Behandlung von fehlenden Werten in den Einflußgrößen von Regressionsmodellen – Ein Überblick

Werner Vach[1], Martin Schumacher[2]

Kurztitel: Fehlende Werten in den Einflußgrößen

Short title: Missing values in the covariates

Schlüsselwörter: Complete Case Analyse, fehlende Werte, Imputationsmethoden, Maximum Likelihood Schätzung, Missing at Random, Regressionsmodelle

Keywords: complete case analysis, imputation methods, maximum likelihood estimation, missing at random, missing values, regression models

[1]Freiburger Zentrum für Datenanalyse und Modellbildung, Universität Freiburg
[2]Institut für Medizinische Biometrie und Medizinische Informatik, Universität Freiburg

Zusammenfassung:

Bei dem Einsatz von Regressionsmodellen in biometrischen Anwendungen bilden fehlende Werte in den Einflußgrößen eine zusätzliche Erschwernis, die in der Praxis häufig auftritt. Die Standardlösung der meisten statistische Softwareprodukte ist die Complete Case Analyse, d.h. die Einschränkung auf Beobachtungseinheiten mit vollständiger Kovariableninformation. Dies ist jedoch offensichtlich nicht effizient und bedingt unter Umständen sogar eine Verzerrung der Effektschätzer. Wir geben einen Überblick über alternative Ansätze zur Behandlung der fehlenden Werte. Dabei zeigen wir die Mängel einiger sehr einfacher Methoden auf und stellen einige neuere Ansäzte vor. Jede effiziente Analyse ist jedoch an Annahmen über die Entstehung der fehlender Werte gekoppelt, die häufig zweifelhaft sind. Wir geben daher für einige praxisrelevante Situationen die entscheidenen Überlegungen an, die die Wahl einer adäquaten Methode beeinflussen.

Summary:

In biometrical applications the use of regression models is often complicated by the occurence of missing values in the covariates. Most statistical software packages just exclude all units with incomplete covariate information from the analysis. This complete case analysis is not only inefficient, it can be also a source of bias. We give an overview of alternative methods to deal with incomplete covariate data. We demonstrate the deficiencies of some simple methods and present some modern approaches. However, each efficient approach is based on assumptions on the missing value mechanism, which are often doubtful in practice. For some situations relevant in practice we present the essential conditions determining the choice of an adequate method.

Danksagung: Wir danken der Deutschen Forschungsgemeinschaft für die Förderung eines Forschungsprojektes, auf dessen Ergebnissen die vorliegende Arbeit beruht.

1. Einleitung

Die Untersuchung des simultanen Einflusses mehrerer Variablen auf eine Zielgröße ist eine grundlegende Aufgabe in vielen Anwendungsgebieten der Biometrie. Regressionsmodelle bilden hierfür einen geeigneten Rahmen, und die entsprechenden Methoden der statistischen Inferenz gehören zum Standardrepertoire des angewandten Statistikers. Grundlage jeder Analyse ist dabei die Erhebung aller Einflußgrößen und der Zielgröße an unabhängigen Beobachtungseinheiten. In der Praxis ist es jedoch oft nicht möglich, für alle Beobachtungseinheiten tatsächlich alle Einflußgrößen zu erfassen. Die Gründe für derartige Lücken in den Daten können sehr unterschiedlich sein und hängen oft eng mit dem Erhebungsmechanismus zusammen. Einige Beispiele sollen dies näher beleuchten:

- Werden Einflußgrößen über eine persöhnliche Befragung oder einen Fragebogen erhoben, so ist die Antwortverweigerung eine typische Quelle fehlender Werte, insbesondere wenn es um „sensible“ Fragen wie Drogenkonsum oder Einkommen geht.
- Im Gegensatz zur absichtlichen Antwortverweigerung können fehlende Werte bei einer Befragung auch durch Gedächtnislücken entstehen, z.B. bei der Frage nach Kinderkrankheiten.
- Bei der Benutzung von Dokumenten wie z.B. Krankenhausunterlagen können fehlende Werte durch lückenhafte oder widersprüchliche Angaben entstehen.
- Die Messung von Laborparametern ist manchmal nur ab einer bestimmten Menge von Blut, Urin oder Gewebe möglich, die nicht immer beigebracht werden kann.
- Ist die Messung einer Kovariable sehr teuer, so kann es notwendig sein, die Erhebung dieser Kovariablen auf eine Teilpopulation einzuschränken.
- Auch in einer gut geplanten und durchgeführten Studie kann es zu kleinen „Unfällen“ kommen. Ein Reagenzglas zerbricht, eine Probe oder ein Formblatt geht verloren, eine Untersuchung wird vergessen, oder die Fehladjustierung eines Meßgerätes wird zu spät entdeckt. Jeder solche Unfall verursacht fehlende Werte.

Die Standardmethode zur Behandlung fehlender Werte, die die meisten statistischen Programmpakete benutzen, ist die Complete Case Analyse, d.h. die Auswertung wird auf die Beobachtungseinheiten mit vollständiger Kovariableninformation eingeschränkt. Dieses Vorgehen ist offensichtlich nicht sehr effizient, da die ausgeschlossenen Beobachtungseinheiten auch Information über den Zusammenhang zwischen Einfluß- und Zielgrößen tragen. Die vorliegende Arbeit soll einen Überblick über die Möglichkeiten und Grenzen bei der Behandlung von fehlenden Werten in den Einflußgrößen geben. Dabei werden sowohl einfache Ansätze betrachtet, die sich innerhalb bestehender Programmpakete realisieren lassen (Kapitel 5), als auch neuere, kompliziertere Ansätze, die die methodische Forschung der letzten Jahre hervorgebracht hat (Kapitel 6).

Wir werden unsere Betrachtungen auf den Fall einer Zielgröße Y und zweier Kovariablen X_1, X_2 beschränken, wobei nur die zweite Kovariable mit fehlenden Werten behaftet ist. Die Beobachtbarkeit von X_2 wird durch die Zufallsvariable R_2 indiziert, d.h.

$$R_2 := \begin{cases} 1 & \text{falls } X_2 \text{ beobachtet} \\ 0 & \text{falls } X_2 \text{ „missing"} \end{cases} .$$

In dieser Situation liegt also das Hauptinteresse darin, bei der Schätzung des Effektes von X_1 die Beobachtungseinheiten mit $R_2 = 0$ zu nutzen, die von der Complete Case Analyse vernachlässigt werden. Eine Verbesserung der Schätzung des Effektes von X_2 selbst ist nicht zu erwarten. Auf zusätzliche Schwierigkeiten, die sich bei komplexeren Situationen mit mehreren Kovariablen und beliebigem Muster der Fehlstellen ergeben, werden wir gegebenenfalls bei den einzelnen Verfahren eingehen.

Die Ausagen der vorliegenden Arbeit gelten in weiten Teilen für beliebige Regressionsmodelle, bei denen die bedingte Verteilung von Y gegeben X_1 und X_2, die wir mit $\mathcal{D}(Y|x_1, x_2)$ bezeichnen, zumindest partiell modelliert wird. Teilweise ergeben sich für spezielle Modelle wie der Gaußschen Regression, der logistischen Regression oder dem Cox-Modell besondere Aussagen.

2. Mechanismen für fehlende Werte

Die Eigenschaften von Methoden zur Behandlung fehlender Werte hängen von dem Wahrscheinlichkeitsgesetz ab, nach dem die fehlenden Werte entstehen. Es ist daher notwendig, zuerst die entsprechenden Mechanismen zu betrachten, die allgemein durch die bedingten Beobachtungswahrscheinlichkeiten

$$q(y, x_1, x_2) := P(R_2 = 1|Y = y, X_1 = x_1, X_2 = x_2)$$

beschrieben werden können. Wir beginnen mit der Betrachtung einiger Situationen, die typisch sein können für eine Abhängigkeit der Beobachtungs- bzw. Fehlwahrscheinlichkeiten von der Zielgröße Y, der vollständig beobachtbaren Kovariablen X_1 bzw. dem wahren Wert der unvollständig beobachtbaren Kovariablen X_2.

- In einer Fall-Kontroll-Studie ist es häufig notwendig, unterschiedliche Erhebungsmechanismen für Fälle und Kontrollen zu benutzen. Dies kann zu erheblichen Unterschieden in den Fehlwahrscheinlichkeiten führen. In einer Studie über Risikofaktoren für die Entstehung von Hirntumoren (Schlehofer et al. 1992) konnte z.B. bei über 90% der Fälle die Blutgruppe anhand von Krankenakten ermittelt werden. Bei den Kontrollen war nur eine mündliche Befragung möglich, so daß die entsprechende Information nur bei 54% der Kontrollen vorlag. Auch wenn auf gleiche Erhebungsmechanismen bei Fällen und Kontrollen geachtet wird, sind Unterschiede in den Fehlraten möglich, weil z.B. Fälle sich intensiver mit der Problematik auseinandergesetzt haben, oder umgekehrt, weil Fälle aufgrund ihrer Krankheit weniger belastbar sind.

- Bei der Durchführung von Befragungen und Interviews hängt die Antwortbereitschaft häufig von Alter und Geschlecht der Befragten ab. Wird also X_2 auf diese Weise erhoben, und ist X_1 Alter oder Geschlecht, so hängt die Fehlwahrscheinlichkeit für X_2 von X_1 ab. Ein anderes Beispiel ergibt sich bei der Untersuchung prognostischer Faktoren von Krebspatienten, wo die Bestimmbarkeit eines biochemischen Parameters (X_2) von der Größe oder dem Stadium des Tumors (X_1) abhängt.
- Das Standardbeispiel für eine Abhängigkeit der Fehlwahrscheinlichkeit vom wahren Wert ist die Antwortverweigerung auf „sensible“ Fragen wie Alkoholkonsum, Sexualverhalten oder Einkommen. Eine zweite Klasse von Beispielen ergibt sich bei Variablen, die auf bestimmte Ereignisse in der Vergangenheit zielen. Hier bedingt das Vorliegen eines Ereignisses häufig ein besseres Erinnern oder eine bessere Dokumentation.

Die geschilderten Situationen können natürlich auch in Kombination auftreten. Dabei sind auch Situationen denkbar, in denen sich die Fehlwahrscheinlichkeiten nicht additiv oder multiplikativ aus den Komponenten zusammensetzen, sondern komplexe Interaktionen vorliegen. Würde z.B. das Alter über eine Befragung erhoben, so könnte es sein, daß der Zusammenhang zwischen dem wahren Alter und der Antwortbereitschaft zwischen Männern und Frauen unterschiedlich ist.

Für die Einschätzung der Eigenschaften von Verfahren zur Behandlung fehlender Werte in den Kovariablen sind zwei verschiedene Bedingungen von zentraler Bedeutung. Bei der ersten Bedingung

$$q(y, x_1, x_2) = q(x_1, x_2) \tag{B1}$$

wird eine Abhängigkeit der Fehlwahrscheinlichkeiten von der Zielgröße ausgeschlossen. Diese Bedingung stellt sicher, daß

$$\mathcal{D}(Y|x_1, x_2, R_2 = 1) = \mathcal{D}(Y|x_1, x_2) \ ,$$

daß also die Tatsache einer vollständigen Kovariableninformation nicht das Regressionsmodell ändert. Die zweite Bedingung

$$q(y, x_1, x_2) = q(y, x_1) \tag{B2}$$

schließt eine Abhängigkeit der Fehlwahrscheinlichkeit vom wahren Wert der Kovariablen aus. Diese Bedingung stellt sicher, daß

$$\mathcal{D}(X_2|y, x_1, R_2 = 0) = \mathcal{D}(X_2|y, x_1, R_2 = 1) \ .$$

Damit ist es also möglich, die Verteilung der unbeobachteten Werte von X_2 aus der Verteilung der beobachteten Werte zu schätzen. Sie ist deshalb die Grundlage aller Verfahren, die die Information der Beobachtungseinheiten mit unvollständiger Kovariableninformation nutzen. Sie wird in der Literatur als *Missing at Random* (MAR) bezeichnet (Rubin 1976).

3. Die Complete Case Analyse

Die Complete Case Analyse, d.h. die Einschränkung der Auswertung auf Beobachtungseinheiten mit vollständiger Kovariableninformation, ist die Standardmethode der meisten statistischen Programmpakete. Als Hauptkritik an ihr wird meistens gesehen, daß sie die in den übrigen Beobachtungeinheiten vorhandene Information vernachlässigt, also nicht effizient ist. Übersehen wird dabei, daß sie auch eine Quelle der Verzerrung von Effektschätzern sein kann. Eine Einschränkung auf Beobachtungseinheiten mit vollständiger Kovariableninformation bedeutet nichts anderes als eine Selektion innerhalb der untersuchten Population, und somit können bei einer Complete Case Analyse all jene unerwünschten Phänomene auftreten, die wir in anderen Zusammenhängen als Selektionseffekte kennen.

Die Effektschätzer einer Complete Case Analyse unterscheiden sich nicht systematisch von denen, die wir bei der Analyse des vollständigen Datensatzes erhalten würden, wenn gilt, daß die bedingte Verteilung von Y gegeben X_1, X_2 unabhängig davon ist, ob wir X_2 beobachten oder nicht. Dies ist erfüllt, wenn die Wahrscheinlichkeit fehlender Werte nur von den Kovariablen selber, aber nicht von der Zielgröße abhängt, also die Bedingung B1 erfüllt ist. Denn dann ändert sich durch die Selektion der Beobachtungseinheiten mit vollständiger Kovariableninformation zwar möglicherweise die Verteilung von (X_1, X_2) in der Population, aber auch nur diese, und nicht das Regressionsmodell. Es ist wichtig zu sehen, daß für eine Complete Case Analyse die MAR Annahme massiv verletzt sein darf. Dies kann keine Quelle der Verzerrung sein, solange wir ausschließen können, daß die Wahrscheinlichkeit des Auftretens fehlender Werte vom Wert der Zielgröße abhängt. Siehe hierzu auch das zweite Beispiel in Kapitel 9.

Bei der Analyse prospektiver Studien, bei denen die Kovariablen X zu einem baseline-Zeitpunkt, die Zielgröße Y aber erst später, z.B. nach Verabreichung einer bestimmten Therapie, gemessen wird, können wir im allgemeinen einen direkten Bezug zwischen dem Wert der Zielgröße und dem Auftreten eines fehlenden Wertes in den Kovariablen ausschließen. Anders ist dies bei retrospektiven Studien, inbesondere bei Fall-Kontroll-Studien. Bei letzteren führen unterschiedliche Erhebungsmechanismen für Fälle und Kontrollen dazu, daß im allgemeinen die Fehlraten zwischen Fällen und Kontrollen unterschiedlich sind. Sofern die Fehlraten nur vom Fall-Kontroll-Status abhängen, ist dies nicht schädlich, da dies nur eine verzerrte Schätzung des Intercepts, nicht aber der Effekte selbst, bedeutet. Allgemein ist eine substantielle Verzerrung nur dann zu erwarten, wenn die Fehlwahrscheinlichkeit sowohl mit der Zielgröße, als auch mit einer oder mehreren Kovariablen korreliert, so daß durch die Selektion der Complete Case Analyse auch die Assoziation zwischen Zielgröße und Kovariable verändert wird. Eine genauere Untersuchung zu diesen Möglichkeiten einer verzerrten Effektschätzung findet sich in Vach & Blettner (1991).

4. Die grundlegende Problematik bei der Ausnutzung von Beobachtungseinheiten mit fehlender Kovariableninformation

Wir haben bereits gesehen, daß die Bedingung B1 sicherstellt, daß sich für Beobachtungseinheiten mit vollständiger Kovariableninformation eine einfache Beziehung zwischen dem „wahren" Regressionsmodell $\mathcal{D}(Y|x_1, x_2, R_2 = 1)$ und dem zu analysierenden Regressionsmodell $\mathcal{D}(Y|x_1, x_2)$ ableiten läßt. Für Beobachtungseinheiten mit unvollständiger Kovariableninformation ist das „wahre" Regresionsmodell

$$\mathcal{D}(Y|x_1, R_2 = 0) = \int \mathcal{D}(Y|x_1, x_2, R_2 = 0) \cdot f(x_2|x_1, R_2 = 0)\, dx_2$$

wobei $f(x_2|x_1, R_2 = 0)$ die bedingte Dichte von X_2 gegeben $X_1 = x_1$ und $R_2 = 0$ bezeichnet. Die Bedingung B1 erlaubt es, $\mathcal{D}(Y|x_1, x_2, R_2 = 0)$ durch $\mathcal{D}(Y|x_1, x_2)$ zu ersetzen. Da nun zum einen das „wahre" Regressionsmodell aber auch von der bedingten Verteilung von X_2 gegeben X_1 bei Nichtbeobachtbarkeit von X_2 abhängt, zum anderen über diese Verteilung keine Information direkt aus den Daten zu gewinnen ist, müssen wir zusätzlich verlangen, daß

$$q(x_1, x_2) = q(x_1)\ , \tag{B3}$$

daß also neben der Bedingung B1 auch die MAR-Annahme erfüllt ist. Denn B3 impliziert, daß

$$f(x_2|x_1, R_2 = 0) = f(x_2|x_1, R_2 = 1) = f(x_2|x_1)\ ,$$

so daß Beobachtungseinheiten mit vollständiger Kovariableninformation genutzt werden können, um die im „wahren" Regressionsmodell der Beobachtungseinheiten mit unvollständiger Kovariableninformation auftretenden mischenden Verteilungen schätzen zu können. Wir erhalten somit unter B3

$$\mathcal{D}(Y|x_1, R_2 = 0) = \mathcal{D}(Y|x_1) = \int \mathcal{D}(Y|x_1, x_2) \cdot f(x_2|x_1)\, dx_2\ .$$

Die verschiedenen Verfahren zur Behandlung von unvollständiger Kovariableninformation unterscheiden sich nur darin, wie mit der Notwendigkeit der Schätzung von $f(x_2|x_1)$ umgegangen wird. Die zuerst betrachteten Verfahren der Auslassung einer Kovariablen, der Behandlung fehlender Werte als eigene Kategorie und der Einsetzung von Mittelwerten versuchen die Notwendigkeit dieser Schätzung zu umgehen, was, zumindest teilweise, ihre unbefriedigenden Eigenschaften erklärt.

5. Ad hoc Methoden

Als ad hoc Methoden bezeichnen wir solche Ansätze, die durch künstliche Manipulation auf die Konstruktion eines vollständigen Datensatzes zielen, der dann mit Standardsoftware ausgewertet werden kann. Die Complete Case Analyse ist eine derartige Methode, hier ist die Manipulation das Weglassen der Beobachtungen mit fehlenden Werten. Das Weglassen der Variablen mit fehlenden Werten ist eine zweite. Etwas sinnvoller erscheint das Ersetzen fehlender Werte durch Mittelwerte oder andere Schätzer; ein ähnliches Ziel verfolgt die Behandlung fehlender Werte als eigene Kategorie.

In diesem Abschnitt gehen wir von generalisierten linearen Modellen ohne Interaktionstermen aus, d.h. es liegt eine Spezifikation des Erwartungswert von Y gemäß

$$E[Y|X_1 = x_1, X_2 = x_2] = \mu(\beta_0 + \beta_1 x_1 + \beta_2 x_2)$$

mit bekannter Linkfunktion μ vor, so daß β_1 den adjustierten Effekt von X_1 und β_2 den adjustierten Effekt von X_2 beschreibt. Die Aussagen gelten aber auch sinngemäß für das Cox-Modell.

5.1 Weglassen von unvollständig beobachteten Kovariablen

Während bei der Complete Case Analyse Beobachtungseinheiten mit fehlenden Werten weggelassen werden, ist es auch denkbar, sich dafür zu entscheiden, die Kovariable mit fehlenden Werten wegzulassen. Damit erhalten wir dann aber einen nicht adjustierten Schätzer für den Effekt von X_1. Sofern $\beta_2 \neq 0$ ist, die zweite Kovariable also einen zusätzlichen Einfluß aufweist, sind adjustierte und nicht adjustierte Effekte im allgemeinen verschieden. Wir wollen auch kurz in Erinnerung rufen, daß in verallgemeinerten linearen Modellen die Unabhängigkeit der Kovariablen nicht ausreicht, um Gleichheit von adjustiertem und nicht–adjustiertem Effekt zu sichern (Gail et al. 1984). Somit stellt dieses Verfahren im allgemeinen keine adäquate Lösung dar. Es ist jedoch wichtig zu sehen, daß jeder allzu naive Ansatz, Beobachtungseinheiten mit $R_2 = 0$ zu nutzen, der Gefahr unterliegt, teilweise nicht-adjustiert zu schätzen, da $\mathcal{D}(Y|x_1, R_2 = 0) = \mathcal{D}(Y|x_1)$ bei Gültigkeit von B1.

5.2 Behandlung fehlender Werte als eigene Kategorie

Bei kategoriellen Variablen ist es eine weitverbreitete Methode, fehlende Werte als eigene Kategorie zu behandeln. Sofern wir von der Gültigkeit der Bedingung B3 ausgehen, ist dies jedoch im allgemeinen eine wenig geeignete Strategie, den adjustierten Effekt der ersten Kovariablen zu schätzen. Die Behandlung fehlender Werte als eigene Kategorie bedeutet nämlich nichts anderes, daß wir für Beobachtungseinheiten mit $R_2 = 0$ den Erwartungswert gemäß

$$E[Y|X_1 = x_1, R_2 = 0) = \mu(\beta_0 + \beta_1 x_1 + \beta_2^*)$$

ansetzen; wir erlauben also eine Verschiebung des Intercepts, aber β_1 modelliert hier den nicht-adjustierten Effekt. Da aber für Beobachtungeinheiten mit $R_2 = 1$ der *gleiche* Parameter β_1 den adjustierten Effekt beschreibt, erhalten wir Schätzer für β_1, die irgendwo zwischen dem adjustierten und dem nicht-adjustierten Effekt liegen. Gerade in epidemiologischen Studien können wir also auf diese Weise die angestrebte Adjustierung für Confounder zunichte machen und Effekte über- oder unterschätzen.

Es kann Situationen geben, in denen die Behandlung fehlender Werte als eigene Kategorie sinnvoll ist, nämlich dann, wenn die Tatsache des Fehlens selber prädiktiv ist. Commenges et al. (1992) berichten von einer Studie zum Vergleich verschiedener Tests zur Diagnose von Dementia bei Screening Maßnahmen. Fehlende Werte für die Ergebnisse zweier Tests zeigten hier einen hohen prädikativen Wert, da sie die totale Unfähigkeit der Probanden, den Test zu verstehen, widerspiegelten.

5.3 Imputationsmethoden

Bei Imputationsmethoden wird jeder fehlende Wert durch eine Schätzung ersetzt. Ist also X_2 eine stetige Kovariable, so besteht der einfachste Ansatz darin, jeden fehlenden Wert durch den Mittelwert der beobachteten Werte von X_2 zu ersetzen, also durch einen Schätzer von EX_2. Eine Verfeinerung dieses Ansatzes besteht darin, zuerst für die Beobachtungen mit vollständiger Kovariableninformation eine Regression von X_2 auf X_1 durchzuführen, und anschließend mit den so ermittelten Schätzungen $\widehat{\alpha}$ und $\widehat{\beta}$ fehlende Werte in X_2 durch die Prädiktionen $\widehat{\alpha} + \widehat{\beta} x_1$ zu ersetzen, also durch einen Schätzer für $E[X_2|X_1 = x_1]$. (Genauer stellt die Bedingung B3 sicher, daß die benutzten Schätzer tatsächlich Schätzer für $E[X_2|R_2 = 0]$ bzw. $E[X_2|X_1 = x_1, R_2 = 0]$ sind.)

Betrachten wir zunächst einmal die klassische Gaußsche Regression mit

$$E[Y|X_1 = x_1, X_2 = x_2] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \ .$$

Hier gilt

$$E[Y|X_1 = x_1] = \beta_0 + \beta_1 x_1 + \beta_2 E[X_2|X_1 = x_1] \ .$$

Dies bedeutet, daß, wenn wir fehlende Werte durch Schätzungen für $E[X_2|X_1 = x_1]$ ersetzen, die Erwartungswertstruktur des Regressionsmodelles annähernd richtig spezifiziert wird, so daß die Schätzer der Regressionsparameter konsistent sind (Gill 1986). Sofern X_1 und X_2 nicht als unabhängig vorausgesetzt werden können, ist die Einsetzung von Schätzern für EX_2 dagegen nicht adäquat und führt mit wachsender Korrelation der Kovariablen zu einer erheblichen Verzerrung der Parameterschätzer. Während also die Erwartungswertstruktur durch Einsetzung geeigneter Schätzer annähernd korrekt spezifiziert werden kann, ist die Varianzstruktur im allgemeinen fehlspezifiziert. Gehen wir nämlich davon aus, daß $Var(Y|X_1 = x_1, X_2 = x_2) = \sigma^2$, so ist $Var(Y|X_1 = x_1) > \sigma^2$, da nun Y einer Mischungsverteilung genügt. Im Rahmen der Kleinst-Quadrat-Schätzung kann versucht werden, dies durch eine entsprechend geringere Gewichtung der Beobachtungseinheiten mit fehlenden Werten auszugleichen (Dagenais 1973, Beale & Little 1975, Gill 1986). Siehe hierzu auch die Überblicksarbeit von Little (1992).

In verallgemeinerten Regressionsmodellen mit einer Linkfunkton, die nicht die Identität ist, gilt im allgemeinen

$$E[Y|X_1 = x_1] \neq \mu(\beta_0 + \beta_1 x_1 + \beta_2 \cdot E[X_2|X_1 = x_1]) ,$$

so daß eine Ersetzung fehlender Werte durch einen Schätzer für $E[X_2|X_1 = x_1]$ zu inkonsistenten Schätzern der Regressionsparameter führt. Sofern β_2 nicht zu groß ist, ist $\mu(\beta_0 + \beta_1 x_1 + \beta_2 \cdot E[X_2|X_1 = x_1])$ jedoch eine recht gute Approximation an $E[Y|X_1 = x_1]$, so daß die Verzerrung möglicherweise relativ klein ist. Für den Spezialfall der logistischen Regression mit zwei binären Kovariablen konnte dies durch entsprechende Berechnungen der asymptotischen Verzerrung bestätigt werden (Vach 1994).

Neben einer möglichen Verzerrung der Schätzer ist ein zweites Problem die inkorrekte Schätzung der Varianz der Parameterschätzer. Diese kann zum einen dadurch entstehen, daß die Annahme einer homoskedastischen Varianz verletzt wird, wie oben für die Gaußsche Regression beschrieben. Zum anderen entsteht durch die Einsetzung von Schätzern für die bedingten Erwartungswerte eine zusätzliche Variabilität, die bei der Varianzschätzung nicht berücksichtigt wird. Bei der logistischen Regression tritt nur der zweite Fall auf, da hier die Varianz eine Funktion des Erwartungswertes ist. Im Spezialfall zweier binärer Kovariablen zeigten Simulationsstudien, daß eine substantielle Unterschätzung der Varianz nur bei extremen Fehlraten auftritt.

Bei diesen, teilweise recht vorteilhaften Ergebnissen, ist allerdings zu beachten, daß die bedingten Erwartungswerte auch tatsächlich konsistent geschätzt werden müssen. Da der Zusammenhang zwischen X_1 und X_2 nicht notwendigerweise linear ist, ist eine einfache Regression von X_1 auf X_2 möglicherweise nicht ausreichend. Daher sollte zumindest noch ein quadratisches Glied benutzt werden. Imputationsmethoden lassen sich auch in einfacher Weise bei kategoriellen Variablen benutzen, indem eine Dummykodierung benutzt wird. Bei binären Variablen wird in naheliegender Weise von „Probability Imputation“ (Schemper & Smith 1990) gesprochen. Imputationsmethoden lassen sich in naheliegender Weise auf beliebig viele Kovariablen und beliebige Muster der Fehlstellen verallgemeinern.

5.4 Eine Illustration zur Gefahr einer verzerrten Schätzung durch ad-hoc-Methoden

Wir möchten noch einmal kurz an einer kleinen beispielhaften Berechnung zeigen, in welcher Größenordnung die durch die Benutzung einer ad-hoc-Methode entstehende Verzerrung sein kann. Dazu betrachten wir ein logistisches Regressionsmodell $P(Y = 1|X_1 = x_1, X_2 = x_2) = \Lambda(\beta_0 + \beta_1 x_1 + \beta_2 x_2)$ mit zwei binären Kovariablen. Wir gehen davon aus, daß 30% der Beobachtungen von X_2 zufällig fehlen, also $q(y, x_1, x_2) \equiv 0.7$, und vergleichen die asymptotische Verzerrung bei der Schätzung von β_1 für die vier Methoden des Weglassens von X_2, der Behandlung fehlender Werte als eigene Kategorie, sowie der Ersetzung fehlender Werte durch die relative Häufigkeit von X_2 bzw. der bedingten relativen Häufigkeit gegeben X_1. In Abbildung 5.1 ist die asymptotische Verzerrung von $\widehat{\beta}_1$ für den Fall balancierter Kovariablen in Abhängigkeit der Odds Ratios von X_1 und X_2 aufgezeigt, wobei $\beta_1 = 1.0$ und $\beta_2 = 1.5$ ist, X_2 also einen großen Einfluß besitzt. Deutlich wird, daß das Weglassen der Kovariablen X_2 zur größten Verzerrung führt, während die Verzerrung beim Einsetzen bedingter relativer Häufigkeiten sehr gering ist. Die Einsetzung der einfachen relativen Häufigkeit oder die Behandlung fehlender Werte als eigene Kategorie führt zu annähernd der gleichen Verzerrung; beide Verfahren stimmen ja auch darin überein, bei der Suche nach einem Ersatz für einen fehlenden Wert die notwendige Berücksichtigung von X_1 zu vernachlässigen. Bei beiden Ansätzen wird schon bei einem OR von 9.0 das wahre β_1 um 22% überschätzt. (Ein OR von 9.0 entspricht einer Korrelation von r=0.5.) Weitere Berechnungen zeigen, daß bei einer Fehlrate von 50% die Überschätzung 35% beträgt, bei einer Fehlrate von 70% sogar 47% beträgt.

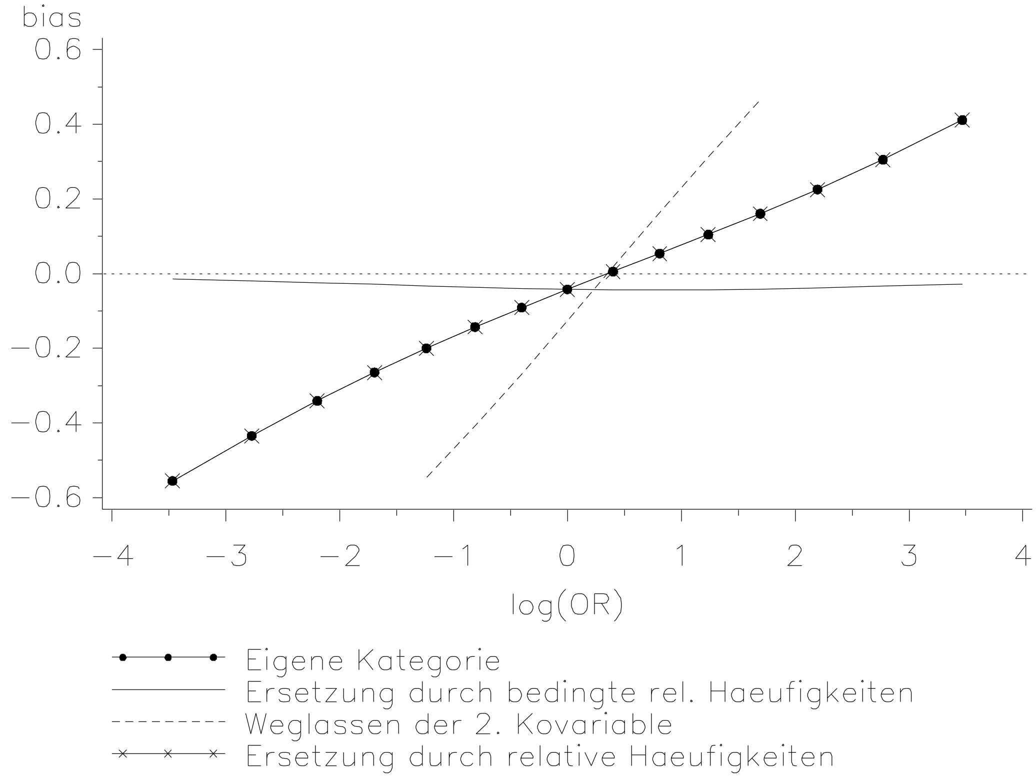

Abbildung 5.1: Asymptotische Verzerrung der Schätzung von β_1 in Abhängigkeit des logarithmierten Odds Ratios zwischen X_1 und X_2. Der Interceptparameter ist so gewählt, daß $P(Y = 1) = 0.5$ gilt.

6. Neuere Ansätze

In den letzten Jahren sind verschiedene neue Ansätze zur Behandlung fehlender Werte entwickelt worden. Gemeinsam ist ihnen, daß sie durch die Benutzung etablierter statistischer Prinzipien wie dem Maximum Likelihood Prinzip sicherstellen, daß die resultierenden Schätzer für die Regressionsparameter unter der Annahme B3 konsistent sind, und somit das wesentliche Problem der ad-hoc-Verfahren umgangen ist. Innerhalb dieser neuen Verfahren bilden dann Überlegungen zur Effizienz und Praktikabilität die entscheidenen Kriterien für eine weitere Auswahl.

6.1 Maximum Likelihood Schätzung

Für die Anwendung des Maximum Likelihood Prinzipes müssen wir die Dichte des Regressionsmodelles geeignet parametrisch darstellen. Gehen wir also davon aus, daß die Dichte der bedingten Verteilung um Y gegeben X_1 und X_2 durch $f_\beta(y|x_1, x_2)$ gegeben ist, so impliziert B3 daß

$$f_\beta(y|x_1, x_2, R_2 = 1) = f_\beta(y|x_1, x_2)$$

und

$$f_{\beta,\theta}(y|x_1, R_2 = 0) = \int f_\beta(y|x_1, x_2) \cdot f_\theta(x_2|x_1)\, dx_2 \ ,$$

d.h. wir brauchen zusätzlich eine geeignete Parametrisierung der bedingten Verteilung von X_2 gegeben X_1. Ist eine derartige Parametrisierung gefunden, so können wir β und θ gemeinsam nach dem ML Prinzip schätzen (Ibrahim 1990) und die sich ergebenden Schätzer sind konsistent, die asymptotische Varianz der Schätzer ist schätzbar und die Schätzer von β sind asymptotisch effizient, also nicht mehr verbesserbar. Allerdings ist schon die Konsistenz der Schätzer davon abhängig, daß das Modell für die bedingte Verteilung von X_2 gegeben X_1 richtig spezifiziert ist; liegt hier eine Fehlspezifikation vor, so kann dies zu verzerrten Schätzern der Regressionsparameter führen. Als weitere Schwierigkeit ergibt sich, daß die Likelihood ein Integral enthält, was unter Umständen spezielle numerische Methoden der Maximierung erforderlich macht.

Sind X_1 und X_2 kategoriell, so ist die Wahl einer Parametrisierung nicht schwierig, wir können direkt die bedingten Wahrscheinlichkeiten $\theta_{jk} = P(X_2 = k|X_1 = j)$ benutzen. Zusätzlich wird aus dem Integral eine Summe, so daß die Likelihood analytisch behandelbar ist und numerische Verfahren wie Newton-Raphson benutzt werden können, die uns über die Matrix der zweiten Ableitungen auch einen Varianzschätzer liefern. Sind X_1 oder X_2 stetig, so ist jede Parametrisierung restriktiv und somit potentiell gefährlich hinsichtlich einer Verzerrung der Schätzung von β. Zusätzlich ergeben sich bei stetigem X_2 Probleme bei der numerischen Berechnung des ML-Schätzers. Zwar kann hier der EM-Algorithmus, das Standardwerkzeug für die ML-Schätzung bei fehlenden Werten (Dempster, Laird & Rubin 1977), benutzt werden, jedoch erfordert jeder E-Step unter Umständen numerische Integration.

Im Falle mehrerer Kovariablen mit beliebigen Mustern der Fehlstellen ist es notwendig, die gemeinsame Verteilung der Kovariablen zu spezifizieren. Sofern stetige Kovariablen auftreten, erscheint es nahezu unmöglich, eine multivariate Verteilung zu spezifizieren, die ausreichende Flexibilität besitzt, um die praktisch relevanten Formen der marginalen Verteilung (Schiefe, Bimodalität) als auch der Abhängigkeiten (nicht-linear, Interaktionen) abzudecken. Nur wenn alle Kovariablen kategoriell sind, erscheint eine nicht allzu restriktive Modellierung, z.B. über loglineare Modelle, möglich (Vach & Blettner 1995).

6.2 Schätzung der Likelihood

Um parametrische Annahmen über die bedingte Verteilung von X_2 gegeben X_1 zu vermeiden, ist eine naheliegende Idee, diese Verteilungen nichtparametrisch zu schätzen, und dann die geschätzte Likelihood $\widehat{L}(\beta)$ basierend auf $\widehat{f}_\beta(y|x_1, R_2 = 0) = \int f_\beta(y|x_1, x_2)\, d\widehat{F}^{X_2|X_1=x_1}$ zu betrachten. Pepe & Fleming (1991) betrachten den Fall eines kategoriellen X_1, so daß $\widehat{F}^{X_2|X_1=x_1}$ gerade die empirischen Verteilungen von X_2 in den durch X_1 definierten Schichten sind. Carroll & Wand (1992) betrachten den Fall eines stetigen X_1, so daß die Verwendung von Kernschätzern oder ähnlichen Methoden nötig wird. In beiden Fällen werden Schätzer für die Varianz von $\widehat{\beta}$ hergeleitet, die aber komplexere Berechnungen erfordern. Der wesentliche Nachteil dieses Ansatzes besteht darin, daß er für den Fall mehrerer Kovariablen mit beliebigem Muster von fehlenden Werten kaum praktikabel ist, da nun nichtparametrische Schätzer für multivariate, bedingte Verteilungen mit mehrdimensionalen Bedingungen benötigt werden.

6.3 Schätzung der Scorefunktion

Bei einer Beobachtungseinheit mit $R_2 = 0$ ist der Beitrag $S_\beta(y, x_1, x_2)$ zur Scorefunktion unbekannt. Es ist jedoch für die Konsistenz der Schätzer von β unschädlich, den unbekannten Beitrag durch $E[S_\beta(y, x_1, X_2)|Y = y, X_1 = x_1]$ zu ersetzen. Dann ist es naheliegend, diesen Erwartungswert durch Methoden der nichtparametrischen Regression aus den Beobachtungen mit $R_2 = 1$ zu schätzen. Dieser Ansatz wurde bisher nur für den Spezialfall, in dem Y und X_1 kategoriell sind, untersucht (Reilly & Pepe 1995), indem er jedoch äquivalent zum Ansatz des nächsten Abschnittes ist. Es ist etwas einfacher auf den Fall mehrerer Kovariablen mit beliebigem Muster der Fehlstellen zu verallgemeinern, da die bedingten Erwartungswerte komponentenweise geschätzt werden; es verbleibt nur das Problem der Wahl eines geeigneten Abstandsmaßes bei mehrdimensionalen Bedingungen.

6.4 Gewichtung reziprok zur Beobachtungswahrscheinlichkeit („Weighting“)

Aus der Stichprobentheorie ist es wohlbekannt, daß bei einer Stichprobe mit unterschiedlicher Auswahlwahrscheinlichkeit je Objekt der Grundgesamtheit die adäquate Schätzung des Mittelwertes der Grundgesamtheit den Horvitz-Thompson Schätzer erfordert, bei dem jedes Element der Stichprobe umgekehrt proportional zur Auswahlwahrscheinlichkeit gewichtet wird (Cochran 1977, S. 259). Diese Idee läßt sich hier auf die Scorefunktion anwenden, d.h. nur Beobachtungseinheiten mit $R_2 = 1$ tragen zur Scorefunktion bei, diese werden aber gewichtet reziprok zu $\widehat{q}(Y, X_1)$, wobei $\widehat{q}$ ein Schätzer für die Beobachtungswahrscheinlichkeit ist, der z.B. durch logistische Regression von R_2 auf Y und X_1 gewonnen werden kann. Die eigentliche Berechnung von $\widehat{\beta}$ kann dann durch Standardsoftware erfolgen, sofern diese die Angabe beliebiger Gewichte erlaubt.

Es läßt sich einfach zeigen, daß die resultierenden Schätzer für $\widehat{\beta}$ konsistent sind, sofern $\widehat{q}$ konsistent geschätzt wird. Die Schätzer für die Varianz von $\widehat{\beta}$, die sich bei der gerade beschriebenen Benutzung von Standardsoftware ergeben, sind im allgemeinen zu klein, und die Unterschätzung ist wesentlich massiver als bei Imputationmethoden. Sofern $q(y, x_1)$ durch parametrische Modelle geschätzt wird, sind korrekte Varianzschätzer einfach herleitbar, aber nicht mit Standardsoftware berechenbar. Für den Spezialfall daß Y und X_1 kategoriell sind finden sich entsprechende Ergebnisse bei Flanders & Greenland (1991) und Pepe & Reilly (1995).

Der Ansatz läßt sich in naheliegender Wiese auf die Situation mehrerer Kovariablen mit beliebigem Muster der Fehlstellen verallgemeinern. Ein wesentlicher Nachteil besteht darin, daß er im Vergleich zu einer ML-Schätzung einen hohen Effizienzverlust aufweist, wenn nicht $q(y, x_1, x_2) \equiv q$ angenommen werden kann, und daß er unter Umständen sogar weniger effizient als eine Complete Case Analyse sein kann (Vach 1994, S. 46).

6.5 Semiparametrisch effiziente Schätzung

Die Behandlung fehlender Werte in den Einflußgrößen bei Regressionsmodellen ist ein Problem, welches mit der Theorie der semiparametrischen Modelle (Bickel et al. 1994) behandelt werden kann, da wir das Regressionsmodell parametrisch angeben, die Verteilung der Kovariablen aber unspezifiziert lassen wollen. Innerhalb dieser Theorie lassen sich für unser Problem (unter der Bedingung B2) konsistente Schätzer herleiten, die im folgenden Sinne effizient sind: Die asymptotische Varianz ist genauso groß wie diejenige der „ungünstigsten“ parametrischen Modellierung $f_\theta(x_2|x_1)$, d.h. derjenigen mit größter asymptotischer Varianz von $\widehat{\beta}$.

Robins et al. (1994) zeigen, daß ein derartiger Schätzer als Lösung einer Schätzgleichung gewonnen werden kann. Dabei ist der Beitrag einer einzelnen Beobachtungseinheit

$$S_\beta(Y, X_1, X_2) = \frac{h(X_1, X_2)(y - \mu_\beta(x_1, x_2))}{q(Y, X_1)} - \frac{\phi(Y, X_1)(1 - q(Y, X_1))}{q(Y, X_1)}$$

falls $R_2 = 1$, und

$$S_\beta(Y, X_1) = -\frac{\phi(Y, X_1)(0 - q(Y, X_1))}{q(Y, X_1)}$$

falls $R_2 = 0$. Diese Schätzgleichungen kombinieren also einen Beitrag, der auf der Differenz zwischen Y und seinem Erwartungswert basiert und wie wir ihn aus den generalisierten linearen Modellen kennen, mit einem Beitrag, der in ähnlicher Weise R_2 mit seinem Erwartungwert $q(Y, X_1)$ in Verbindung setzt, und zusätzlich wird ähnlich wie beim Weighting-Ansatz reziprok mit der Beobachtungswahrscheinlichkeit gewichtet. Die Schätzgleichungen lassen sich einfach durch numerische Verfahren wie Newton-Raphson lösen und die asymptotische Varianz durch einen Sandwich-Estimator schätzen. Die Funktion $\phi(y, x_1)$ hängt direkt von $h(x_1, x_2)$ ab, die Funktion $h(x_1, x_2)$ hat aber nicht die übliche Gestalt $\frac{\partial}{\partial\beta}\mu_\beta(x_1, x_2)V^{-1}(\mu_\beta(x_1, x_2))$, sondern ist wesentlich komplizierter und hängt auch von der wahren Verteilung von (X_1, X_2) ab. Daher ist eine adaptive Prozedur notwendig, indem h zuerst unter einer parametrischen Annahme geschätzt wird, und dann mit dem geschätzten h die Schätzung von β erfolgt. Dabei gilt aber, im Gegensatz zum ML-Ansatz, daß auch bei Fehlspezifikation der gemeinsamen Verteilung von (X_1, X_2), die Schätzung von β konsistent bleibt. Ist die Spezifikation korrekt, so ändert sich durch die Schätzung von h nicht die asymptotische Varianz von $\widehat{\beta}$, wir sind also genauso effizient wie bei a-priori-Kenntnis von h. Ebenso hat die Ersetzung von q durch geeignete Schätzer keinen Einfluß auf die asymptotische Varianz. Eine zusätzliche Schwierigkeit ergibt sich dadurch, daß bei stetigen Kovariablen die Funktion h nur implizit angegeben werden kann. Die zur Lösung dieses Problemes notwendigen numerischen Verfahren sind in Robins et al. (1994) aber prinzipiell schon angegeben. Ebenso wurden dort Verallgemeinerungen für den Fall beliebig vieler Kovariablen und beliebiger Muster von Fehlstellen angegeben.

7. Die Analyse retrospektiver Studien, insbesondere von Fall-Kontroll-Studien

Wir sind bei der bisherigen Betrachtung von der Bedingung B3 ausgegangen. Wie wir schon in Abschnitt 3 gesehen haben, ist bei retrospektiver Erhebung der Daten, insbesondere bei Fall-Kontroll-Studien, zu erwarten, daß ein Zusammenhang zwischen der Zielgröße und der Beobachtungswahrscheinlichkeit besteht; wir können günstigstenfalls von der Bedingung B2 ausgehen. Tatsächlich reicht diese für alle in Kapitel 6 betrachteten Verfahren aus. Bei den Likelihood-basierten Verfahren liegt dies daran, daß für die gemeinsame Verteilung von Y, X_1 und X_2 gilt

$$\begin{aligned} f(y, x_1, x_2 | R_2 = 1) &\propto f_\beta(y|x_1, x_2) \cdot f_\theta(x_2|x_1) \text{ und} \\ f(y, x_1 | R_2 = 0) &\propto \int f_\beta(y|x_1, x_2) \cdot f_\theta(x_2|x_1)\, dx_2 \,, \end{aligned}$$

die Likelihood also die gleiche Form annimmt, die sich bei Betrachtung der bedingten Verteilung der Zielgröße nur unter B3 ergibt.

Es ergibt sich allerdings das Problem, daß unter B2 im Allgemeinen die bedingte Verteilung von X_2 gegeben X_1 nicht mehr direkt aus den Beobachtungseinheiten mit $R_2 = 1$ schätzbar ist, da $\mathcal{D}(X_2|x_1, R_2 = 0) \neq \mathcal{D}(X_2|x_1, R_2 = 1)$. Dies betrifft den Ansatz mit der Schätzung der Likelihood als auch die Imputationsmethoden mit Einsetzung einer Vorhersage $\widehat{X}_2$. Das Problem ist prinzipiell lösbar (Vach & Schumacher 1993), die Imputationsmethode kann aber dennoch zu einer erheblichen Verzerrung führen, so daß sie bei retrospektiven Studien nicht empfehlenswert ist.

Bei der Analyse von Fall-Kontroll-Studien ist weiter zu beachten, daß prospektive und retrospektive Likelihood nicht notwendigerweise übereinstimmen. Allgemein darf die prospektive Likelihood ohne Verlust der Konsistenz der Schätzer benutzt werden, die Benutzung der retrospektiven Likelihood kann aber effizienter sein (Carroll et al. 1995). Schätzmethoden basierend auf der retrospektiven Likelihood werden diskutiert in Breslow & Cain (1988) und Schill et al. (1993).

Es sei daran erinnert, daß bei retrospektiven Studien eine Complete Case Analyse zu verzerrten Schätzern führen kann. Daher ist hier die Benutzung der Methoden aus Kapitel 6 schon zur Vermeidung verzerrter Schätzer notwendig. Das „Weighting" als einfachste dieser Methoden kann aber gerade bei einer Abhängigkeit der Beobachtungswahrscheinlichkeit von der Zielgröße sehr ineffizient im Vergleich zu den anderen Ansätzen sein.

8. Ein wichtiger Spezialfall

Es gibt einen wichtigen Spezialfall, in dem es möglich ist, unter der MAR Annahme einen semiparametrisch effizienten Schätzer und seine Varianz in einfacher Weise zu berechnen. Es handelt sich um die Situation einer logistischen Regression, bei der eine kategorielle Kovariable X_1 bei allen Beobachtungseinheiten beobachtet wird, eine weitere Kovariable X_2 beliebigen Niveaus aber nur an einem Teil der Beobachtungseinheiten. Sollen aus diesen unvollständigen Daten die Parameter des Regressionsmodelles

$$P(Y=1|X_1=j, X_2=x_2) = \Lambda(\beta_0 + \beta_{1j} + \beta_2 x_2)$$

mit der Nebenbedingung $\beta_{1j}=0$ in effizienter Weise geschätzt werden, so ergibt sich für $\widehat{\beta}_2$ der Complete Case Schätzer, und $\widehat{\beta}_{1j}$ und $\widehat{\beta}_0$ können wie folgt aus dem Complete Case Schätzer $\widehat{\beta}^{CC}$ berechnet werden (Cain & Breslow 1988, Carroll et al. 1995):

$$\widehat{\beta}_0 = \widehat{\beta}_0^{CC} + \log\left(\frac{Q_{01}}{Q_{11}}\right); \quad \widehat{\beta}_{1j} = \widehat{\beta}_{1j}^{CC} + \log\left(\frac{Q_{11}Q_{0j}}{Q_{1j}Q_{01}}\right)$$

wobei $Q_{ij} = \frac{M_{ij}}{N_{ij}}$. Dabei ist N_{ij} die Gesamtanzahl der Beobachtungseinheiten mit $Y = i, X_1 = j$ und M_{ij} die Anzahl davon mit vollständiger Kovariableninformation. Die Varianz der Schätzer läßt sich gewinnen aus

$$\begin{aligned}
\widehat{Var}\widehat{\beta}_0 &= \widehat{Var}\widehat{\beta}_0^{CC} + \frac{1}{N_{01}} + \frac{1}{N_{11}} - \frac{1}{M_{01}} - \frac{1}{M_{11}} - \frac{1}{N_{1\cdot}} - \frac{1}{N_{0\cdot}} \\
\widehat{Var}\widehat{\beta}_{1j} &= \widehat{Var}\widehat{\beta}_{1j}^{CC} + \frac{1}{N_{01}} + \frac{1}{N_{11}} + \frac{1}{N_{0j}} + \frac{1}{N_{1j}} - \frac{1}{M_{01}} - \frac{1}{M_{11}} - \frac{1}{M_{0j}} - \frac{1}{M_{1j}}
\end{aligned}$$

Zu beachten ist, daß in diesem Spezialfall auch prospektive und retrospektive Likelihood übereinstimmen.

Y (Fall/Kontrolle)	0		1	
X_1 (Exposition: Bestrahlung)	0	1	0	1
$X_2 = 0$ (keine Östrogentherapie)	32	502	21	264
$X_2 = 1$ (mit Östrogentherapie)	22	150	17	104
$X_2 =?$ (Östrogentherapie unbekannt)	65	788	28	404
Fehlrate	54.6%	54.7%	42.4%	52.3%

Tabelle 9.1: Daten zum ersten Beispiel

9. Zwei Beispiele

Im ersten Beispiel betrachten wir Daten aus der "International Radiation Study on Cervical Cancer Patients" (Boice et al. 1989), einer Fall-Kontroll-Studie, bei der der Zusammenhang zwischen Bestrahlung und dem Auftreten von Brustkrebs bei Patientinnen, die wegen Cervixcarcinoms behandelt wurden, untersucht wurde. Ein möglicher Confounder ist dabei eine Östrogenersatztherapie. Ob eine solche durchgeführt wurde oder nicht, ließ sich jedoch nur bei der Hälfte der Patientinnen ermitteln. Wir haben also hier die Situation, daß der Effekt einer vollständig beobachteten Exposition (X_1) nach einem unvollständig beobachteten Confounder (X_2) adjustiert werden soll.

Die Daten für diese Fragestellung sind in Tabelle 9.1 zusammengefaßt, die Ergebnisse einer Complete Case Analyse sowie einer ML-Schätzung unter Einbeziehung der Subjekte mit unbekanntem Östrogenersatzstatus zeigt Tabelle 9.2. Die Differenz in der Schätzung des Effektes der Exposition ist hier auf die spezifische Konstellation der Fehlraten zurückzuführen; sie unterscheiden sich in Abhängigkeit von der Exposition nur innerhalb der Fälle, nicht innerhalb der Kontrollen. Die Schätzung von β_1 der Complete Case Analyse ist daher verzerrt, sie verleitet irrtümlich dazu über einen möglichen protektiven Effekt der Strahlentherapie nachzudenken. In dieser Berechnung ist aber möglicherweise die Bedingung B2 verletzt, da es einen Zusammenhang zwischen der Durchführung einer Östrogenersatztherapie und ihrer Dokumentation geben kann, so daß auch der ML-Schätzer fragwürdig ist. Diese Situation ist bezeichnend für die Analyse unvollständiger Daten: Ohne Annahmen über den die fehlenden Werte bedingenden Mechanismus ist ein abschließende Analyse nicht möglich.

Das zweite Beispiel greift eine Studie zur Therapie maligner Gliome auf, welche schon in mehreren Arbeiten für illustrative Zwecke eingesetzt wurde (Ulm et al. 1989, Edler and Quintero 1989, Sauerbrei und Schumacher 1992, Schmoor et al. 1993). In der erstgenannten Arbeit wurden unter anderem mit Hilfe des Cox-Modelles der simultane Einfluß verschiedener Faktoren auf die Überlebenszeit untersucht, und dabei unter anderem eine Wechselwirkung zwischen Karnofskyindex und Therapie herausgearbeitet. Wir betrachten hier diejenigen Faktoren, die in der genannten Arbeit für ein Endmodell selektiert wurden.

		Schätzer	Standard-abweichung	95%-Konfidenz-intervall
Complete Case	β_1	-0.175	0.223	[-0.61,0.26]
Analyse	β_2	0.264	0.140	[-0.01,0.54]
Maximum Likelihood	β_1	0.012	0.162	[-0.31,0.33]
Schätzer	β_2	0.264	0.140	[-0.01,0.54]

Tabelle 9.2: Ergebnisse zur Schätzung der Parameter des logistischen Regressionmodelles $P(Y = 1|X_1 = x_1, X_2 = x) = \Lambda(\beta_0 + \beta_1 x_1 + \beta_2 x_2)$ aus den Daten von Tabelle 9.1.

Obwohl das Vorhandensein eines Tumorrest nach der operativen Behandlung vor Beginn der Chemoherapie vermutlich von prognostischem Wert ist, wurde diese Variable aufgrund ihrer hohen Rate fehlender Werte in der Analyse von Ulm et al. nicht berücksichtigt. In Tabelle 9.3 geben wir die Ergebnisse zweier Analysen unter Einbeziehung der Variablen Tumorrest an und stellen sie einer Anlyse unter Auslassung dieser Variablen gegenüber. Die Schätzer für die Effekte der einzelnen prognostischen Faktoren basieren jeweils auf einem logistischen Modell für gruppierte Überlebenszeiten (mit 6, 12, 18 und 24 Monaten als Cutpoints) und wurden nach dem Maximum Likelihood Prinzip bestimmt. In der ersten Analyse werden nur diejenigen 215 der 374 Patienten betrachtet, für die ein Wert für die Variable Tumorrest vorliegt (Complete Case Analyse, vgl. Abschnitt 3.). In die zweite Analyse gehen alle Patienten, also auch jene mit einem fehlenden Wert in der Variablen Tumorrest, ein, und es wird der in Abschnitt 6.1 beschriebene Ansatz benutzt. Die für diese Analyse notwendige Spezifikation der gemeinsamen Verteilung aller Kovariablen wurde durch ein log-lineares Modell mit Wechselwirkungen erster Ordnung realisiert.

Wie zu erwarten, zeichnen sich die Schätzer der Complete Case Analyse aufgrund des geringeren Stichprobenumfanges durch höhere Standardabweichungen aus. Mit Blick auf die beiden Analysen, in die alle Patienten einbezogen werden, ergibt sich nach Adjustierung für die Variable Tumorrest ein wesentlich geringerer Einfluß für den Resektionsstatus, was durch die hohe Korrelation mit der Variablen Tumorrest erklärt werden kann (Odds Ratio 4.5); die Variablen entsprechen sich hier nicht ein-eindeutig, da der Resektionsstatus die Einschätzung des Chirurgen nach der Operation, der Tumorrest aber den Befund im CT widerspiegelt. Auch andere Effektschätzer ändern sich durch die Adjustierung.

Prognostischer Faktor	Kodierung	Complete Case Analyse		Analyse aller Patienten: mit Variable Tumorrest		Analyse aller Patienten: ohne Variable Tumorrest	
Alter	0=≤50; 1=51–60; 2=60<	-0.699	(0.165)	-0.655	(0.128)	-0.597	(0.121)
Therapie	0=BCNU; 1=BCNU+VW26	-1.126	(0.421)	-0.470	(0.302)	-0.505	(0.298)
Karnofsky Index^0	0=<70; 1=70–80; 2=>80	-0.536	(0.249)	-0.261	(0.181)	-0.198	(0.176)
Karnofsky Index^1	0=<70; 1=70–80; 2=>80	0.860	(0.243)	0.577	(0.191)	0.641	(0.187)
Malignitätsgrad	0=3; 1=3-4; 2=4;	-0.572	(0.167)	-0.577	(0.132)	-0.575	(0.130)
Resektion	0=partiell; 1=total	0.230	(0.271)	0.192	(0.212)	0.361	(0.189)
Cortison	0=nein; 1=ja	-0.576	(0.263)	-0.502	(0.200)	-0.609	(0.191)
Epileptische Anfälle	0=nein; 1=ja	0.167	(0.275)	0.442	(0.210)	0.547	(0.199)
Org. Psychosyndrom	0=nein; 1=ja	-0.350	(0.293)	-0.437	(0.239)	-0.409	(0.235)
Bewußtseinsstörung	0=nein; 1=ja	-0.484	(0.346)	-0.325	(0.278)	-0.340	(0.275)
Tumorrest	0=nein; 1=ja	-0.592	(0.301)	-0.556	(0.304)	–	

Tabelle 9.3: Ergebnisse zum zweiten Beispiel. Angegeben sind die Maximum Likelihood Schätzer der Regressionsparameter; in Klammern die geschätzten Standardfehler. Karnofsky Index^0: Karnofsky Index für Patienten mit Therapie BCNU; Karnofsky Index^1: Karnofsky Index für Patienten mit Therapie BCNU+VW26;

Wir haben hier statt dem häufig verwendeten Cox-Modell zur Analyse von Überlebenszeiten ein logistisches Modell mit gruppierten Überlebenszeiten benutzt, um das Maximum-Likelihood-Prinzip anwenden zu können. Gemäß Abbott (1985) können die gewonnen Schätzer als gute Approximation an diejenigen eines Cox Modelles angesehen werden. Im Gegensatz zur Analyse von Ulm et al. haben wir das Kollektiv um die Patienten mit Biopsie verringert, da hier ein nahezu kausaler Zusammenhang mit dem Vorliegen eines Tumorrestes besteht, was die Modellierung der gemeinsamen Verteilung komplizierter macht. Die präsentierten Ergebnisse sollen illustrieren, daß es möglich ist, auch unvollständig beobachtete Kovariablen in Regressionsanalysen einzubeziehen, ohne den Preis der verminderten Schätzgenauigkeit einer Complete Case Analyse bezahlen zu müssen. Für eine adäquate Analyse dieser Daten sind weitergehende Untersuchungen notwendig; in Vach (1996) finden sich zusätzliche Überlegungen zur Modellierung der Verteilung der Kovariablen, zur Abhängigkeit der Ergebnisse von der Gültigkeit der MAR Annahme sowie zum Unterschied zwischen den Ergebnissen der Complete Case Analyse und der 2. Analyse von Tabelle 9.3.

10. Überlegungen zur Auswahl der geeigneten Methode zur Behandlung fehlender Werte

Unsere bisherigen Ausführungen lassen sich wie folgt zusammenfassen: Bei der Wahl der geeigneten Methode sind zwei Fragen von zentraler Bedeutung: 1. Besteht ein direkter Zusammenhang zwischen der Zielgröße Y und der Beobachtbarkeit der Kovariablen? Bei prospektiven Studien können wir diese Frage im allgemeinen verneinen, da die Kovariablen zu einem Baselinezeitpunkt, die Zielgröße jedoch erst zu einem späteren Zeitpunkt gemessen werden. Bei retrospektiven Studien ist dies meistens nicht ausschließbar. 2. Besteht ein Zusammenhang zwischen dem wahren Wert einer Kovariablen und seiner Beobachtbarkeit? Diese Frage ist in der Praxis schwierig zu beantworten; es gibt Situationen wie die Antwortverweigerung, wo offensichtlich ein Zusammenhang vorliegt, und einige wenige, wo er ausgeschlossen werden kann, z.B. wenn eine Variable erst während der Durchführung einer Studie zusätzlich in die Erhebung eingeschlossen wird. Letztendlich sind hier meist nur vage Vermutungen möglich.

Müssen wir bei der zweiten Frage von einem Zusammenhang, d.h. von einer Verletzung der MAR Annahme ausgehen, so bleibt bei einer prospektiven Studie, wo wir die erste Frage verneinen können, wenigstens noch die Complete Case Analyse, um konsistente Schätzer der Regressionsparameter zu erhalten, während bei retrospektiven Studien alle Ansätze fragwürdig sind. Können wir dagegen auf die MAR Annahme bauen, so steht uns das Spektrum der Methoden aus Kapitel 6 zur Verfügung, und zwar sowohl bei prospektiven Studien als auch bei retrospektiven Studien. Da für die meisten dieser Ansätze noch keine entsprechende Software zur Verfügung steht, stellt sich die Frage nach geeigneten Approximationen. Hier kann höchstens eine eingeschränkte Empfehlung für die Ersetzung fehlender Werte durch Vorhersagen unter Einbeziehung der anderen Kovariablen ausgesprochen werden, wobei es aber auch hier zu verzerrten Schätzern und Unterschätzung der Varianz kommen kann, und diese Methode nicht bei retrospektiven Studien benutzt werden sollte. Die Behandlung fehlender Werte als eigene Kategorie ist nur in Ausnahmefällen sinnvoll.

Für das bei der Analyse von Überlebenszeiten häufig verwendete Cox Modell sind zusätzliche Überlegungen notwendig. Zum einen ist es hier denkbar, daß Zusammenhänge zwischen dem Zensierungsmechanismus und dem Mechanismus zur Erzeugung fehlender Werte gibt, wodurch eine zusätzliche Möglichkeit der Verzerrung gegeben ist (Pugh et al. 1993). Zum anderen ist keiner der in Kapitel 6 beschriebenen Ansätze direkt anwendbar, da die Inferenz im Cox-Modell weder auf einer vollständigen Likelihood basiert noch die Scorefunktion sich aus Beiträgen der einzelnen Beobachtungseinheiten additiv zusammensetzt. Daher sind entsprechende Adaptionen notwendig (Lin & Ying 1993, Pugh et al. 1993, Zhou & Pepe 1995, Robins et al. 1994).

11. Ausblick

Um die methodischen Fortschritte bei der Behandlung fehlender Werte in die Praxis umsetzen zu können, ist es notwendig, geeignete Implementierungen den Anwendern zur Verfügung zu stellen. Dabei müssen praxisorientierte Situationen mit mehreren Kovariablen und beliebigen Mustern der fehlenden Werte bearbeitbar sein. Der Erstautor bereitet derzeit ein derartiges Programm für die Durchführung einer logistischen Regression bei fehlstellenbehafteten, kategoriellen Kovariablen vor, wobei das ML-Prinzip benutzt wird. Langfristig erscheint der allgemeine Ansatz von Robins et al. (1994) am vielversprechendsten, da er die Gefahr einer verzerrten Schätzung der Regressionsparameter durch eine Mißspezifikation der Kovariablenverteilung vermeidet, die Behandlung stetiger Kovariablen erlaubt und hinsichtlich der Effizienz andere semiparameterischen Methoden (siehe Abschnitt 6.2 bis 6.4) übertrifft. Allerdings muß immer berücksichtigt werden, daß alle effizienten Ansätze auf der MAR-Annahme beruhen, die in der Praxis häufig zweifelhaft sind. Ein pragmatischer Ansatz zu diesem Problem sind Sensitivitätsanalysen, bei denen eine Schätzung der Regressionsparameter unter verschiedenen Annahmen über die Verletzung der MAR-Annahme durchgeführt wird. Dies erscheint insbesondere dann sinnvoll, wenn Vorinformationen über die Art der Abweichung vorliegen, z.B. läßt sich bei der Frage nach dem Alkoholkonsum annehmen, daß die Fehlwahrscheinlichkeit mit wahrem Wert ansteigt. Die resultierende Variation der Schätzer erlaubt es dann zu beurteilen, wie sich eine mögliche Verletzung der MAR-Annahme auswirkt. Beispiele für dieses Vorgehen finden sich in Vach & Blettner (1995).

Es gibt allerdings Bereiche, in denen die MAR-Annahme auf jeden Fall erfüllt ist, nämlich dann, wenn von vorneherein eine Kovariable nur für einen Teil der Beobachtungseinheiten erhoben wird. Dies kann dann sinnvoll sein, wenn die Messung der Kovariable sehr aufwendig oder teuer ist. Häufig ist es zusätzlich möglich, eine einfacher zu erhebende Surrogatvariable mit einzubeziehen. Z.B. kann bei einer ernährungsepidemiologischen

Studie für einen Teil der Subjekte die exakte aufgenommene Nahrungsmenge mittels eines Wiegeprotokolles erhoben werden, gleichzeitig eine allgemeine Befragung nach den Ernährunggewohnheiten für alle Subjekte durchgeführt werden. Die Methoden des Kapitels 6 sind alle auch in einer solchen Situation anwendbar bzw. teilweise ursprünglich hierfür entwickelt. Werden in einem ersten Schritt die Surrogatvariablen für alle Subjekte erhoben, so ist es möglich, die Stichprobenumfänge für die nachfolgende Erhebung der exakten Messungen nach Kosten-Nutzengesichtspunkten zu optimieren. Insbesondere ist es sinnvoll, bei seltenen Expositionen bzw. schiefen Verteilungen in der zweiten Erhebung gezielt im dünnbesetzen Bereich des Kovariablenraumes zu erheben, wodurch erhebliche Effizienzgewinne im Vergleich zu einer zufälligen Stichprobenziehung möglich sind. Hiermit ergeben sich dann ganz neue Aspekte bei der Planung von Studien.

Es sei abschließend noch einmal betont, daß nicht erwartet werden darf, daß die mathematische Statistik alleine die Probleme der fehlenden Werte lösen kann. Eine adäquate Behandlung fehlender Werte ist nur bei Vorliegen zusätzlicher Kenntnisse über die zugrundeliegenden Mechanismen möglich, und häufig führt der Mangel an diesen Kenntnissen dazu, daß die ursprüngliche Fragestellung nicht mehr adäquat beantwortet werden kann. Deshalb darf das Bestreben der Biometrie nach vollständigen Daten und guter Datenqualität nicht nachlassen, es sei denn der Mechanismus der Entstehung von Fehlstellen ist kontrolliert oder zumindest nachvollziehbar.

Literatur

Abbott, R.D. (1985) *Logistic regression in survival analysis.* American Journal of Epidemiology 121, 465-471

Beale, E.M.L. & Little, R.J.A. (1975) *Missing values in multivariate analysis.* Journal of the Royal Statistical Society B 37, 129-145

Bickel, P.J., Klaassen, C.A., Ritov, Y., Wellner, J.A. (1993) *Efficient and adaptive estimation for semiparametric models.* John Hopkins University Press, Baltimore

Boice, J.D., Blettner, M., Kleinerman, R., et al. (1989) *Radiation dose and breast cancer risk in patients treated for cancer of the cervix.* International Journal of Cancer 44, 7-14

Breslow, N.E. & Cain, K.C. (1988) *Logistic regression for two-stage case-control data.* Biometrika 75, 11-20

Cain, K.C. & Breslow, N.E. (1988) *Logistic regression analysis and efficient design for two-stage studies.* American Journal of Epidemiology 128, 1198-1206

Carroll, R.J. & Wand, M.P. (1991) *Semiparametric estimation in logistic measurement error models.* Journal of the Royal Statistical Society B 53, 573-585

Carroll, R.J., Wang, S.. & Wang, C.Y. (1995) *Prospective analysis of logistic case-control studies.* Journal of the American Statistical Association 90, 157-169

Cochran, W.G (1977) *Sampling techniques.* 3rd edition, Wiley, New York

Commenges, D., Gagnon M., Letenneur, L., Dartigues, J.F., Barbarger-Gateau, P., & Salamon R. (1992) *Improving screening for dementia in the elderly using mini-mental state examination subcores, Benton's visual retention test, and Isaacs' set test.* Epidemiology 3, 185-188

Dagenais, M.G. (1973) *The use of observations in the multiple regression analysis: A generalized least squares approach.* Journal of Econometrics 1, 317-328

Dempster, A.P., Laird, N.M., & Rubin, D.B. (1977) *Maximum likelihood estimation from incomplete data via EM algorithm (with discussion).* Journal of the Royal Statistical Society B 39, 1-38

Edler, L. & Quintero, C. (1991) *Kommentar zu K. Ulm, C. Schmoor, W. Sauerbrei, G. Kemmler, Ü. Aydemir, B. Müller und M. Schumacher (1989).* Biometrie und Informatik in Medizin und Biologie 22, 59-85

Flanders, W.D. & Greenland, S. (1991) *Analytic methods for two-stage case-control studies and other stratified designs.* Statistics in Medicine 10, 739-747

Gail, M.H., Wieand, S., & Piantadosi, S. (1984) *Biased estimation of treatment effect in randomized experiments with nonlinear regressions and omitted covariates.* Biometrika 71, 431-444

Gill, R.D. (1986) *A note on some methods for regression analysis with incomplete observations.* Sankhya B 48, 19-30

Ibrahim, J.G. (1990) *Incomplete data in generalized linear models.* Journal of the American Statistical Association 85, 765-769

Lin, D.Y. & Ying, Z. (1993) *Cox regression with incomplete covariate measurements.* JASA 88, 1341-1349

Little, R.J.A. (1992) *Regression with missing X's: A review.* JASA 87, 1227-1237

Pepe, M.S. & Fleming, T.R. (1991) *A nonparametric method for dealing with mismeasured covariate data.* Journal of the American Statistical Association 86, 108-113

Pugh, M., Robins, J., Lipsitz, S. & Harrington, D. (1993) *Inference in the Cox proportional hazards model with missing covariate data.* Technical report 758Z, Division of Biostatistics, Dana-Farber Cancer Institute, Boston

Reilly, M. & Pepe, M. (1995) *A Mean Score method for missing and auxiliary covariate data in regression models.* Biometrika 82, 299-314

Robins, J.M., Rotnitzky, A. & Zhao, L.P. (1994) *Estimation of regression coeffecients when some regressors are not always observed.* JASA 89, 846-866

Rubin, D.B. (1976) *Inference and missing data.* Biometrika 63, 581-592

Sauerbrei, W. & Schumacher, M. (1992) *A bootstrap resampling procedure for model building: Application to the Cox regression model.* Statistics in Medicine 11, 2093-2109

Schemper, M. & Smith, T.L. (1990) *Efficient evaluation of treatment effects in the presence of missing covariate values.* Statistics in Medicine 9, 777-784

Schill, W., Jöckel, K.H., Drescher, K., & Timm, J. (1993) *Logistic analysis in case-control studies under validation sampling.* Biometrika 80, 339-352

Schmoor, C., Ulm, K., & Schumacher, M. (1993) *Comparison of the Cox model and the* *regression tree procedure in analysing a randomized clinical trial.* Statistics in Medicine 12, 2351-2366

Ulm, K., Schmoor, C., Sauerbrei, W., Kemmler, G., Aydemir, Ü., Müller, B., & Schumacher, M. (1989) *Strategien zur Auswertung einer Therapiestudie mit der Überlebenszeit als Zielkriterium.* Biometrie und Informatik in Medizin und Biologie 20, 171-205

Schlehofer, B., Blettner, M., Becker, N., Martinsohn, C. & Wahrendorf, J. (1992) *Medical risk factors and the development of brain tumor.* Cancer 69, 2541-2547

Vach, W. (1994) *Logistic regression with missing values in the covariates.* Lecture Notes in Statistics 86, New York, Springer

Vach, W. (1996) *Some issues in estimating the effect of prognostic factors from incomplete covariate data.* Statistics in Medicine (eingereicht)

Vach, W. & Blettner, M. (1991) *Biased estimation of the odds ratio in case-control studies due to the use of ad-hoc methods of correcting for missing values for confounding variables.* American Journal of Epidemiology 134, 895-907

Vach, W. & Blettner, M. (1995) *Logistic regression with incompletely observed categorical covariates – Investigating the sensitivity against violation of the missing at random assumption.* Statistics in Medicine 14, 1315-1329

Vach, W. & Schumacher, M. (1993) *Logistic regression with incompletely observed categorical covariates – A comparison of three approaches.* Biometrika 80, 353-362

Zhou, H. & Pepe, M.S. (1995) *Auxiliary covariate data in failure time regression.* Biometrika 82, 139-149

Anschrift der Verfasser: Dr. Werner Vach, Zentrum für Datenanalyse und Modellbildung, Universität Freiburg, Albertstr. 26/28, 79104 Freiburg; Prof. Dr. Martin Schumacher, Institut für Medizinische Biometrie und Medizinische Informatik, Universität Freiburg, Stefan Meier Str. 26, 79104 Freiburg.