annahme einer konstanten Ausfallwahrscheinlichkeit abzulehnen ist.

Tabelle 3.3: Unter der Annahme bedingt zufälliger Ausfälle (MAR) geschätzter Ausbildungsstatus – Zeilenprozentwerte

Beschäftigungs-dauer 1996 (*X*)	Ausbildungsstatus im April 1997 (*Y*) 1 Auszub.	2 Abschluss	3 Abbruch	Insgesamt
	mit Angaben zum Ausbildungsstatus (*R* = 1)			
1 ≤12 Mon.	85,2	3,9	11,0	493
2 13-24 Mon.	67,3	22,4	10,3	419
3 25+ Mon.	11,8	79,1	9,0	211
Insgesamt	64,7	24,9	10,3	1.123
	ohne Angaben zum Ausbildungsstatus (*R* = 0)			
1 ≤12 Mon.	85,2	3,9	11,0	92
2 13-24 Mon.	67,3	22,4	10,3	85
3 25+ Mon.	11,8	79,1	9,0	81
Insgesamt	56,3	33,6	10,1	258
Insgesamt (*R*=1,0)	63,2	26,6	10,3	1.381

Datenbasis: Tabelle 3.1. Modellspezifikation (Logit-Modell): *Y*|*X* {*XY*} *R*|*XY* {*RX*};
Devianz $G^2 = 0$; d.f. = 0; Log-Likelihood = -2.191,57; Anzahl log-linearer Parameter = 12.

3.3.3 Nicht ignorierbarer Ausfall (NINR)

Wenn Jugendliche am Ende der Ausbildung, d. h. nach etwa drei Jahren und nach erfolgreichem Abschluss i. d. R. über ein höheres Einkommen verfügen können, dürften sie gegenüber Auszubildenden eine höhere Umzugsneigung aufweisen. Ähnlich dürfte die Umzugsneigung von Ausbildungsabbrechern steigen, sofern sie den bisherigen Ausbildungsbetrieb verlassen. Es liegt deshalb nahe, einen Zusammenhang zwischen dem Ausfall bzw. Wegzug und dem Ausbildungsstatus und damit nicht ignorierbare Ausfälle (NINR bzw. MNAR) anzunehmen.

Ist nun die Ausfallwahrscheinlichkeit nicht für alle *Y*-Werte konstant, sondern hängt von den fehlenden Werten selbst ab $P(R \mid Y_{obs}, Y_{mis}) \neq P(R \mid Y_{obs})$, wird bei der Schätzung des NINR-Modells *Y*|*X* {*XY*} *R*|*XY* {*RY*} davon ausgegangen, dass

die geschätzten Spaltenprozentwerte m_{ij} / m_{+j} für die Gruppen ohne und mit Ausfall gleich sind (Little und Rubin 2002: 344):

$$m_{ij0} / m_{+j0} = m_{ij1} / m_{+j1} .$$

Zusammen mit der Bedingung, dass bei der Ausfallgruppe die geschätzten Randverteilungen mit den beobachteten Werten übereinstimmen müssen:

$$\sum_j m_{ij0} = m_{i0}$$

(siehe Gleichungen 7-9 unten), ergibt sich, dass die Verhältnisse (Odds) der geschätzten Zellenbesetzungen für die Ausfallgruppe (m_{ij0}), den Odds der Zellenbesetzungen der vollständig beobachteten Tabelle (m_{ij1}) entsprechen sollen (Gleichungen 1-6). Dies veranschaulicht, dass damit die Verteilungen von Y der vollständig beobachteten Daten (Y_{obs}) auf die Daten der Ausfälle (Y_{mis}) übertragen werden:

(R1) $m_{210}/m_{110} = m_{211}/m_{111}$
(R2) $m_{310}/m_{110} = m_{311}/m_{111}$
(R3) $m_{220}/m_{120} = m_{221}/m_{121}$
(R4) $m_{320}/m_{120} = m_{321}/m_{121}$
(R5) $m_{230}/m_{130} = m_{231}/m_{131}$
(R6) $m_{330}/m_{130} = m_{331}/m_{131}$
(R7) $m_{110} + m_{120} + m_{130} = n_{1+0}$
(R8) $m_{210} + m_{220} + m_{230} = n_{2+0}$
(R9) $m_{310} + m_{320} + m_{330} = n_{3+0}$

Da in diesem Beispiel für die Gruppe ohne Ausfälle im Strukturmodell ein saturiertes Modell $\{XY\}$ geschätzt wird, entsprechen die geschätzten Fallzahlen den beobachteten Werten. Die obigen Restriktionen (R1-R9) bilden die Matrix **A**.[12]

[12] Für diesen Hinweis danke ich Siegfried Gabler.

$$\begin{pmatrix} 0{,}6714 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0{,}0595 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 4{,}9474 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 8{,}7895 & 0 & 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0{,}7963 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0{,}3519 & 0 & 0 & 0 & 0 & 0 & -1 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} m_{110} \\ m_{120} \\ m_{130} \\ m_{210} \\ m_{220} \\ m_{230} \\ m_{310} \\ m_{320} \\ m_{330} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 92 \\ 85 \\ 81 \end{pmatrix}$$

$$\mathbf{A}\ \mathbf{m} = \mathbf{n}$$

Das in der ersten Restriktion R1 relevante Verhältnis der Zellenbesetzungen der vollständigen Tabelle beträgt $n_{211}/n_{111} = 282/420 = 0{,}6714$. Damit lässt sich R1 als $m_{210} = 0{,}6714 * m_{110}$ formulieren. Eine weitere Umformung entspricht der ersten Zeile des obigen linearen Gleichungssystems: $0{,}6714 * m_{110} - m_{210} = 0$. Analog dazu werden die anderen Restriktionen bis R6 dargestellt. Die restlichen Zeilen der Matrix **A** entsprechen den Restriktionen R7 bis R9. Löst man dieses Gleichungssystem mit $\mathbf{m} = \mathbf{A}^{-1}\ \mathbf{n}$, erhält man unter bestimmten Bedingungen die gesuchten Zellenbesetzungen.

Mit dieser Beispieltabelle, in der die Odds der beobachteten Randverteilung ausgefallener Daten nahezu gleich sind ($m_{210}/m_{110} = 0{,}92$ und $m_{310}/m_{110} = 0{,}88$), werden allerdings, wie es häufig bei Datenausfall vorkommen kann, negative Zellenbesetzungen geschätzt. Dies liegt daran, dass die Odds der Randverteilung nicht im Intervall der Odds beobachteter Daten liegen (z. B. $m_{211}/m_{111} = 0{,}67$ und $m_{311}/m_{111} = 0{,}06$). Dann müssen eine oder mehrere Zellenbesetzungen auf Null gesetzt werden und es liegt eine sogenannte „boundary solution“ oder Stutzung vor (Baker et al. 1992; Little und Rubin 2002: 344).

Mithilfe des EM-Algorithmus können allerdings in diesen Fällen immer Lösungen gefunden werden, deren Werte größer als Null sind, also im zulässigen Wertebereich liegen (Baker und Laird 1988). Tabelle 3.4 enthält die entsprechenden Ergebnisse des LEM-Programms.

Für Absolventen wird durch das NINR-Modell die Ausfallwahrscheinlichkeit $P(R=0 \mid Y=j)$ auf 23 Prozent), für Ausbildungsabbrecher auf 60 Prozent geschätzt.[13] Von der Ausfallgruppe werden dementsprechend die meisten Personen den Abbrechern zugeordnet. Bei den vollständig beobachteten Daten weichen aufgrund der Schätzprobleme die geschätzten von den beobachteten Odds etwas ab, weshalb die Devianz G^2 geringfügig von Null verschieden ist. Um Schätzungen der Gesamttabelle inkl. der systematischen Ausfälle zu erhalten, können die Daten der beobachteten Werte n_{ij1} auch mit dem Kehrwert der durch dieses Modell geschätzten Antwortwahrscheinlichkeiten $\left(1-P(R=0 \mid Y=j)\right)$ gewichtet werden.

Für Auszubildende wird durch das NINR-Modell die Ausfallwahrscheinlichkeit auf nahezu Null Prozent geschätzt (siehe Tabelle 3.4: $(0{,}1\,\% * 258)/(52{,}7\,\% * 1.381)$). Geht man davon aus, dass Jugendliche auch während der Ausbildung umziehen, erscheint dieses Teilergebnis wenig plausibel. Es liegt deshalb nahe, zusätzlich Zusammenhänge zwischen dem Ausfall und der Beschäftigungsdauer zu schätzen. Weil aber sowohl das obige MAR-Modell als auch das NINR-Modell bereits alle Freiheitsgrade verbraucht haben, kann der nicht ignorierbare Ausfall für diese Beispieldaten nicht zusätzlich zur MAR-Annahme modelliert werden. Stehen aber weitere erklärende Variablen zur Verfügung, können Freiheitsgrade durch Restriktionen bzw. den Ausschluss höherer Interaktionen gewonnen werden.

[13] Diese Anteile werden vom LEM-Programm berichtet. Sie können auch anhand der Angaben in Tabelle 3.4 berechnet werden; z. B. für die Absolventen: $(32{,}2\,\% * 258)/(26{,}3\,\% * 1.381) = 23$ Prozent.

Tabelle 3.4: Unter der Annahme nicht ignorierbarer Ausfälle (NINR) geschätzter Ausbildungsstatus – Zeilenprozentwerte

Beschäftigungs-dauer 1996 (*X*)	Ausbildungsstatus im April 1997 (*Y*) 1 Auszub.	2 Abschluss	3 Abbruch	Insgesamt
	mit Angaben zum Ausbildungsstatus (*R* = 1)			
1 ≤12 Mon.	84,8	3,9	11,3	495
2 13-24 Mon.	67,9	22,4	9,8	415
3 25+ Mon.	11,8	79,0	9,2	213
Insgesamt	64,7	24,9	10,3	1.123
	ohne Angaben zum Ausbildungsstatus (*R* = 0)			
1 ≤12 Mon.	0,1	6,3	93,5	90
2 13-24 Mon.	0,1	31,1	68,8	89
3 25+ Mon.	0,0	62,9	37,1	79
Insgesamt	0,1	32,2	67,7	258
Insgesamt (*R*=1,0)	52,7	26,3	21,0	1.381

Datenbasis: Tabelle 3.1. Modellspezifikation (Logit-Modell): $Y|X\ \{XY\}\ R|XY\ \{RY\}$;
Devianz $G^2 = 0{,}48$; d.f. = 0; Log-Likelihood = -2.191,81; Anzahl log-linearer Parameter = 12;
„boundary solution“.

Bei jeweils null Freiheitsgraden sind die Devianzen der MAR- und NINR-Modelle nahezu gleich (siehe Tab. 3.3 und 3.4). Wie häufig bei Daten mit fehlenden Werten (Little und Rubin 2002: 344), liegen auch in diesem Beispiel für eine Entscheidung zwischen den verschiedenen Modellannahmen praktisch keine statistischen Kriterien vor. Man kann nun die Schätzergebnisse auf ihre Plausibilität hinterfragen.

In dem obigen NINR-Modell z. B. wird für Auszubildende eine Ausfallwahrscheinlichkeit von fast null Prozent geschätzt. Dies erscheint wenig plausibel. Für solche Überlegungen sind Sensitivitätsanalysen hilfreich, mit denen gezeigt werden kann, in welchem Bereich die Ergebnisse von verschiedenen für plausibel gehaltenen Modellen variieren (Allison 2002; Little und Rubin 2002: 344; Molenberghs und Verbeke 2005: 575ff.). Doch letztlich kann über die Gültigkeit der Ergebnisse nur spekuliert werden, falls keine externen Daten als Prüfstandard („gold standard“) verfügbar sind.

3.3.4 Modelle mit Referenzdaten der Beschäftigtenstichprobe

Es gibt verschiedene Vorschläge, wie externe Referenzdaten bei Ausfallanalysen berücksichtigt werden können. Z. B. verwenden Fitzmaurice et al. (1996) Querschnittsdaten bei der Analyse von Panelausfällen zur Verbesserung der Identifikation von Koeffizienten zu nicht ignorierbaren Ausfällen bei der Analyse von Wahlabsichten. Querschnittsinformationen des Mikrozensus wurden ebenfalls im Verbundprojekt zum Mikrozensuspanel zur Validierung eingesetzt. Darüber hinaus wurde für die Validierung u. a. das Sozioökonomische Panel (SOEP) genutzt (Basic und Rendtel 2005: 8ff.). Durch den Vergleich räumlich immobiler Personen mit den gesamten SOEP-Daten ergaben sich einerseits Hinweise auf den im Mikrozensuspanel bei der Analyse räumlich Immobiler zu erwartenden Bias. Andererseits lieferten die mit den SOEP-Daten beobachteten Verläufe Restriktionen, die in log-linearen Modellen zum Ausfallgeschehen im Mikrozensuspanel zur Identifikation von Parametern verwendet werden konnten (Basic et al. 2005). Dabei müssen allerdings Abweichungen aufgrund unterschiedlicher Stichprobenpläne sowie evtl. Befragungsausfälle im SOEP hingenommen werden.

In dieser Arbeit werden SOEP-Daten nicht verwendet, da damit einerseits Besucher der gymnasialen Oberstufe nicht äquivalent zum Mikrozensus abgegrenzt werden können. Andererseits wären die Fallzahlen des SOEP für Analysen zu beruflichen Ausbildungsverläufen zu klein. Als Referenzdaten dienen Querschnittsangaben der Bildungsstatistik bzw. die Beschäftigtenstichprobe (siehe Kapitel 2, Abschnitte 2 und 3).

In der Beschäftigtenstichprobe liegt zwar im Unterschied zum SOEP keine Information über die räumliche Mobilität von Auszubildenden vor, jedoch können die Daten als Stichprobe der Gesamtheit betrachtet werden, die nicht durch die bei freiwilligen Umfragen auftretenden Probleme von Befragungsausfällen beeinträchtigt ist. Wie die Verlaufsangaben der Beschäftigtenstichprobe und des Mikrozensuspanels in log-linearen Modellen zur Untersuchung selektiver Ausfälle eingesetzt werden können, wird für die Beispieldaten im Folgenden skizziert. Die

Konstruktion der hier verwendeten Daten aus der Beschäftigtenstichprobe ist in Abschnitt 6.1 beschrieben.

Vergleicht man die Verteilungen in der Beschäftigtenstichprobe mit denen des Mikrozensuspanels (ohne Ausfälle), ist die Beschäftigungsdauer näherungsweise gleich verteilt. In der Beschäftigtenstichprobe finden sich jedoch etwas weniger Ausbildungsabsolventen (5.095 / 23.070 = 22 %) und mehr Abbrecher (14 %) als im Mikrozensuspanel (25 % bzw. 10 %; siehe Tab. 3.5). Stärkere Abweichungen sind bei einer Beschäftigungsdauer von wenigstens 25 Monaten zu beobachten. Während in der Beschäftigtenstichprobe für diese Gruppe die Anteile 61 Prozent für Abschluss und 26 Prozent für Abbruch betragen, liegen sie im Mikrozensuspanel bei 79 bzw. neun Prozent. Es ist zu vermuten, dass diese Differenzen mit dem Ausfall zusammenhängen.

Der Indikator für den Datentyp (*D*: 1 = Beschäftigtenstichprobe, 2 = Mikrozensuspanel) kann beim Vergleich wie eine unabhängige Variable behandelt werden. Bei der Modellierung sind zwei mögliche Vorgehensweisen zu unterscheiden.

Analog zu den obigen Beispielen kann man lediglich das Insgesamt ohne Berücksichtigung des Datentyps verwenden. Allerdings wird dabei nicht berücksichtigt, dass die hier verwendeten Daten der Beschäftigtenstichprobe aufgrund der Recodierungen (per Definition) keine Ausfälle enthalten (siehe S. 120 und Fußnote 60). Um dies zu berücksichtigen, sind Ausfälle in der Beschäftigtenstichprobe als strukturelle Nullzellen zu behandeln. Das lässt sich durch die Interaktion *RD* formulieren.

Tabelle 3.5: Ausbildungsstatus im April 1997 nach Beschäftigungsdauer für Auszubildende im April 1996 im Mikrozensuspanel 1996-1999 und in der Beschäftigtenstichprobe – Fallzahlen und Zeilenprozentwerte

Beschäftigungsdauer 1996 (X)	mit Angaben zum Ausbildungsstatus (Y) 1 Auszubildender	2 Abschluss	3 Abbruch	Insges. ($R = 1$)	ohne Ang. Insges. ($R = 0$)	Insgesamt
			Beschäftigtenstichprobe ($D = 1$)			
1 ≤12 Mon.	8.710	399	841	9.950		9.950
	(87,5)	(4,0)	(8,5)	(100)		
2 13-24 Mon.	5.453	1.377	876	7.706		7.706
	(70,7)	(17,9)	(11,4)	(100)		
3 25+ Mon.	668	3.319	1.427	5.414		5.414
	(12,3)	(61,3)	(26,4)	(100)		
Insgesamt	14.831	5.095	3.144	23.070		23.070
	(64,3)	(22,1)	(13,6)	(100)		
			Mikrozensuspanel ($D = 2$)			
1 ≤12 Mon.	420	19	54	493	92	585
	(85,2)	(3,9)	(11,0)	(100)		
2 13-24 Mon.	282	94	43	419	85	504
	(67,3)	(22,4)	(10,3)	(100)		
3 25+ Mon.	25	167	19	211	81	292
	(11,8)	(79,1)	(9,0)	(100)		
Insgesamt	727	280	116	1.123	258	1.381
	(64,7)	(24,9)	(10,3)	(100)		
			Insgesamt			
1 ≤12 Mon.	9.130	418	895	10.443	92	10.535
2 13-24 Mon.	5.735	1.471	919	8.125	85	8.210
3 25+ Mon.	693	3.486	1.446	5.625	81	5.706
Insgesamt	15.558	5.375	3.260	24.193	258	24.451

Quelle: Beschäftigtenstichprobe (IABS-R01) und Mikrozensuspanel 1996-1999; eigene Berechnungen (siehe Abschnitt 6.1).

Dementsprechend sind dann die Modelle bei den verschiedenen Annahmen zum Ausfallzusammenhang wie folgt zu spezifizieren:

Vollständig zufälliger Ausfall (MCAR): $Y|DX\ \{XY\}\ R|DXY\ \{RD\}$

Bedingt zufälliger Ausfall (MAR): $Y|DX\ \{XY\}\ R|DXY\ \{RDX\}$

Nicht ignorierbarer Ausfall (NINR): $Y|DX\ \{XY\}\ R|DXY\ \{RDY\}$

Aufgrund des erheblich größeren Stichprobenumfangs der Beschäftigtenstichprobe werden die Schätzungen des Strukturmodells im Wesentlichen durch die Zusammenhänge in der Referenzstatistik geprägt sein. Mit dem Teilmodell *Y|DX* {*XY*} wird angenommen, dass die strukturellen Zusammenhänge zwischen *Y* und *X* in beiden Stichproben gleich sind. Da die Daten aber auf ganz unterschiedliche Weise entstanden sind (Umfragedaten vs. prozessproduzierte Daten) ist dies eine kritische Annahme. In Bezug auf die Konstruktion vergleichbarer Variablen und Abgrenzungen in beiden Stichproben stellen sich deshalb hohe Anforderungen und es kann grundsätzlich nicht ausgeschlossen werden, dass sich die strukturellen Zusammenhänge zwischen dem Status (*Y*) und den erklärenden Variablen unterscheiden. Dennoch bietet diese Verwendung externer Referenzdaten gegenüber Sensitivitätsanalysen lediglich auf Basis des Mikrozensuspanels entscheidende Vorteile, da Sensitivitätsanalysen nur einen Eindruck über die Variabilität der Schätzergebnisse unter verschiedenen möglichen Ausfallannahmen geben können.

Für den Teil des Strukturmodells gilt zwar die obige Annahme der Verteilungsgleichheit, doch werden im MAR-Ausfallmodell aufgrund der Interaktion *RDX* eventuelle Verteilungsabweichungen der erklärenden Variablen in beiden Daten durch das spezifizierte Logit-Modell kontrolliert, da alle Interaktionen im Modell implizit enthalten sind.

Tabelle 3.6 zeigt die Ergebnisse, die unter der Annahme bedingt zufälliger Ausfälle im Mikrozensuspanel geschätzt werden, wenn im Strukturmodell zusätzlich die Daten der Beschäftigtenstichprobe enthalten sind.

Gegenüber einem Modell mit der Annahme vollständig zufälliger Ausfälle (MCAR; $G^2 = 68{,}89$; d.f. = 8; ohne Darstellung) wird durch das MAR-Modell *Y|DX* {*XY*} *R|DXY* {*RDX*} mit $G^2 = 50{,}04$ (d.f. = 6) eine statistisch signifikant bessere Anpassung an die beobachteten Verteilungen erzielt. Im Vergleich zum MAR-Modell, das nur die Daten des Mikrozensuspanels enthält (siehe Tab. 3.3), zeigen sich insbesondere bei den Anteilen für die Absolventen und Ausbildungsabbrecher mit der höchsten Beschäftigungsdauer deutliche Abweichungen in

Höhe von jeweils rund 17 Prozentpunkten, die auf die Berücksichtigung der Referenzdaten zurückzuführen sind. Während hier von den ausgefallenen Personen mit einer Beschäftigungsdauer von wenigstens 25 Monaten 62 Prozent den Absolventen und 26 Prozent den Abbrechern zugeordnet werden, sind es in Tabelle 3.3 79 bzw. neun Prozent.

Tabelle 3.6: Unter der Annahme bedingt zufälliger Ausfälle (MAR) geschätzter Ausbildungsstatus – Zeilenprozentwerte

Beschäftigungs-dauer 1996 (*X*)	Ausbildungsstatus im April 1997 (*Y*) 1 Auszub.	2 Abschluss	3 Abbruch	Insgesamt
	mit Angaben zum Ausbildungsstatus (*R* = 1)			
	Beschäftigtenstichprobe (*D*=1)			
1 ≤12 Mon.	87,4	4,0	8,6	9.950
2 13-24 Mon.	70,6	18,1	11,3	7.706
3 25+ Mon.	12,3	62,0	25,7	5.414
Insgesamt	64,2	22,3	13,5	23.070
	mit Angaben zum Ausbildungsstatus (*R* = 1)			
	Mikrozensuspanel (*D* = 2)			
1 ≤12 Mon.	87,4	4,0	8,6	493
2 13-24 Mon.	70,6	18,1	11,3	419
3 25+ Mon.	12,3	62,0	25,7	211
Insgesamt (*R* = 1)	67,0	20,2	12,8	1.123
	ohne Angaben zum Ausbildungsstatus (*R* = 0)			
	Mikrozensuspanel (*D* = 2)			
1 ≤12 Mon.	87,4	4,0	8,6	92
2 13-24 Mon.	70,6	18,1	11,3	85
3 25+ Mon.	12,3	62,0	25,7	81
Insgesamt (*R* = 0)	58,3	26,8	14,9	258
	Mikrozensuspanel insgesamt (*D* = 2)			
1 ≤12 Mon.	87,4	4,0	8,6	585
2 13-24 Mon.	70,6	18,1	11,3	504
3 25+ Mon.	12,3	62,0	25,7	292
Insgesamt (*R*=1,0)	65,4	21,4	13,2	1.381
	Beschäftigtenstichprobe und Mikrozensuspanel			
Insgesamt	64,2	22,3	13,5	24.451

Datenbasis: Tabelle 3.5. Modellspezifikation (Logit-Modell): *Y|DX {XY} R|DXY {RD RDX}*; Devianz G^2 = 50,04; d.f. = 6; Log-Likelihood = -47.033,12; Anzahl log-linearer Parameter = 15.

Tabelle 3.7 zeigt, welche Verteilungen unter der Annahme nicht ignorierbarer Ausfälle im Mikrozensuspanel mit dem Modell *Y|DX {XY} R|DXY {RDY}* geschätzt werden.

Tabelle 3.7: Unter der Annahme nicht ignorierbarer Ausfälle (NINR) geschätzter Ausbildungsstatus – Zeilenprozentwerte

Beschäftigungsdauer 1996 (*X*)	Ausbildungsstatus im April 1997 (*Y*) 1 Auszub.	2 Abschluss	3 Abbruch	Insgesamt
	mit Angaben zum Ausbildungsstatus (*R* = 1)			
	Beschäftigtenstichprobe (*D* = 1)			
1 ≤12 Mon.	87,3	4,0	8,7	9.950
2 13-24 Mon.	70,5	18,1	11,5	7.706
3 25+ Mon.	12,3	61,7	26,0	5.414
Insgesamt	64,1	22,2	13,7	23.070
	mit Angaben zum Ausbildungsstatus (*R* = 1)			
	Mikrozensuspanel (*D* = 2)			
1 ≤12 Mon.	89,8	4,1	6,2	486
2 13-24 Mon.	73,2	18,5	8,3	414
3 25+ Mon.	13,4	66,8	19,8	227
Insgesamt (*R* = 1)	68,3	22,0	9,7	1.127
	ohne Angaben zum Ausbildungsstatus (*R* = 0)			
	Mikrozensuspanel (*D* = 2)			
1 ≤12 Mon.	75,4	3,7	20,9	99
2 13-24 Mon.	57,8	16,0	26,2	90
3 25+ Mon.	8,1	44,0	47,9	65
Insgesamt (*R* = 0)	52,0	18,3	29,7	255
	Mikrozensuspanel insgesamt (*D* = 2)			
1 ≤12 Mon.	87,3	4,0	8,7	585
2 13-24 Mon.	70,5	18,1	11,5	504
3 25+ Mon.	12,3	61,7	26,0	292
Insgesamt (*R*=1,0)	65,3	21,3	13,4	1.381
	Beschäftigtenstichprobe und Mikrozensuspanel			
Insgesamt	64,1	22,2	13,7	24.451

Datenbasis: Tabelle 3.5. Modellspezifikation (Logit-Modell): *Y|DX {XY} R|DXY {RD RDY}*; Devianz G^2 = 51,03; d.f. = 6; Log-Likelihood = -47.033,62; Anzahl log-linearer Parameter = 15; „boundary solution“.

Im Vergleich zum NINR-Modell, das ausschließlich auf Daten des Mikrozensuspanels (siehe Tab. 3.4) beruht, ergeben sich andere Verteilungen. Der wohl auffälligste Unterschied ist darin zu erkennen, dass nun bei den Ausfällen auch Anteile von Auszubildenden geschätzt werden und somit das vorherige unplausible Ergebnis nicht mehr auftritt.

Weil die Beschäftigtenstichprobe definitionsgemäß keine Ausfälle enthält, werden die NINR-Schätzungen (siehe Abschnitt 3.3.3) nur auf Basis der Odds der geschätzten Verteilungen der Personen ohne Ausfall des Mikrozensuspanels ($D = 2$, $R = 1$) berechnet. Die Odds des Mikrozensuspanels und der Beschäftigtenstichprobe unterscheiden sich deshalb. Auch wenn für die Ausfallgruppe in Tabelle 3.7 keine Zellenbesetzungen mit Null erkennbar sind, liegt dennoch eine sogenannte „boundary solution“ vor, die damit zusammenhängt, dass mit der im Abschnitt 3.3.3 beschriebenen Lösung für Ausbildungsabsolventen negative Zellenbesetzungen geschätzt werden.

Die noch verbleibenden sechs Freiheitsgrade können dazu genutzt werden, mit dem Modell $Y|DX\ \{XY\}\ R|DXY\ \{RD\ RDX\ RDY\}$ ein vermutlich angemesseneres Modell zu schätzen, das neben bedingt zufälligen, von der Beschäftigungsdauer abhängigen Ausfällen auch nicht ignorierbare, mit dem Ausbildungsstatus verbundene Ausfälle berücksichtigt. Die Devianz dieses Modells beträgt $G^2 = 13{,}03$ (d.f. = 4; siehe Tab. 3.8). Sie ist deutlich niedriger als die Devianzen des MAR-Modells ($G^2 = 50{,}04$) und des ersten NINR-Modells ($G^2 = 51{,}03$). Da durch dieses Modell eine signifikant bessere Anpassung an die beobachteten Werte erreicht werden kann, ist es zu präferieren.

Betrachtet man die geschätzten Anteile des (nicht beobachteten) Ausbildungsstatus der Ausfallgruppe ($R = 0$) in Tabelle 3.8, wird für 64 Prozent der Ausgefallenen der Status eines Auszubildenden geschätzt; dies entspricht etwa den Randverteilungen der beobachteten Daten und ist im Wesentlichen auf die Annahme bedingt zufälliger Ausfälle (MAR) zurückzuführen. Jedoch werden mit diesem Modell insgesamt nur knapp fünf Prozent der Ausfälle den Absolventen zugeord-

net, die bei den beobachteten Daten der Beschäftigtenstichprobe rund 22 Prozent und im Mikrozensuspanel 25 Prozent umfassen.

Tabelle 3.8: Unter der Annahme bedingt zufälliger (MAR) und nicht ignorierbarer Ausfälle (NINR) geschätzter Ausbildungsstatus – Zeilenprozentwerte

Beschäftigungsdauer 1996 (*X*)	Ausbildungsstatus im April 1997 (*Y*) 1 Auszub.	2 Abschluss	3 Abbruch	Insgesamt
	mit Angaben zum Ausbildungsstatus (*R* = 1)			
	Beschäftigtenstichprobe (*D* = 1)			
1 ≤12 Mon.	87,4	4,0	8,6	9.950
2 13-24 Mon.	70,7	17,9	11,4	7.706
3 25+ Mon.	12,5	61,3	26,2	5.414
Insgesamt	64,2	22,1	13,7	23.070
	mit Angaben zum Ausbildungsstatus (*R* = 1)			
	Mikrozensuspanel (*D* = 2)			
1 ≤12 Mon.	87,6	4,6	7,7	495
2 13-24 Mon.	68,9	21,3	9,7	418
3 25+ Mon.	8,1	79,8	12,1	211
Insgesamt (*R* = 1)	65,8	24,9	9,3	1.124
	ohne Angaben zum Ausbildungsstatus (*R* = 0)			
	Mikrozensuspanel (*D* = 2)			
1 ≤12 Mon.	86,4	0,3	13,4	90
2 13-24 Mon.	79,1	1,4	19,5	86
3 25+ Mon.	23,9	13,3	62,8	81
Insgesamt (*R* = 0)	64,2	4,8	31,0	257
	Mikrozensuspanel insgesamt (*D* = 2)			
1 ≤12 Mon.	87,4	4,0	8,6	585
2 13-24 Mon.	70,7	17,9	11,4	504
3 25+ Mon.	12,5	61,3	26,2	292
Insgesamt (*R*=1,0)	65,5	21,2	13,4	1.381
	Beschäftigtenstichprobe und Mikrozensuspanel			
Insgesamt	64,3	22,0	13,7	24.451

Datenbasis: Tabelle 3.5. Modellspezifikation (Logit-Modell): *Y|DX {XY} R|DXY {RD RDX RDY}*; Devianz G^2 = 13,03; d.f. = 4; Log-Likelihood: -47.014,62; Anzahl log-linearer Parameter = 17; „boundary solution“.

Für 31 Prozent der Ausgefallenen wird geschätzt, dass sie die Ausbildung abgebrochen haben (siehe Tab. 3.8). Die entsprechenden Anteile liegen bei den beobachteten Daten ($R = 1$) des Mikrozensuspanels bzw. bei der Beschäftigtenstichprobe mit neun bzw. 14 Prozent deutlich niedriger. Diese Schätzungen der Anteile der Absolventen und Ausbildungsabbrecher unter den Ausgefallenen konnten ansatzweise schon in dem NINR-Modell zum Mikrozensuspanel ohne Berücksichtigung der Beschäftigtenstichprobe (siehe Tab. 3.7) beobachtet werden. Sie sind somit im Wesentlichen auf die Modellierung nicht ignorierbarer Ausfälle zurückzuführen.

Das für die Ausfälle mit diesen einfachen Beispieldaten erzielte Modellergebnis lässt sich dahingehend interpretieren, dass erfolgreiche Ausbildungsabsolventen eine geringere, Ausbildungsabbrecher aber eine größere Ausfallwahrscheinlichkeit infolge räumlicher Mobilität aufweisen. Ob dieses vorläufige Ergebnis auch Bestand hat, wenn zusätzliche erklärende Variablen betrachtet werden, bleibt den späteren Analysen vorbehalten.

Zusammenfassend kann festgehalten werden, dass diese Modelle nicht nur die Prüfung substanzwissenschaftlicher Hypothesen zu Zusammenhängen in der Gesamttabelle (inkl. Ausfällen), sondern auch die Überprüfung von Annahmen über Zusammenhänge zwischen dem Ausfall und den Analysevariablen erlauben (Allison 2002; Baker und Laird 1988; Fay 1986; Little 1985; Molenberghs und Verbeke 2005; Schafer 1997; Toutenburg et al. 2004; Vermunt 1997a; Winship et al. 2002). Es handelt sich dabei um ein in der Praxis bewährtes Verfahren, das zudem mit dem Programm LEM leicht anzuwenden ist. Wie oben gezeigt, können dabei auch Verteilungen von Referenzdaten direkt in den Modellen berücksichtigt werden. In den Beispielen wurde nur eine abhängige Variable für den Übergang vom Ausbildungsstatus im Jahre 1996 auf den Ausbildungsstatus im Jahre 1997 (Y) behandelt. Mit einer Erweiterung können auch weitere Übergänge (z. B. Y_t, Y_{t+1}) und entsprechende zeitabhängige Ausfallzusammenhänge modelliert werden (Conaway 1992, 1993; Rendtel 1995: 274-281; siehe hierzu auch die Anwendung in Abschnitt 5.4.3).

Literatur (Auszug)

Allison, Paul D., 2002: Missing Data. Sage University Papers Series on Quantitative Applications in the Social Sciences, 07-136. Thousand Oaks, CA: Sage.

Baker, Stuart G., 1994: Missing Data. Composite Linear Models for Incomplete Multinomial Data. Statistics in Medicine 13: 609-622.

Baker, Stuart G., und Nan M. Laird, 1988: Regression Analysis for Categorical Variables With Outcome Subject to Nonignorable Nonresponse. Journal of the American Statistical Association 83(401): 62-69.

Baker, Stuart G., William F. Rosenberger und Rebecca DerSimonian, 1992: Closed-form estimates for missing counts in two-way contingency tables. Statistics in Medicine 11(5): 643-657.

Basic, Edin, Ivo Marek und Ulrich Rendtel, 2005: The German Microcensus as a tool for longitudinal data analysis: An evaluation using SOEP data. Methodenverbund „Aufbereitung und Bereitstellung des Mikrozensus als Panelstichprobe“, Arbeitspapier Nr. 3. Berlin: Freie Universität. URL: http://www.destatis.de/jetspeed/portal/cms/Sites/destatis/Internet/DE/Content/Wissenschaftsforum/MethodenVerfahren/Mikrozensus/Arbeitspapiere/Arbeitspapier3,property=file.pdf; 29. 06. 2007.

Basic, Edin, und Ulrich Rendtel, 2005: Estimation strategies in the presence of non-coverage in the German Microcensus-Panel: An evaluation using SOEP data. Methodenverbund „Aufbereitung und Bereitstellung des Mikrozensus als Panelstichprobe“, Arbeitspapier Nr. 8. Berlin: Freie Universität. URL: http://www.destatis.de/jetspeed/portal/cms/Sites/destatis/Internet/DE/Content/Wissenschaftsforum/MethodenVerfahren/Mikrozensus/Arbeitspapiere/Arbeitspapier8,property=file.pdf; 29. 06. 2007.

Chambers, Ray L., und Alan H. Welsh 1993: Log-linear models for survey data with non-ignorable non-response. Journal of the Royal Statistical Society (B) 55: 151-170.

Conaway, Mark R., 1992: The Analysis of Repeated Categorical Measurements Subject to Nonignorable Nonresponse. Journal of the American Statistical Association 87(419): 817-824.

Conaway, Mark R., 1993: Non-ignorable Non-response Models for Time-ordered Categorical Variables. Applied Statistics 42(1): 105-115.

Copeland, Kennon R., 2004: Panel survey estimation in the presence of late reporting and nonresponse. Dissertation. College Park, MD: University of Maryland. URL: https://drum.umd.edu/dspace/bitstream/1903/1762/1/umi-umd-1740.pdf.

Dempster, Arthur P., Nan M. Laird und Donald B. Rubin, 1977: Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society (B) 39: 1-38.

Fay, Robert E., 1986: Causal Models for Patterns of Nonresponse. Journal of the American Statistical Association 81(394): 354-365.

Fitzmaurice, Garrett M., Peter Clifford und Anthony F. Heath, 1996: Logistic Regression Models for Binary Panel Data with Attrition. Journal of the Royal Statistical Society A 159(2): 249-263.

Forster, Jonathan J., und Peter W.F. Smith, 1998: Model-based inference for categorical survey data subject to non-ignorable non-response. Journal of the Royal Statistical Society B 60(1): 57-70.

Glynn, Robert J., Nan M. Laird und Donald B. Rubin, 1986: Selection modeling versus mixture modeling with non-ignorable nonresponse. S. 115-152 in Howard Wainer: Drawing Inferences from Self-selected Samples. New York: Springer.

Goodman, Leo A., 1973: The analysis of multidimensional contingenty tables when some variables are posterior to others: a modifed path analysis approach. Biometrika 60: 179-192.

Heckman, James J., 1976: The common structure of statistical models of truncation, sample selection and limited dependent variables, and a simple estimation method for such models. Annals of Economic and Social Measurement 5: 475–492.

Little, Roderick J.A., 1985: Nonresponse Adjustments in Longitudinal Surveys: Models for Categorical Data. Bulletin of the International Statistical Institute 15: 1-15.

Little, Roderick J. A., 1993: Pattern-mixture models for multivariate incomplete data. Journal of the American Statistical Association 88: 125–134.

Little, Roderick J.A., und Donald B. Rubin, 2002: Statistical Analysis with Missing Data. 2. Auflage. New York: Wiley.

Molenberghs, Geert, und Geert Verbeke, 2005: Models for Discrete Longitudinal Data. New York: Springer.

Molenberghs, Geert, und Els Goetghebeur, 1997: Simple Fitting Algorithms for Incomplete Categorical Data. Journal of the Royal Statistical Society B (Methodological) 59(2): 401-414.

Molenberghs, Geert, Els J. T. Goetghebeur, Stuart R. Lipsitz und Michael G. Kenward, 1999: Nonrandom Missingness in Categorical Data: Strengths and Limitations. The American Statistician 53(2): 110-118.

Park, Taesung, und Morton B. Brown 1994: Models for categorical data with non-ignorable nonresponse. Journal of the American Statistical Association 89: 44-52.

Rendtel, Ulrich, 1995: Lebenslagen im Wandel: Panelausfälle und Panelrepräsentativität. Frankfurt: Campus.

Rindskopf, David, 1992: A general approach to categorical data analysis with missing data, using generalized linear models with composite links. Psychometrika 57(1): 29-42.

Rubin, Donald B., 1976: Inference and Missing Data. Biometrika 63(3): 581-592.

Schafer, Joseph L., 1997: Analysis of Incomplete Multivariate Data. London: Chapman & Hall.

Schimpl-Neimanns, Bernhard, 2006b: Berufliche Ausbildungsverläufe bis zum Übergang ins Erwerbsleben - Analysen zur Stichprobenselektivität des Mikrozensuspanels 1996-1999. ZUMA-Arbeitsbericht 2006/02. Mannheim: ZUMA. URL: http://www.gesis.org/fileadmin/upload/forschung/ publikationen/gesis_reihen/zuma_arbeitsberichte/AB06_02_Schimpl.pdf.

Statistisches Bundesamt (Hg.), 2006a: Handbuch Mikrozensus-Panel 1996-1999. Methodenverbund Aufbereitung und Bereitstellung des Mikrozensus als Panelstichprobe. Version 0.2 - Juli 2006. Bonn: Statistisches Bundesamt, Gruppe VIII C (Mikrozensus).

Stolzenberg, Ross M., und Daniel A. Relles, 1997: Tools for Intuition about Sample Selection Bias and its Correction. American Sociological Review 62(3): 494-507.

Toutenburg, Helge, Christian Heumann und Thomas Nittner, 2004: Statistische Methoden bei unvollständigen Daten. Discussion Paper 380. München: Ludwig-Maximilians-Universität. URL: http://www.stat.uni-muenchen.de/sfb386 /papers/dsp/paper380.ps.

Vermunt, Jeroen, 1996: Causal log-linear modeling with latent variables and missing data. S. 35-60 in: Uwe Engel und Jost Reinecke (Hg.): Analysis of Change: Advanced Techniques in Panel Data Analysis. Berlin: de Gruyter.

Vermunt, Jeroen, 1997a: Log-linear models for event histories. Advanced Quantitative Techniques in the Social Sciences; 8. Thousand Oaks: Sage.

Vermunt, Jeroen, 1997b: LEM: A general program for the analysis of categorical data. Tilburg University.

Winship, Christopher, Robert D. Mare und John R. Warren, 2002: Latent Class Models for Contingency Tables with Missing Data. S. 408-432 in: Jacques A. Hagenaars und Allan L. McCutcheon (Hg.): Applied Latent Class Analysis. Cambridge, NY: Cambridge University Press.

Auszug aus: Schimpl-Neimanns, Bernhard: Bildungsverläufe und Stichprobenselektivität - Analysen zur Stichprobenselektivität des Mikrozensuspanels 1996 - 1999 am Beispiel bildungsstatistischer Fragestellungen. GESIS-Forschungsberichte: Reihe Sozialwissenschaftliche Datenanalyse; Bd. 1. Bonn: GESIS, 2008: Kapitel 3 (S. 25-42) und Anhang (S. 177-180).

Anhang: LEM-Programme zu den Selektionsmodellen in Kapitel 3

1 MCAR-Programm zum Beispiel in Abschnitt 3.3.1, Tabelle 3.2 (tab3_2.inp)

Kommando		Kommentar
res 1	*	Anzahl Responsevariablen (R)
man 2	*	Anzahl manifester Variablen (X, Y)
dim 2 3 3	*	Anzahl Ausprägungen der Variablen R, X, Y
lab R X Y	*	Variablenlabel
sub XY X	*	Subgruppen mit Angaben zu XY bzw. nur zu X
mod Y|X {XY} R|XY {R}	*	Logit-Modell zum Ausfalltyp MCAR
* mod {XY R}	*	Alternative Spezifikation als log-lineares Modell
rec 12	*	Anzahl der Datenzeilen in „tab3_1.dat“
rco	*	Fallzähler (freq) in „tab3_1.dat“
dat tab3_1.dat	*	Datentabelle „tab3_1.dat“, s. u.
dum 1 1 1	*	Referenzkategorien der Variablen R, X, Y
wma RXY tab3_2.fre	*	Ausgabe geschätzter Zellenbesetzungen

Datentabelle „tab3_1.dat“ (siehe Tabelle 3.1, Mikrozensuspanel 1996-1999)

X	Y	freq		Beschäftigungsdauer 1996 in Monaten (X)	Ausbildungsstatus im April 1997 (Y)	Subgruppe
1	1	420	*	<=12	Azubi	XY
1	2	19	*	<=12	Abschluss	
1	3	54	*	<=12	Abbruch	
2	1	282	*	13-24	Azubi	
2	2	94	*	13-24	Abschluss	
2	3	43	*	13-24	Abbruch	
3	1	25	*	>=25	Azubi	
3	2	167	*	>=25	Abschluss	
3	3	19	*	>=25	Abbruch	
1	0	92	*	<=12	Missing	X
2	0	85	*	13-24	Missing	
3	0	81	*	>=25	Missing	

2 MAR-Programm zum Beispiel in Abschnitt 3.3.2, Tabelle 3.3 [tab3_3.inp]

Kommando		Kommentar
res 1	*	Anzahl Responsevariablen (R)
man 2	*	Anzahl manifester Variablen (X, Y)
dim 2 3 3	*	Anzahl Ausprägungen der Variablen R, X, Y
lab R X Y	*	Variablenlabel
sub XY X	*	Subgruppen mit Angaben zu XY bzw. nur zu X
mod Y|X {XY} R|XY {RX}	*	Logit-Modell zum Ausfalltyp MAR
* mod {XY RX}	*	Alternative Spezifikation als log-lineares Modell
rec 12	*	Anzahl der Datenzeilen in „tab3_1.dat“
rco	*	Fallzähler (freq) in „tab3_1.dat“
dat tab3_1.dat	*	Datentabelle „tab3_1.dat“, s. o.
dum 1 1 1	*	Referenzkategorien der Variablen R, X, Y
wma RXY tab3_3.fre	*	Ausgabe geschätzter Zellenbesetzungen

3 NINR-Programm zum Beispiel in Abschnitt 3.3.3, Tabelle 3.4 [tab3_4.inp]

Kommando		Kommentar
res 1	*	Anzahl Responsevariablen (R)
man 2	*	Anzahl manifester Variablen (X, Y)
dim 2 3 3	*	Anzahl Ausprägungen der Variablen R, X, Y
lab R X Y	*	Variablenlabel
sub XY X	*	Subgruppen mit Angaben zu XY bzw. nur zu X
mod Y|X {XY} R|XY {RY}	*	Logit-Modell zum Ausfalltyp NINR
* mod {XY RY}	*	Alternative Spezifikation als log-lineares Modell
rec 12	*	Anzahl der Datenzeilen in „tab3_1.dat“
rco	*	Fallzähler (freq) in „tab3_1.dat“
dat tab3_1.dat	*	Datentabelle „tab3_1.dat“, s. o.
dum 1 1 1	*	Referenzkategorien der Variablen R, X, Y
wma RXY tab3_4.fre	*	Ausgabe geschätzter Zellenbesetzungen

4 MAR-Programm zum Beispiel in Abschnitt 3.3.4, Tabelle 3.6 [tab3_6.inp]

```
Kommando                          Kommentar
res 1                          *  Anzahl Responsevariablen (R)
man 3                          *  Anzahl manifester Variablen (D, X, Y)
dim 2 2 3 3                    *  Anzahl Ausprägungen der Variablen R, D, X, Y
lab R D X Y                    *  Variablenlabel
sub DXY DX                     *  Subgruppen mit Angaben zu DXY bzw. DX
mod Y|DX {XY}                  *  Logit-Modell zum Ausfalltyp MAR
  R|DXY {cov(RD,1), wei(RD),
  cov(RDX,2), wei(RDX)}
rec 21                         *  Anzahl der Datenzeilen in „tab3_5.dat“
rco                            *  Fallzähler (freq) in „tab3_5.dat“
dat tab3_5.dat                 *  Datentabelle „tab3_5.dat“, s. u.
des [ 0 0 0 1                  *  Designmatrix für Effekte:  cov(RD,1)
      0 0 0 0 0 0 0 0 0 0 1 0  *    cov(RDX,2): R=2, D=2, X=2
      0 0 0 0 0 0 0 0 0 0 0 1 ]*    cov(RDX,2): R=2, D=2, X=3
sta wei(RD) [ 1 1 0 1 ]        *  Startwerte für Effekte RD und RDX:
sta wei(RDX)                   *  Gewichtung struktureller Nullzellen in den
        [ 1 1 1 1 1 1 0 0 0 1 1 1 ]   IABS-Daten
dum 1 1 1 1                    *  Referenzkategorien der Variablen R, D, X, Y
wma RDXY tab3_6.fre            *  Ausgabe geschätzter Zellenbesetzungen
```

Datentabelle „tab3_5.dat“ (siehe Tabelle 3.5,
Beschäftigtenstichprobe (IABS-R01) (D=1), Mikrozensuspanel 1996-1999 (D=2)

D	X	Y	freq		Beschäftigungsdauer 1996 in Monaten (X)	Ausbildungsstatus im April 1997 (Y)	Subgruppe
1	1	1	8710	*	<=12	Azubi	DXY
1	1	2	399	*	<=12	Abschluss	
1	1	3	841	*	<=12	Abbruch	
1	2	1	5453	*	13-24	Azubi	
1	2	2	1377	*	13-24	Abschluss	
1	2	3	876	*	13-24	Abbruch	
1	3	1	668	*	>=25	Azubi	
1	3	2	3319	*	>=25	Abschluss	
1	3	3	1427	*	>=25	Abbruch	
2	1	1	420	*	<=12	Azubi	
2	1	2	19	*	<=12	Abschluss	
2	1	3	54	*	<=12	Abbruch	
2	2	1	282	*	13-24	Azubi	
2	2	2	94	*	13-24	Abschluss	
2	2	3	43	*	13-24	Abbruch	
2	3	1	25	*	>=25	Azubi	
2	3	2	167	*	>=25	Abschluss	
2	3	3	19	*	>=25	Abbruch	
2	1	0	92	*	<=12	Missing	DX
2	2	0	85	*	13-24	Missing	
2	3	0	81	*	>=25	Missing	

5 NINR-Programm zum Beispiel in Abschnitt 3.3.4, Tabelle 3.7 [tab3_7.inp]

Kommando		Kommentar
res 1	*	Anzahl Responsevariablen (R)
man 3	*	Anzahl manifester Variablen (D, X, Y)
dim 2 2 3 3	*	Anzahl Ausprägungen der Variablen R, D, X, Y
lab R D X Y	*	Variablenlabel
sub DXY DX	*	Subgruppen mit Angaben zu DXY bzw. DX
mod Y|DX {XY}	*	Logit-Modell zum Ausfalltyp NINR
R|DXY {cov(RD,1), wei(RD),		
cov(RDY,2), wei(RDY)}		
rec 21	*	Anzahl der Datenzeilen in „tab3_5.dat“
rco	*	Fallzähler (freq) in „tab3_5.dat“
dat tab3_5.dat	*	Datentabelle „tab3_5.dat“, s. o.
des [0 0 0 1	*	Designmatrix für Effekte: cov(RD,1)
0 0 0 0 0 0 0 0 0 0 1 0	*	cov(RDY,2): R=2, D=2, Y=2
0 0 0 0 0 0 0 0 0 0 0 1]	*	cov(RDY,2): R=2, D=2, Y=3
sta wei(RD) [1 1 0 1]	*	Startwerte für Effekte RD und RDY:
sta wei(RDY)	*	Gewichtung struktureller Nullzellen in den
[1 1 1 1 1 1 0 0 0 1 1 1]		IABS-Daten
dum 1 1 1 1	*	Referenzkategorien der Variablen R, D, X, Y
wma RDXY tab3_7.fre	*	Ausgabe geschätzter Zellenbesetzungen

6 MAR- und NINR-Programm zum Beispiel in Tabelle 3.8 [tab3_8.inp]

Kommando		Kommentar
res 1	*	Anzahl Responsevariablen (R)
man 3	*	Anzahl manifester Variablen (D, X, Y)
dim 2 2 3 3	*	Anzahl Ausprägungen der Variablen R, D, X, Y
lab R D X Y	*	Variablenlabel
sub DXY DX	*	Subgruppen mit Angaben zu DXY bzw. DX
mod Y|DX {XY}	*	Logit-Modell zum Ausfalltyp MAR und NINR
R|DXY {cov(RD,1), wei(RD),		
cov(RDY,2), wei(RDY)		
cov(RDX,2),wei(RDX)}		
rec 21	*	Anzahl der Datenzeilen in „tab3_5.dat“
rco	*	Fallzähler (freq) in „tab3_5.dat“
dat tab3_5.dat	*	Datentabelle „tab3_5.dat“, s. o.
des [0 0 0 1	*	Designmatrix für Effekte: cov(RD,1)
0 0 0 0 0 0 0 0 0 0 1 0	*	cov(RDY,2): R=2, D=2, Y=2
0 0 0 0 0 0 0 0 0 0 0 1	*	cov(RDY,2): R=2, D=2, Y=3
0 0 0 0 0 0 0 0 0 0 1 0	*	cov(RDX,2): R=2, D=2, X=2
0 0 0 0 0 0 0 0 0 0 0 1]	*	cov(RDX,2): R=2, D=2, X=3
sta wei(RD) [1 1 0 1]	*	Startwerte für Effekte RD und RDY:
sta wei(RDY)	*	Gewichtung struktureller Nullzellen in den
[1 1 1 1 1 1 0 0 0 1 1 1]		IABS-Daten
dum 1 1 1 1	*	Referenzkategorien der Variablen R, D, X, Y
wma RDXY tab3_8.fre	*	Ausgabe geschätzter Zellenbesetzungen