Statistische Methoden bei unvollständigen Daten

H. Toutenburg, C. Heumann, T. Nittner

Department für Statistik
Ludwig–Maximilians–Universät München
Ludwigstr. 33, D–80539 München

May 6, 2004

Abstract

Dieser Artikel gibt einen Überblick über die Problematik fehlender Daten im Rahmen der statistischen Datenanalyse. Im Prinzip sollte er auch Lesern mit geringem mathematischen und statistischen Wissen dienlich sein und sie mathematisch nicht überfordern. Gegebenenfalls kann über allzu theoretische Komponenten hinweggelesen werden.

1 Einleitung

Wir weisen ausdrücklich darauf hin, dass die Problematik fehlender Daten ein umfassendes Gebiet ist, das sowohl in andere Gebiete (etwa Informatik, Psychologie) hineinreicht als auch in der Statistik selbst schwer abgrenzbar ist. Ein Versuch der Abgrenzung sollte mit dieser Abhandlung gelungen sein.

Struktur und Inhalt gliedern sich wie folgt: Beginnend mit einem einfachen Beispiel (Abschnitt 1.1) werden zwei zentrale Begriffe bzw. Werkzeuge eingeführt, ohne die eine Betrachtung des Problems fehlender Daten selbst als unvollständig zu bezeichnen ist, die Fehlendmuster (1.2) und die Fehlendmechanismen (1.3). Allgemeine Verfahren zum Umgang mit fehlenden Daten werden in Abschnitt 2 dargestellt, im Mittelpunkt dabei steht insbesondere die Imputation (2.3). Verschiedene Schätzverfahren, etwa die Likelihood–Methodik und die Bayes–Inferenz findet der Leser in Abschnitt 3. Bisher losgelöst von der Spezifikation eines analytischen Modells werden in Abschnitt 4 einige 'gängige' Modelle der Dependenzanalyse vorgestellt. Nach dem linearen Regressionsmodell (4.1) betrachten wir die sehr allgemeine Klasse der nicht– und semiparametrischen Modelle (4.2). Jedes Kapitel endet mit einem Abschnitt 'Hinweise und weiterführende Literatur'.

1.1 Ein einfaches Beispiel

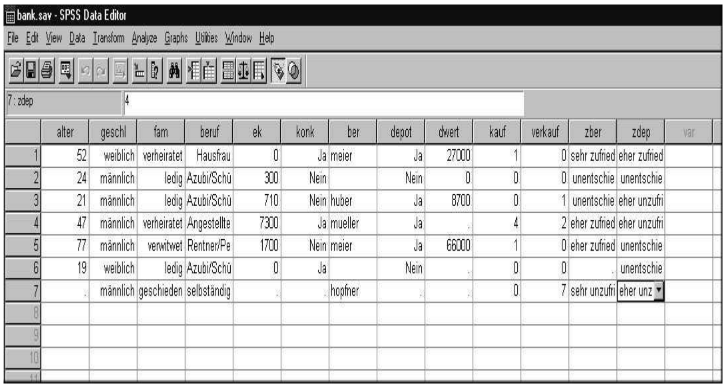
Im Rahmen einer Kundenbefragung einer Bank werden die Kunden gebeten, Auskunft über ihr persönliches Portfolio bzw. ihre Zufriedenheit mit dem Service ihrer Bank zu geben. Neben einigen demographischen Merkmalen, etwa Alter, Geschlecht, Familienstand oder Adresse, gilt es also, relativ spezielle Informationen abzufragen. Das Ziel des Kreditinstituts besteht letztlich darin, Ursachen für Kundenwanderungen frühzeitig identifizieren zu können. Dies können schlecht positionierte Produkte ebenso sein wie unzulängliche Beratung bei komplexeren Produkten. Wir könnten uns also etwa die Erhebung der in Tabelle 1.1 zusammengefassten Merkmale je Kunden vorstellen. Tabelle 1.1 ist sicherlich nicht vollständig und soll nur zur Veranschaulichung von Problemen dienen, die im Zusammenhang mit fehlenden Daten stehen. Unter der Annahme, dass die Kunden über ausführliche Informationen verfügen, was die Definition der Merkmale, deren Skalierung und Kategorisierung betrifft, wäre ein Datensatz, wie er in Abbildung 1.1 im SPSS–Format auszugsweise dargestellt ist, durchaus denkbar. Der hier vorliegende Auszug enthält die Daten von sieben Kunden, anhand derer wir nun die Problematik fehlender Daten bzw. deren potenzieller Ursachen einleitend betrachten wollen. Zunächst stellen wir fest, dass die Kunden 1 und 3 sowie der Rentner/Pensionär in Zeile 5 vollständig geantwortet haben. Der 47–jährige Angestellte hingegen machte keine Angabe zu seinem Depotwert, die Angaben der beiden letzten Kunden sind ebenso unvollständig. Stellen wir nun einige Überlegungen zu

Variablenname	Bedeutung
`alter`	Alter
`geschl`	Geschlecht
`fam`	Familienstand
`beruf`	Beruf
`ek`	monatliches Bruttoeinkommen
`konk`	Kunde auch bei anderer Bank
`ber`	Name des Beraters
`depot`	Depot (Ja/Nein)
`dwert`	Depotwert in EUR
`kauf`	Anzahl der Transaktionen (Kauf) im letzten Geschäftsjahr
`verkauf`	Anzahl der Transaktionen (Verkauf) im letzten Geschäftsjahr
`zber`	Zufriedenheit mit der persönlichen Beratung
`zdep`	Zufriedenheit mit der Entwicklung des Depotbestandes

Table 1.1: Merkmale (Variable) und ihre Bedeutung.

diesen drei Kunden an, so stellen wir fest, dass

- Kunde 2 keine Angabe zu seinem Berater machte,
- Kunde 4 das höchste monatliche Bruttoeinkommen besitzt, seinen Depotwert verschweigt, und sowohl mit der Betreuung wie auch mit der Entwicklung des Depotbestandes eher unzufrieden ist,



bank.sav - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

7 : zdep 4

	alter	geschl	fam	beruf	ek	konk	ber	depot	dwert	kauf	verkauf	zber	zdep	var
1	52	weiblich	verheiratet	Hausfrau	0	Ja	meier	Ja	27000	1	0	sehr zufried	eher zufried	
2	24	männlich	ledig	Azubi/Schü	300	Nein		Nein	0	0	0	unentschie	unentschie	
3	21	männlich	ledig	Azubi/Schü	710	Nein	huber	Ja	8700	0	1	unentschie	eher unzufri	
4	47	männlich	verheiratet	Angestellte	7300	Ja	mueller	Ja	.	4	2	eher zufried	eher unzufri	
5	77	männlich	verwitwet	Rentner/Pe	1700	Nein	meier	Ja	66000	1	0	eher zufried	unentschie	
6	19	weiblich	ledig	Azubi/Schü	0	Ja		Nein	.	0	0	.	unentschie	
7	.	männlich	geschieden	selbständig	.	.	hopfner	.	.	0	7	sehr unzufri	eher unz	
8														
9														
10														

Figure 1.1: Auszugsweiser Datensatz der Kundenbefragung (SPSS–Format).

- Kundin 6 kein Einkommen besitzt, ebenso den Namen ihres Berater nicht nennt, keinen Depotwert angegeben hat und auch bei der Zufriedenheit mit der persönlichen Beratung keine Angabe machte, und schließlich
- Kunde 7 Alter, Einkommen, und Depotwert verschwieg; zudem machte er keine Angabe zur Frage, ob er überhaupt ein Depot besitzt.

Bereits bei der Betrachtung stellt sicherlich ein Großteil der Leser Vermutungen darüber an, warum Werte fehlen und hat sicherlich ad hoc einige mögliche Ursachen parat. Das bedeutet, dass offensichtlich der Datenverlust von Interesse ist und zwar dahingehend, Kausalzusammenhänge zu knüpfen, die das Zustandekommen der fehlenden Angaben mehr oder weniger plausibel begründen. Intuitiv–logische Überlegungen gestatten etwa die Vermutung, dass

- Kunde 2 entweder keinen persönlichen Berater hat oder dessen Namen nicht nennen kann,
- Kunde 4 seinen Depotwert nicht kennt oder nicht nennen will,
- Kundin 6 möglicherweise keinen Berater hat, keinen Depotwert angegeben hat, weil sie kein Depot besitzt und demzufolge auch kein Urteil über die Zufriedenheit mit der Beratung abgeben kann, und
- Kunde 7 möglicherweise lediglich seinen Unmut artikulieren will, ohne durch eventuell gemachte Angaben seine Anonymität zu verlieren.

Bei Kundin 6 könnte man zumindest den Depotwert auf 0 setzen, denn sie hatte angegeben, keines zu besitzen. Sollte sie tatsächlich keinen persönlichen Berater haben, ist der fehlende Wert bei 'zber' kein Informationsverlust und die Tatsache, dass sie eine Beurteilung der Zufriedenheit der Depotentwicklung angegeben hat, obwohl sie keines besitzt, eher ein Problem der Validität der vorhandenen Daten bzw. bei der Spezifikation und Definition der zu erhebenden Variablen bzw. Merkmale. Bei den restlichen Kunden mit unvollständigen Angaben kann man entweder von zufälligen Ursachen sprechen oder von systematischen, was hier bedeutet, dass aufgrund anderer Angaben entsprechend geantwortet wurde. Die in Abschnitt 2.1 noch vorzustellende *complete case analysis*, ein Standardverfahren zum Umgang mit fehlenden Daten, würde bei der Analyse lediglich die Kunden 1, 3 und 5 verwenden, was einem Kundeninformationsverlust von 4/7, also etwa 57% der Daten gleichkäme. Dieser Informationsverlust ist im Allgemeinen schwer erträglich und führt zu verzerrten Resultaten, Interpretationen und Entscheidungen. Eine Identifikation der Ursachen, wenn auch nur schemenhaft, erlaubt jedoch zumindest einen partiellen Informationsgewinn. Wir haben bisher gesehen, dass die Struktur der fehlenden Daten—es fehlen eines oder mehrere Merkmale—ebenso von Bedeutung sein kann wie die Identifikation von Ursachen. In den folgenden beiden Abschnitten wollen wir diese beiden Charakteristiken nun formalisieren und anhand unseres Beispieles veranschaulichen.

1.2 Fehlendmuster

Die sogenannten Fehlendmuster, die *missing data pattern*, sind in der einschlägigen Literatur bekannt und teils in statistischer Software (Solas, SPSS) implementiert. Die Abbildungen 1.2–1.5 repräsentieren verschiedene Fehlendmuster. Abbildung 1.2 ist ein Spezialfall des sogenannten monotonen Patters aus Abbildung 1.3, das oftmals auch durch Vertauschen von Zeilen oder/und Spalten erreicht werden kann. Abbildung 1.4 ist ein Fehlendmuster, das beispielsweise entsteht, wenn zwei Stichproben zusammengesetzt werden. Das ist etwa der Fall, wenn eine Bankfiliale alle Merkmale bis auf den Namen des persön-

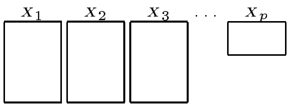


Figure 1.2: Univariates Fehlendmuster.

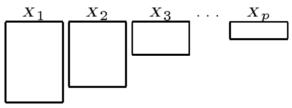


Figure 1.3: Monotones Fehlendmuster.

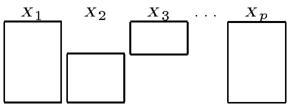


Figure 1.4: Spezielles Fehlendmuster.

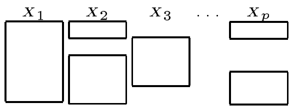


Figure 1.5: Allgemeines Fehlendmuster.

lichen Beraters erfragt, eine andere Filiale alle Merkmale bis auf den Depotwert. Ein Zusammenfügen der beiden Datensätze würde zu Muster 1.4 führen. Schließlich zeigt Abbildung 1.5 ein Muster ohne erkennbare Struktur. Wie zuvor bereits erwähnt, besteht etwa in SPSS die Möglichkeit der Generierung dieser Pattern, jedoch nicht in einer wie hier dargestellten, klaren Form. In Solas, einem Programm speziell zur Analyse unvollständiger Datensätze, besteht hingegen die Möglichkeit eines durchaus anschaulichen Patterns. Fehlendmuster dienen lediglich zur Visualisierung der Fehlendstruktur, geben also Aufschluss über das Ausmaß fehlender Werte in einzelnen Variablen oder Fällen. Sortiert man die Daten beispielsweise der Größe des monatlichen Bruttoeinkommens nach und betrachtet dann die

Fehlendmuster, könnte man—bei größerer Fehlendwahrscheinlichkeit für höhere Einkommen—gar einen potenziellen Kausalzusammenhang vermuten. Höherdimensionale oder komplexere (funktionale) Abhängigkeiten sind jedoch grafisch nicht identifizierbar. Zur Charakterisierung dieser Abhängigkeiten ganz allgemein dienen die sogenannten Fehlendmechanismen, die im nächsten Abschnitt vorgestellt werden.

1.3 Fehlendmechanismen

Die Fehlendmechanismen, auch *missing data mechanisms* genannt, gehen zurück auf Little and Rubin (1987) bzw. deren Neuauflage Little and Rubin (2002). Wir wollen diese nun ganz allgemein definieren und in Anlehnung an unser Beispiel veranschaulichen. Das Ziel besteht letztendlich darin, Abhängigkeiten zwischen den fehlenden und den beobachteten Daten identifizieren zu können. Aufgrund der Tatsache, dass einige Daten nicht vorhanden sind, ist es notwendig, ein Modell zu betrachten, das die Situation a priori charakterisiert: Man fasst demzufolge die Situation als Zufallsexperiment auf, basierend auf den Daten, wie sie beobachtet worden wären und partitioniert diese in einen beobachteten Teil Z_{beob} und einen fehlenden Teil Z_{fehl}. Die Wahrscheinlichkeit dafür, dass ein Wert fehlt ist nun die bedingte Wahrscheinlichkeit einer Zufallsvariablen R ('R' für *reported*) gegeben die Daten ohne fehlende Werte, also $Z = (Z_{\text{beob}}, Z_{\text{fehl}})$. Definieren wir die Zufallsvariable R als binäre

Variable gemäß

$$R = \begin{cases} 1 & \text{falls Wert beobachtet} \\ 0 & \text{falls Wert fehlt} \end{cases} \quad . \tag{1.1}$$

Unser Ziel, die Abhängigkeiten der fehlenden von den beobachteten Daten zu beschreiben, kann nun über die Betrachtung der bedingten Wahrscheinlichkeiten von $P(R \mid Z)$ erfolgen. In Anlehnung an Little and Rubin (2002) unterscheiden wir

1. vollständig zufälliges Fehlen, *missing completely atrandom (MCAR)*, wenn

$$P(R \mid Z_{\text{beob}}, Z_{\text{fehl}}) = P(R)\,, \tag{1.2}$$

2. zufälliges Fehlen, *missing at random (MAR)*, wenn

$$P(R \mid Z_{\text{beob}}, Z_{\text{fehl}}) = P(R \mid Z_{\text{fehl}}) \text{ und} \tag{1.3}$$

3. nicht–zufälliges Fehlen, *missing not at random (MNAR)*, wenn

$$P(R \mid Z_{\text{beob}}, Z_{\text{fehl}}) = P(R \mid Z_{\text{beob}}, Z_{\text{fehl}})\,, \tag{1.4}$$

   wobei hier zumindest die Abhängigkeit von Z_{fehl} gegeben sein muß.

Diesen im Grunde sehr theoretischen Teil wollen wir anhand unseres Beispiels nochmals veranschaulichen. Wenn wir davon ausgehen, dass Kunde 2 den Namen seines Beraters zum Zeitpunkt der Bearbeitung des Fragebogens nicht parat hat, so fehlt der Name des Beraters zufällig und damit gemäß MCAR. Auch wenn alle anderen Kunden ein besseres Gedächtnis haben sollten sprechen wir hier von vollständig zufälligem Fehlen, denn offensichtlich sind wir nicht in der Lage, aus den vorhandenen Daten Gründe für das Fehlen identifizieren zu können. Ein typisches Beispiel für zufälliges Fehlen, also gemäß MAR, könnte das Fehlen des Depotwertes bei Kunde 4 sein; man würde etwa vermuten, dass Kunden mit höherem Einkommen ihren Depotwert bewußt verschweigen. Nicht–zufälliges Fehlen läge etwa dann vor, wenn Kunden aufgrund des hohen Depotwertes diese Angabe verweigern. Im übertragenen Sinne bedeutet MCAR also, dass das Fehlen eines Wertes rein zufällig ist, d.h. dass die fehlenden Daten eine Substichprobe der beobachteten Daten sind. MAR bedeutet, dass das Fehlen eines Wertes von den Werten einer anderen Variablen abhängt; hingegen bedeutet MNAR, dass das Fehlen von der Variablen (bzw. deren Werten) selbst abhängt.

1.4 Hinweise und weiterführende Litaratur

Neben der allgemeinen Problematik fehlender Daten und einem einfachen Beispiel wurden hier zwei wesentliche Werkzeuge vorgestellt—Fehlendmuster und Fehlendmechanismen. Während Fehlendmuster einen ersten Überblick über das Ausmaß des Fehlens innerhalb eines Datensatzes geben können, dienen die Fehlendmechanismen zur Identifikation möglicher Ursachen. Damit eng in Verbindung steht die Qualität statistischer Verfahren bzw. deren Anwendbarkeit. Sie sind zuweilen von großer praktischer Bedeutung, denn teils ist man lediglich daran interessiert, wovon das Fehlen abhängt, ohne an der Datenlage selbst etwas ändern zu wollen, z.B. durch Imputation. Von nicht minderer Bedeutung sind die Fehlendmuster, die bei großen Datenmengen in der Praxis einen einleitenden Überblick verschaffen können. Zudem ist insbesondere die Struktur eines monotonen Musters wie es in Abbildung 1.3 dargestellt ist, in der Theorie der Likelihood–Methodik von Bedeutung, da aufgrund der Monotonie Faktorisierungen der Dichten vorgenommen werden können. Bereits erwähnt wurden Little and Rubin (1987) bzw. deren Neuauflage (Little and Rubin, 2002) als Einführung in die statistische Analyse unvollständiger Datensätze. Speziell bei multivariaten statistischen Modellen empfiehlt sich das Studium von Schafer (1997). Eher praktisch orientiert, primär auf Stichproben und Gewichtungsverfahren abzielend, ist das 2002 erschienene Buch von Groves, Dillman, Eltinge and Little (2002). Als Beispiel für einen weiteren Überblick speziell in der linearen Regressionsanalyse mit unvollständigen Kovariablen ist der Artikel von Little (1992) zu nennen. Ein wichtiger Themenbereich ist die Diagnose von Fehlendmechanismen basierend auf statistischen Tests. Grundlegend ist das Werk von Cohen and Cohen (1983). Im Kontext etwa binärer Variablen kann mithilfe einfacher Binomialtests zumindest auf nicht–MCAR getestet werden, unter Zuhilfenahme von Simulation auch auf non–MAR, siehe dazu Nittner and Toutenburg (2004).

2 Allgemeine Verfahren

In diesem Abschnitt stellen wir die bekanntesten Verfahren zum Umgang mit unvollständigen Datensätzen vor. Er gibt einen detaillierten Überblick über die Einteilung von Methoden zum Umgang mit fehlenden Werten. In Anlehnung an Little and Rubin lassen sich die Verfahren ganz allgemein untergliedern.

Verfahren basierend auf den vollständigen Fällen: Reduzierung der Analyse auf die vollständigen Daten; es werden alle Zeilen entfernt, die mindestens einen fehlenden Wert beinhalten (*complete case analysis*); bei der Analyse der verfügbaren Fälle (*available case analysis*) richtet sich die Fallzahl nach den jeweils in die Analyse mit einbezogenen Variablen.

Verfahren basierend auf Imputation: Ersatzwerte für alle fehlenden Werte; für alle unvollständigen Fälle werden unterschiedliche oder konstante Werte eingesetzt.

Verfahren basierend auf Gewichtung: Verwendung von Gewichten; inverse Auswahlwahrscheinlichkeit im Rahmen einer Zufallsstichprobe ist ein Beispiel.

Verfahren basierend auf Modellen: Definition eines Modells für die fehlenden Daten; Bildung der gemeinsamen Likelihoodfunktion.

In den nächsten beiden Abschnitten gehen wir kurz auf die *complete case analysis* und auf die *available case analysis* ein, die Imputationsverfahren werden in einem eigenen Abschnitt zusammengefasst.

2.1 Complete Case Analysis (CCA)

Die Analyse der vollständigen Fälle, bekannt auch unter dem Begriff *complete case analysis (CCA)*, ist als gängigstes Verfahren im Umgang mit fehlenden Daten zu bezeichnen. SPSS beispielsweise greift automatisch darauf zurück, wenn fehlende Werte vorliegen. Das Verfahren ist auch unter der Bezeichnung *listwise deletion* bekannt, weil sich die Datenmatrix um diejenigen Zeilen reduziert, die einen oder mehrere fehlende Werte enthalten. In unserem Beispiel würde das bedeuten, dass nur die Kunden 1, 3 und 5 mit in die Analyse eingehen. Man kann sich vorstellen, dass dies zuweilen einen großen Informationsverlust mit sich bringen kann. Beispielsweise fehlt bei Kunde 2 der Beratername, was bei der CCA bedeuten würde, dass der Kunde etwa bei einer Schätzung des durchschnittlichen Depotwertes nicht mitberücksichtigt wird. Problematisch ist zudem die durch mögliche Schichtungseffekte einhergehende Verzerrung von Schätzungen. Schafer (1997) quantifiziert die Grenze des Fehlendanteils, bis zu dem die Analyse der vollständigen Fälle ein geeignetes Verfahren darstellt mit etwa 5%. Im Regressionskontext, also einer als abhängige Variable definierte Größe y und einer Matrix von unabhängigen Variablen X, sind die Schätzungen der complete case analysis erwartungstreu, wenn das Fehlen von

Werten in der Kovariablenmatrix X *nicht* von y abhängt.

2.2 Available Case Analysis (ACA)

Ein erster Schritt zur Verminderung des Informationsverlustes durch die CCA besteht in der Verwendung der sogenannten Analyse der verfügbaren Fälle, *available case analysis (ACA)*. Sie unterscheidet sich von der complete case analysis durch die Hinzunahme derjenigen Fälle, die für eine spezielle Analyse—beispielsweise bei der Schätzung des durchschnittlichen Depotwertes—vollständig beobachtet sind. In unserem Beispiel würde also Kunde 2 mit in die Berechnungen einfließen. Interessieren wir uns etwa für die durchschnittliche Zufriedenheit der Kunden mit ihrem Berater, so wird lediglich Kundin 6 von der Analyse ausgeschlossen. Problematisch bei Verwendung der *available case analysis* ist die Möglichkeit unterschiedlicher Stichprobenumfänge. Das ist für die Analyse einzelner Variablen oder Merkmale nicht von Bedeutung; sobald jedoch Maßzahlen verglichen werden, bei denen der Stichprobenumfang mit eingeht, ist mit Verzerrungen zu rechnen. Selbst unter MCAR sind einfache Größen wie etwa Kovarianzen zu modifizieren (siehe (Little and Rubin, 1987)). Die ACA ist im Sinne ihrer Implementierung nicht als praxisrelevant zu bezeichnen, weswegen auf die einschlägige Literatur verwiesen wird. Im nächsten Abschnitt wird Imputation als Verfahren vorgestellt, der Probmelatik fehlender Daten zu begegnen.

2.3 Imputation

Imputation bedeutet, dass für alle fehlende Werte Ersatzwerte eingesetzt werden und—sofern nicht gesondert erwähnt—der auf diese Weise vervollständigte Datensatz analysiert wird wie ursprünglich vorgesehen. Das heißt, dass der Tatsache der Imputation bei der Analyse nicht Rechnung getragen wird. Erneut in Anlehnung an Little and Rubin kann allgemein folgende Unterscheidung vorgenommen werden:

1. *mean imputation:* Ersatzwert Mittelwert, Median oder Modalwert bei metrischem, ordinalem oder nominalem Skalenniveau;

2. *hot deck imputation:* Imputation bereits während der Datenerhebung; Ersetzung durch Untersuchungseinheiten, die den Kriterien der Stichprobe Rechnung tragen, jedoch nicht in die Stichprobe gelangt sind; man beachte, dass die Ersatzwerte auf Untersuchungseinheiten basieren, die sich von den 'Nonrespondents' schon dahingehend unterscheiden mögen, *weil* sie antworten;

3. *cold deck imputation:* Ersatzwert ist ein konstanter Wert aus einer externen Datenquelle;

4. *regression imputation:* Ein Regressionsmodell der fehlenden Variablen auf die beobachteten Variablen für die vollständigen Fälle dient zur Vorhersage;

5. *stochastic regression imputation:* Vorhersagen aus der *regression imputation* werden durch einen Fehlerterm addi-

tiv überlagert;

6. *composite models:* Kombinieren Ideen aus verschiedenen Ansätzen.

7. *multiple imputation:* M–maliges Ersetzen;

Die multiple Imputation wird gesondert im Abschnitt über Inferenz abgehandelt. An dieser Stelle sei lediglich erwähnt, dass über die multiple Imputation versucht wird, den Fehler, den man bei einer einfachen Imputation mit großer Wahrscheinlichkeit begeht, etwas zu mindern. In der Folge wollen wir auf einige der Verfahren näher eingehen.

2.3.1 Mittelwertimputation

Die Mittelwertimputation, auch *zero order regression (ZOR)* oder *unconditional mean imputation* genannt, geht auf Wilks (1932) zurück und ist unter anderem in SPSS im Rahmen der linearen Regressionsanalyse implementiert. In unserem Beispiel fehlen die Depotwerte der Kunden 4, 6 und 7. Das arithmetische Mittel der beobachteten Daten zu 'dwert' ist trivialerweise

$$\frac{1}{4} \cdot (27000 + 0 + 8700 + 66000) = 25425\,.$$

Das würde bedeuten, dass die Kunden 4, 6 und 7 bei Imputation jeweils einen Depotwert von EUR 25425 zugewiesen bekämen. Daran ersehen wir bereits mögliche Fehler: Etwa kann bei Kunde 6 (Azubi/Schüler/Student) nicht davon ausgegangen werden, dass dieser Wert der Realität entspricht. Jedoch ist mit derartig intuitiven Vermutungen äußerst vorsichtig umzugehen. Man kann lediglich behaupten, dass mit großer Wahrscheinlichkeit die eingesetzten Werte nicht der Realität entsprechen. Quantifizieren läßt sich ein anderes Problem—das der Varianzunterschätzung durch die ZOR. Bezeichne $\bar{x}_{\mathrm{CC}}$ das arithmetische Mittel der vollständigen Fälle und $\bar{x}_{\mathrm{imp}}$ das des durch ZOR vervollständigten Datensatzes. Dann gilt für die Schätzungen der Varianz der vollständigen Fälle, mit der Indexmenge $\Phi = 1, 2, 3, 5$, die die Indizes der beobachteten Depotwerte darstellt,

$$\mathrm{Var} = \frac{1}{4} \sum_{i \in \Phi} (x_i - \bar{x}_{\mathrm{CC}})^2 = 643741875.0 \qquad (2.1)$$

und für die Varianz des vervollständigten Datensatzes

$$\mathrm{Var} = \frac{1}{7} \sum_{i=1}^{7} (x_i - \bar{x}_{\mathrm{imp}})^2 = 367852500.0\,. \qquad (2.2)$$

Wir sehen, dass die Varianz des vervollständigten Datensatzes kleiner ist, was offensichtlich ist, denn die Anzahl der in die Varianz eingehenden Boebachtungen wächst um die Anzahl der fehlenden Werte, wohingegen die zugehörigen Summanden Null sind (da $x_i = \bar{x}_{CC} = \bar{x}_{imp} \forall\ i \notin \Phi$). Man spricht demzufolge auch von einer Varianzunterschätzung durch die Mittelwertimputation. Man kann leicht zeigen, dass bei $n - m$ vollständigen und m fehlenden Werten die Varianz des vervollständigten Datensatzes um den Faktor $(n - m)/n$ kleiner ist als die Varianz der vollständigen Fälle. Es sei jedoch erwähnt, dass bei metrischem

Skalenniveau durchaus eine Wahrscheinlichkeit ($\neq 0$) gegeben ist, dass die zu ersetzenden Werte tatsächlich dem Mittelwert entsprechen. Zu beachten ist das Skalenniveau des betreffenden Merkmals. Grundsätzlich unterscheidet man metrisch, ordinal und nominal, wobei ein Merkmal einer höheren Skalierung durch Maßzahlen niedrigerer Skalierung charakterisiert werden darf. Mittelwertimputation muß demzufolge nicht die Ersetzung durch das arithmetische Mittel implizieren. Ein Beispiel für ein *composite model* ist die stochastische Mittelwertimputation, bei der der eingesetzte Mittelwert durch einen zufälligen Fehler additiv überlagert wird. Im Falle nicht–metrischer Skalierung ist darauf zu achten, dass aus Gründen der Interpretierbarkeit Kategorien 'imputiert' werden, die tatsächlich auch definiert sind.

2.3.2 Bedingte Mittelwertimputation

Die bedingte Mittelwertimputation, auch *first order regression (FOR)*, *conditional mean imputation* oder ganz allgemein *regression imputation* genannt, benötigt ein Hilfs(regressions)modell zur Ersetzung der fehlenden Werte. Die Idee besteht darin, den jeweils fehlenden Wert basierend auf einem Hilfsregressionsmodell vorherzusagen; dabei nutzt man die Schätzungen eines Modells, bei dem die unvollständige Variable auf zu bestimmende unabhängige Variablen regressiert wird, ausgehend von der Stichprobe der vollständigen Fälle. Man könnte im übertragenen Sinne davon sprechen, die Struktur innerhalb der Variablen im Sinne korrelativer Information ausnutzen zu wollen. In unserem Beispiel würde man etwa den fehlenden Depotwert mit Hilfe des bekannten Einkommens vorhersagen wollen, d.h. basierend auf den vollständigen Fällen postuliert man das Modell

$$\texttt{dwert} = \alpha + \beta \cdot \texttt{ek} \qquad (2.3)$$

und schätzt die Parameter α und β für die Kunden 1, 2, 3 und 5. Kunde 7 ist hier nicht relevant, weil bei diesem sowohl der Depotwert wie auch das Einkommen unbekannt sind. SPSS errechnet $\hat{\alpha} = 5086.43$ und $\hat{\beta} = 30.02$. Für Kunde 4 errechnen wir demzufolge

$$\texttt{dwert}_4 = 5086.43 + 30.02 \cdot 7300 = 224232.43\,.$$

Wiederum ist der imputierte Wert stark abhängig von der Information, die in den vollständigen Fällen enthalten ist. Intuitiv offensichtliche Probleme existieren sowohl bei der Wahl des Regressionsmodells selbst wie auch bei der Wahl der eingehenden Variablen. Weder ein lineares Modell noch mit der unvollständigen Variablen korrelierende Variablen garantieren eine 'bessere' Imputation. Auch hier empfiehlt sich eine gewisse Diagnostik, was die Fehlendmechanismen betreffen. Im llgemeinen resultieren aus der FOR erhöhte Residuenvarianzen für die unvollständigen Fälle. Ebenso bewirkt die FOR eine stärkere Korrelation zwischen den unvollständigen Beobachtungen. Little (1992) empfiehlt eine niedrigere Gewichtung der vervollständigten Fälle über den weighted least squares–Ansatz (WLS).

2.3.3 Single Imputation

Beim Begriff *single imputation (SI)* muß zunächst darauf hingewiesen werden, dass die bisher genannten Imputationsverfahren alle *single imputations* sind, in dem Sinne, dass fehlende Werte nur durch *einen* Wert ersetzt werden. Als Imputationsverfahren an sich bedeutet *single imputation*, dass aus einer Verteilung eine Pseudo–Zufallszahl gezogen und imputiert wird. Die Variationsmöglichkeiten sind vielfältig, deren wissenschaftliche Fundierung nicht. Beispielsweise könnten wir die fehlenden Depotwerte auch dadurch ersetzen, indem wir Mittelwert und Varianz aus den vorhandenen Daten schätzen und uns per Zufallszahlengenerator Ersatzwerte generieren lassen, die einer Normalverteilung mit diesen Parametern entstammen. Doch selten ist man in der Lage, Typus und Parameter der (unbekannten) Verteilung derart exakt schätzen oder bestimmen zu können. Eng damit verbunden ist auch die Frage, inwieweit das Fehlen zufällig ist oder nicht. Üblicherweise sind etwa Einkommensdaten Pareto–verteilt und weisen fehlende Werte insbesondere in den hohen und niederen Einkommen(sklassen) auf. Neben diesen Problemen existiert natürlich auch hier das Problem des Fehlers bei der (einmaligen) Imputation.

2.3.4 Nearest Neighbor Imputation

Ein in der Praxis eher selten angewendetes Verfahren ist die *nearest neighbor imputation*, die Imputation des nächsten Nachbarn. Gehen wir erneut von m fehlenden Werten für $i = n -$

$m+1,\dots,n$ aus, also einer Situation

$$\underbrace{x_1,\dots,x_{n-m}}_{\text{beobachtet}},\underbrace{x_{n-m+1},\dots,x_n}_{\text{fehlend}} \quad \text{und} \tag{2.4}$$

$$\underbrace{y_1,\dots,y_{n-m},y_{n-m+1},\dots,y_n}_{\text{beobachtet}} \quad . \tag{2.5}$$

Dann wird ein fehlender Wert x_j, $j=n-m+1,\dots,n$, durch denjenigen Wert x_i ersetzt, $1\le i\le n-m$, der der nächste Nachbar—im Sinne einer zu spezifizierenden Metrik—von x_j ist. In einem einfachen Zusammenhang zweier Variablen bezieht sich das Distanzmaß zur Bestimmung des nächsten Nachbarn auf y–Werte, so dass i die Bedingung

$$\mid y_i - y_j \mid \;=\; \min_{1\le l\le n-m} \mid y_l - y_j \mid \tag{2.6}$$

erfüllt. Gegenüber den bisher genannten Verfahren hat die NNI einige Vorteile: Etwa werden Werte imputiert, die in der beobachteten Stichprobe tatsächlich vorkommen und demzufolge eine gewisse Sinnhaftigkeit offenbaren—vorausgesetzt, man hat meßfehlerfreie Daten. Angenehme asymptotische Eigenschaften bei einfachen Stichproben, etwa bei der Schätzung von Erwar--tungswert und Quantilen wurden bereits nachgewiesen (siehe Chen and Shao (2000)). Als nichtparametrisches Verfahren ist auch eine gewisse Robustheit gegenüber Verletzungen von Modellannahmen zu erwarten. Betrachten wir die NNI nun im Hinblick auf unser Beispiel. Angenommen wir suchen Ersatzwerte für die beiden fehlenden Depotwerte der Kunden 4 und 6. Als maßgebliche Variable zur Identifizierung des nächste Nachbarn diene das monatliche Bruttoeinkommen. Ordnen wir die jeweiligen beobachteten Einkommen der Größe nach aufsteigend, so erhalten wir den Vektor

$$\texttt{ek} = (0, 300, 710, 1700)^t .$$

Das monatliche Bruttoeinkommen von Kunde 4 beträgt EUR 7300, dessen nächster Nachbar aufgrund der 1700 Euro Kunde 5 ist. Damit würden wir im klassischen Sinne der NNI auch für Kunde 4 einen Depotwert von EUR 66000 ansetzen. Der fehlende Depotwert von Kundin 6 wird imputiert durch die 27000 Euro von Kundin 1, die ebenso kein monatliches Einkommen besitzt. Alternativ zum klassischen Ansatz ist auch eine Alternative denkbar, die sich eine feste Anzahl k zunutze macht, um etwa deren Mittelwert als Imputationswert zu verwenden. Für Kunde 4 ergäben sich für $k = 3$ die Depotwerte der Kunden 2, 3 und 5, also EUR 300, 710 und 1700, was uns über die Mittelung einen Ersatzwert von EUR $2710/3 = 903.33$ liefert, siehe dazu Nittner (2003a). Jedoch gestaltet sich eine Erweiterung der NNI auf den mehrdimensionalen Fall nicht einfach und ist für den Laien demzufolge auch schwer zu implementieren. Zudem kann auch die lokale Beschränktheit der NNI Probleme mit sich bringen, die eine Überarbeitung erforderlich macht.

2.3.5 Weitere Verfahren

Neben diesen im Kontext der linearen Regressionsrechnung bekanntesten Verfahren existieren sowohl Erweiterungen wie auch alternative Verfahren. Auf Erweiterungen wird in Abschnitt 2.5 eingegangen, die *nearest neighbor imputation* findet ihre Anwendung im Abschnitt über nicht– und semiparametrische Modelle (4.2).

2.4 Verteilungsbasierte Modelle

Nimmt man eine Verteilung für die Variablen an, bei denen Werte fehlen, so kann likelihoodbasierte oder Bayesianische Inferenz erfolgen, die auch auf Modelle mit nicht–ignorierbarem Fehlen ausgedehnt werden kann. Diese Möglichkeiten werden in Abschnitt 3 beschrieben.

2.5 Hinweise und weiterführende Litaratur

In diesem Kapitel habe wir einige Verfahren vorgstellt, die zum Umgang mit fehlenden Daten verwendet werden können und größtenteils auch Verwendung finden. Wir haben insbesondere darauf Wert gelegt, die Darstellung nicht zu theoretisch zu gestalten und nicht–statistische Überlegungen zuzulassen. Neben den bereits erwähnten Standardwerken wollen wir im einzelnen noch weitere Literaturhinweise geben: *complete* und *available case analysis* werden im Großteil der Litaratur zur Problematik fehlender Daten behandelt. Speziell die Mittelwertimputation bei nicht–stetigem Skalenniveau wurde im Rahmen einiger Simulationsstudien für das lineare Regressionsmodell in Toutenburg and Nittner (2002) behandelt. Ebenfalls im Rahmen des linearen Regressionsmodells wird in Little (1992) die Imputation bedingter Mittelwerte dargestellt und in Simulationen mit bekannten Verfahren verglichen. Weitere Ausführungen findet man in Buck (1960), Afifi and Elashoff (1966) oder Dagenais (1973). Ein neuerer nichtparametrischer Ansatz zur Schätzung des bedingten Mittelwertes wird in Nielsen (2001) beschrieben. Anstatt einer linearen Regressionsrechnung wird hier ein lokaler Ansatz gewählt. Toutenburg, Fieger and Srivastava (1999) beziehen den Responsevektor in die Hilfsregression mit ein und benutzen zusätzlich unterschiedliche Gewichte für beobachtete und fehlende Daten. Die *single imputation* wird relativ kurz in Little and Rubin erwähnt, speziell die Varianzschätzung bei der SI wird in Groves, Dillman, Eltinge and Little (2002) behandelt. Wie zuvor bereits erwähnt, erfährt die *nearest neighbor imputation (NNI)* in der Praxis trotz ihrer relativ guten Eigenschaften wenig Anwendung. Ein Beispiel für ihre Relevanz ist etwa die Volkszählung durch das US Census Bureau (Fay (1999)). Ein neuerer und empfehlenswerter Artikel ist Chen and Shao (2001). Simulationsstudien zum Vergleich der konventionellen Verfahren mit der NNI speziell unter MAR werden in einem univariaten additiven Modell behandelt Nittner (2003b).

3 Inferenz basierend auf der Likelihood

In diesem Abschnitt gehen wir davon aus, dass die Annahme einer parametrischen Verteilung für die erhobenen Variablen gerechtfertigt ist. Häufig werden viele Variablen Z an einem Subjekt erhoben. Es ist dann jeweils genau zu betrachten, für welchen Teil der Variablen eine Verteilungsannahme getroffen wird. Betrachtet man alle Variablen symmetrisch, so ist eine Verteilungsannahme für alle Variablen gewünscht. Sofern alle betrachteten Variablen stetig sind, bietet sich die multivariate Normalverteilung (Mardia, Kent and Bibby, 1979) an. Sind alle Variablen qualitativ (binäre, nominale oder ordinale Merkmale), so bieten sich loglineare Modelle (Agresti, 1990) als grundlegende Modellklasse an. Hier stehen insbesondere auch abgeleitete Modelle zur Verfügung, die ordinale Variablen speziell berücksichtigen. Enthält Z allerdings gemischt stetige und qualitative Variablen , so ist diese Annahme nicht haltbar und die möglichen Modelle werden schnell komplex. In der Praxis trifft man deshalb meist auf Regressionsmodelle, wo eine (univariater Fall) oder mehrere (multivariater Fall) Variablen als Antwortvariablen (response Variable) ausgezeichnet sind und der Einfluß der restlichen Variablen (die Regressoren) auf diese Antwortvariablen untersucht werden soll. Mit Y bezeichnet man die Antwortvariablen, mit X die einflußnehmenden Regressoren. Der Vorteil solcher Modelle liegt (sofern sie für die Fragestellung geeignet sind!) darin, dass jetzt nur für die Antwortvariablen eine Verteilungsannahme getroffen werden muß, nicht aber für die Menge der Regressoren X. Bei den parametrischen Verteilungsannahmen wird man sich im Fall von einer univariaten Antwortvariable der Theorie der verallgemeinerten linearen Modelle (McCullagh and Nelder, 1989) bedienen. Bei der multivariaten Situation kann man wiederum unterscheiden, ob es sich um zu einem bestimmten Zeitpunkt gemessene verschiedene

Variablen handelt (multivariate Querschnittsdaten), ob es sich um die wiederholte Messung ein -und derselben Variable(n) handelt (Paneldaten, Längsschnittanalyse), oder ob es sich um die Messung einer (oder mehrerer) Variablen an abhängigen Individuen handelt (Clusterdaten, Familienstudien). Multivariate Analysen können bereits im Fall vollständiger Daten sehr kompliziert werden und die Zahl möglicher Modelle ist groß. Als Beispiele seien hier nur aufgeführt: marginale Modelle (Liang and Zeger, 1989) für Panel– und Clusterdaten, Modelle mit Zufallseffekten (Verbeke and Molenberghs, 2000) für Panel– und Clusterdaten, bedingte Modelle, graphische Modelle (Whittaker, 1990) für multivariate Daten. Für Modelle im Regressionskontext geben Fahrmeir and Tutz (2001) eine gute Übersicht. Durch fehlende Daten wird die Komplexität dementsprechend noch erhöht. Im Folgenden wollen wir uns Y der Einfachheit wegen als eine univariate Antwortvariable vorstellen. Eine etwaige Bedingung auf X wird zur Vereinfachung der Notation weggelassen. Das angenommene Datenmodell wird jetzt durch die Angabe einer Verteilung bzw. Dichte für Y, $f(ymid\theta)$, festgelegt. Die Dichten können meist in Form einer Exponentialfamilie dargestellt werden, so zum Beispiel die Normalverteilung, die Poissonverteilung oder die Binomialverteilung (Gegenbeispiel: Gleichverteilung). Die Möglichkeit der Darstellung als Exponentialfamilie erleichtert beipielsweise die konkrete Implementation des EM–Algorithmus nach Dempster, Laird and Rubin (1977), eines Algorithmus zur Gewinnung von Maximum–Likelihood Schätzungen bei Problemen mit fehlenden Daten, der in Abschnitt 3.6 näher beschrieben wird. Generell bezeichnet man als Likelihood $L(\theta \mid y)$ eine beliebige Funktion in θ für festes y, die proportional zu $f(y \mid \theta)$ ist (während die Dichte ja als Funktion in y für festes θ aufgefaßt werden muss). Wir werden der Einfachheit wegen Likelihood und Dichte immer mit f bezeichnen.

3.1 Maximum–Likelihood (ML)

Akzeptiert man die Gültigkeit der MAR–Annahme, so läßt sich das Schätzproblem wegen folgender Überlegung vereinfachen. Dazu erinnern wir uns, dass MAR bedeutet, dass

$$P(R \mid y; \xi) = P(R \mid y_{\mathrm{beob}}, y_{\mathrm{fehl}}; \xi) = P(R \mid y_{\mathrm{beob}}; \xi) \ . \tag{3.1}$$

Werden diese Wahrscheinlichkeiten zum Beispiel in Form eines logistischen Regressionsmodells formuliert, so ist ξ der Vektor der Regressionsparamter dieses Modells. Eine andere Möglichkeit ist beispielsweise das Probit–Modell. Die sogenannte *Likelihood der vollständigen Daten* ist gegeben durch

$$f(R, y \mid \theta, \xi) = f(y_{\mathrm{beob}}, y_{\mathrm{fehl}} \mid \theta) P(R \mid y_{\mathrm{beob}}, y_{\mathrm{fehl}}; \xi) \ . \tag{3.2}$$

Sie stellt also die gemeinsame Verteilung der Daten zusammen mit den zufälligen Indikatorvariablen dar, welche den Fehlendmechanismus beschreiben.

Bemerkung: Im Allgemeinen ist $f(y_{\mathrm{beob}}, y_{\mathrm{fehl}} \mid \theta)$ das Produkt von n Dichten, wenn eine unabhängige, bei Regressionsmodellen i.a. nicht identisch verteilte, Stichprobe vom Umfang n angenommen wird. Jedes individuelle y_i, $i = 1, \ldots, n$, kann dabei durchaus multivariat sein. Unter Verwendung von (3.1) vereinfacht sich (3.2) zu

$$f(R, y_{\text{beob}}, y_{\text{fehl}} \mid \theta, \xi) = f(y_{\text{beob}}, y_{\text{fehl}} \mid \theta) P(R \mid y_{\text{beob}}; \xi) \; . \tag{3.3}$$

Die *Likelihood der beobachteten Daten*, welche die Grundlage der Schätzung bildet, erhält man durch Herausintegrieren der fehlenden Daten in (3.2) bzw. (3.3):

$$\begin{aligned} f(R, y_{\text{beob}} \mid \theta, \xi) &= \int P(R \mid y_{\text{beob}}; \xi) f(y_{\text{beob}}, y_{\text{fehl}} \mid \theta) dy_{\text{fehl}} \\ &= P(R \mid y_{\text{beob}}; \xi) \int f(y_{\text{beob}}, y_{\text{fehl}} \mid \theta) dy_{\text{fehl}} \\ &= P(R \mid y_{\text{beob}}; \xi) f(y_{\text{beob}} \mid \theta) \; . \end{aligned} \tag{3.4}$$

Man erkennt sehr schön, dass die Likelihood der beobachteten Daten nun aus zwei separaten Faktoren besteht, welche jeweils durch einen eigenen Parameter(vektor) bestimmt sind. Bestehen keine restriktiven Abhängigkeiten zwischen den Parametern θ und ξ, so nennt man den Fehlendmechanismus *ignorierbar*. Ist man nämlich nur an einer Parameterschätzung von θ interessiert, so genügt es, den Faktor $f(y_{\text{beob}} \mid \theta)$ zu betrachten, der andere Faktor kann bei der Maximierung ignoriert werden, denn für die Maximierung der Likelihood sind beide Faktoren unabhängig voneinander maximierbar. Man macht sich schnell klar, dass im Falle der Nichtgültigkeit von (3.1) diese Vereinfachung nicht möglich ist:

$$
\begin{aligned}
f(R, y_{\text{beob}} \mid \theta, \xi) \quad = \quad & \int P(R \mid y_{\text{beob}},, y_{\text{fehl}}; \xi) \\
& \cdot f(y_{\text{beob}}, y_{\text{fehl}} \mid \theta) dy_{\text{fehl}} \ ,
\end{aligned}
\tag{3.5}
$$

da y_{fehl} jetzt in beiden Faktoren vorkommt. Eine Maximum–Likelihood Schätzung ist auch hier möglich, aber oft technisch sehr anspruchsvoll. Im Falle, dass die MAR–Annahme nicht gerechtfertigt erscheint, muß die gemeinsame Verteilung der Daten und der zufälligen Indikatorvariablen betrachtet werden. Dies kann auf verschiedene Weisen geschehen. Zwei mögliche Faktorisierungen sind die *Pattern Mixture* Modelle und die *Selection* Modelle, die in den folgenden Abschnitten behandelt werden.

3.2 Pattern Mixture Modelle

Betrachtet man die Faktorisierung der gemeinsamen Dichte von Y und R gemäß

$$
f(R, Y \mid \theta, \xi) = f(Y \mid R, \theta) f(R \mid \xi) \ , \tag{3.6}
$$

so spricht man von einem *Pattern Mixture* Modell. Entscheidend ist hier, dass das Datenmodell bedingt auf das konkrete Fehlendmuster R gewählt wird. Anschließend wird über die Verteilung dieser Fehlendmuster 'gemischt'. Der Vorteil dieses Modellansatzes liegt darin, dass die aus den beobachteten Daten nicht identifizierbaren Parameter sofort sichtbar werden. Dazu wollen wir folgendes einfache Beispiel betrachten.

Beispiel: Sei $(Y_1, \ldots, Y_n)$ eine Zufallsstichprobe unabhängiger und identisch verteilter Zufallsvariablen gemäß einer Bernoulli Verteilung mit Parameter $p = E(Y_i) = P(Y_i = 1)$. Im Falle vollständiger Daten ist der Mittelwert $\bar{Y}$ eine erwartungstreue Schätzung für p, d.h. $E(\bar{Y}) = p$. Wir nehmen nun an, dass die ersten $n - m$ Beobachtungen berichtet sind $(m < n)$, während die letzten m Beoachtungen fehlen. Unter der Annahme, dass der Fehlend–Prozess MCAR oder MAR ist (was in diesem einfachen Fall einer i.i.d. Stichprobe äquivalent ist), ist der Complete Case Schätzer $\hat{p} = \frac{1}{n-m}\sum_{i=1}^{n-m} Y_i$ eine erwartungstreue Schätzung für p. Was passiert aber, wenn der Fehlendmechanismus nicht MCAR (MAR) ist? Eine statistische Formulierung des Problems im Sinne eines Pattern Mixture Modells kann mittels der Darstellung

$$\begin{aligned} p &= P(Y_i = 1) \\ &= P(Y_i = 1 \mid R_i = 1; p_1)P(R_i = 1 \mid \xi) \\ &\quad + P(Y_i = 1 \mid R_i = 0; p_0) \cdot [1 - P(R_i = 1 \mid \xi)] \\ &= p_1\xi + p_0(1 - \xi) \end{aligned}$$

gegeben werden. Die Parameter p_0 und p_1 sind die Erwartungswerte für Y im jeweiligen Pattern, wovon es hier nur zwei gibt: entweder wurde Y_i beobachtet $(R_i = 1)$ oder es fehlt $(R_i = 0)$. Der Parameter ξ entspricht dem Erwartungswert von R_i, $\xi = E(R_i) = P(R_i = 1)$. Es ergeben sich folgende Konsequenzen, falls $p_0 \neq p_1$ (wenn Gleichheit gilt, liegt wieder MCAR vor):

- Der Complete Case Schätzer $\frac{1}{n-m}\sum_{i=1}^{n-m} Y_i$ ist eine erwartungstreue Schätzung für p_1, nicht aber für p_0 bzw. p.
- Aus den beobachteten Daten sind p_1 und ξ, der Anteil der Fälle, die beobachtet wurden, schätzbar. Der Parameter p_0 ist nicht identifizierbar aus den beobachteten Daten.
- Für p können durch Annahme der Extremwerte 0 und 1 für p_0 Grenzen geschätzt werden, da

  $$p_1\xi \le p \le p_1\xi + (1-\xi) = 1 - [(1-p_1)\xi] \ ,$$

  und die Parameter, die in den Grenzen vorkommen, aus den beobachteten Daten geschätzt werden können. Literatur zu diesem für komplexe Modelle schwierigen Ansatz der partiell identifizierten Wahrscheinlichkeitsverteilungen wird im Abschnitt 3.8 angegeben.

Der Vorteil dieser Modellklasse liegt laut Little (1993) darin, die nicht aus den Daten identifizierbaren Parameter eines statistischen Modells explizit auszuweisen. Der Nachteil dieses Modellansatzes liegt allerdings darin, dass man ein anderes Datenmodell wählen muß, als man im Falle vollständiger Daten gewählt hätte. Unangenehm ist außerdem, dass die Zahl der zu schätzenden Parameter mit der Zahl der verschiedenen Pattern ansteigt, sofern nicht wieder vereinfachende Annahmen (Restriktionen) getroffen werden. In diesem Punkt sind die im folgenden Abschnitt besprochenen Selection Modelle besser interpretierbar, als sie einfach das Datenmodell, das man auch im Falle vollständiger Daten gewählt hätte, durch ein Modell für den Fehlendmechanismus ergänzen.

3.3 Selection Modelle

Betrachtet man die Faktorisierung

$$f(R, Y \mid \theta, \xi) = f(R \mid Y; \xi) f(Y \mid \theta) \ , \tag{3.7}$$

so spricht man von einem *Selection* Modell (wobei θ und ξ Parameter mit anderer Bedeutung als im vorigen Abschnitt darstellen). Dieses Modell geht auf Heckman (1976) zurück. Hier wird angenommen, dass das Fehlen davon abhängt, ob eine andere (latente, unbeobachtete) Variable Z mit $Z \sim N(\mu_Z, \sigma_Z^2)$ den Schwellenwert Null überschreitet oder nicht, also

$$P(Y_i \text{ beobachtet}) = P(Z_i \geq 0) \, ,$$

wobei auch $Y \sim N(\mu_Y, \sigma_Y^2)$. Die Abhängigkeit der Fehlend–Wahrscheinlichkeit vom Response wird dadurch modelliert, dass eine Korrelation zwischen Y und Z eingeführt wird, d.h. man trifft die Annahme einer bivariaten Normalverteilung für (Y, Z). Man kann zeigen, dass dieses Modell äquivalent zum folgenden Modell ist:

$$P(Y_i \text{ beobachtet} \mid Y_i = y_i) = \Phi(\gamma_0 + \gamma_1 y_i) \, ,$$

wobei γ_0 und γ_1 geeignet zu wählen sind und Φ für die Verteilungsfunktion der Standardnormalverteilung steht. Eine ausführliche Behandlung dieses Modells findet sich in Amemiya (1985) und Rubin (1987). Man kann insbesondere zeigen, dass die Schätzbarkeit des Modells durch die Nichtlinearität des sog. *Mills ratio* $\phi(y)/(1 - \Phi(y))$ gegeben ist, wobei ϕ die Dichte einer Standardnormalverteilung darstellt. Als Vorteil der Selection Modelle wird gesehen, dass man das Datenmodell wählt, welches man auch im Falle vollständiger Daten gewählt hätte, und dieses Modell lediglich unterstützen muß durch ein Modell für den Fehlendmechanismus. Als Nachteil der Selection Modelle gilt, dass eine Annahme für ein spezielles Modell für den Fehlendmechanismus in der Praxis schwierig ist und man das Problem der Fehlspezifikation hat, welche sich auch negativ auf die eigentlich interessierenden Schätzungen auswirken kann, zum Beispiel in Form eines Bias. Im Allgemeinen wird dem Anwender empfohlen, eine Sensitivitätsanalyse durchzuführen, also die Schätzungen unter verschiedenen Annahmen für den Fehlendmechanismus zu wiederholen. Meist ist dies jedoch einfacher gesagt als getan, insbesondere wenn sehr viele Variablen im Spiel sind.

3.4 Äquivalenz von Pattern Mixture Modellen und Selection Modellen in speziellen Fällen

Betrachten wir nochmals das Beispiel aus Abschnitt 3.2. Wählen wir als Ausgangspunkt das Modell für den Fehlendmechanismus, $P(R_i = 1 \mid Y_i = 1) = \alpha_1$ und $P(R_i = 1 \mid Y_i = 0) = \alpha_0$, so liefert die Anwendung des Satzes von Bayes:

$$E(Y_i \mid R_i = 1) = P(Y_i = 1 \mid R_i = 1) = \frac{p\alpha_1}{p\alpha_1 + (1-p)\alpha_0} \quad (3.8)$$

und

$$
\begin{aligned}
E(Y_i \mid R_i = 0) &= P(Y_i = 1 \mid R_i = 0) \\
&= \frac{p(1-\alpha_1)}{p(1-\alpha_1) + (1-p)(1-\alpha_0)} \,. \qquad (3.9)
\end{aligned}
$$

Im Allgemeinen sind (3.8) und (3.9) verschieden. Betrachte als einfaches Zahlenbeispiel $p = 0.4$, $\alpha_1 = 0.9$ and $\alpha_0 = 0.8$. Dann ist $E(Y_i \mid R_i = 1) = 0.36/0.84 = 0.429$ and $E(Y_i \mid R_i = 0) = 0.04/0.16 = 0.25$. Ebenfalls erkennt man, dass (3.8) und (3.9) gleich und gleich $E(Y_i) = p$ sind, wenn $\alpha_1 = \alpha_0$, d.h. MCAR (MAR) liegt vor. Übrigens: da in diesem Beispiel $E(Y_i \mid R_i = 1) > p$ ist, würde der complete case Schätzer den Populationsparameter p *überschätzen.* Für dieses Beispiel, bei dem MCAR und MAR äquivalente Aussagen sind, konnte man zeigen, dass in diesen Fällen auch die bedingten Erwartungswerte gleich und gleich dem unbedingten Erwartungswert sind und damit Selection Modell und Pattern Mixture Modell sich nicht unterscheiden. Was aber gilt in allgemeineren Fällen? Dazu betrachten wir den Fall, dass neben Y noch eine weitere Variable X mit im Spiel ist, fehlenede Werte aber nur in Y auftreten. Um die Notation kompakter zu halten, lassen wir die Parameter und den Index i im Folgenden weg.

MCAR In diesem Fall gilt

$$
f(R \mid Y, X) = f(R) \,. \qquad (3.10)
$$

Das Selection Modell (3.7) vereinfacht sich also zu

$$f(X,Y,R) = f(R)f(X,Y)\,. \tag{3.11}$$

Für das Pattern Mixture Modell folgt

$$f(Y,X \mid R) = f(X,Y)\,,$$

denn mit (3.11) gilt

$$f(Y,X \mid R) = \frac{f(X,Y,R)}{f(R)} = \frac{f(R)f(X,Y)}{f(R)} = f(X,Y)\,.$$

Man erhält als Ergebnis, dass im Falle von MCAR das Selektionsmodell und das Pattern Mixture Modell zum gleichen Ergebnis führen, nämlich

$$f(X,Y,R) = f(R)f(X,Y) \tag{3.12}$$

Geht man jetzt noch zusätzlich davon aus, dass die beiden Faktoren in (3.12) von unabhängigen Parametern indiziert sind, also

$$\begin{aligned}
f(R) &= f_\psi(R) \qquad \psi \in \Psi \\
f(X,Y) &= f_\theta(X,Y) \qquad \theta \in \Theta \\
f(X,Y,R) &= f_{\theta,\psi}(X,Y,R) = f_\theta(X,Y)f_\psi(R),\ (\theta,\eta) \in \Theta \times \Psi,
\end{aligned}$$

dann erhält man das Ergebnis, dass für die Bestimmung einer Parameterschätzung von θ lediglich der entsprechende Faktor der Likelihood maximiert werden muss, der auf den Beiträgen $f_\theta(X_i, Y_i)$ beruht.

MAR Gemäß der Definition liegt MAR dann vor, wenn

$$f(R \mid Y, X) = f(R \mid X)\,. \tag{3.13}$$

Für das Selektionsmodell gilt dann

$$f(X, Y, R) = f(R \mid X) f(X, Y) \tag{3.14}$$

und für das Pattern Mixture Modell folgt

$$f(Y, X \mid R) = f(X, Y) \frac{f(R \mid X)}{f(R)}\,,$$

denn mit (3.14) gilt

$$f(Y, X \mid R) = \frac{f(X, Y, R)}{f(R)} = \frac{f(R \mid X) f(X, Y)}{f(R)}\,.$$

Obwohl eine mathematische Identität zwischen den beiden Modellen herstellbar ist, so sind sie vom statistischen Standpunkt aus verschieden: Das Pattern Mixture Modell spezifiziert ein Modell für $f(Y, X \mid R)$ und $f(R)$, das Selektionsmodell ein Modell für $f(X, Y)$ und $f(R \mid X, Y)$. Als Resultat erhalten wir also, dass die beiden Modellarten bereits bei MAR zu unterschiedlichen *statistischen Modellen* führen.

3.5 Parametrische Verteilungsannahmen im multivariaten Fall

Bisher wurde stets von univariaten Stichproben bzw. von univariaten Responsevariablen ausgegangen. Hierfür stehen mit den Exponentialfamilien und den generalisierten linearen Modellen gut erforschte statistische Verfahren zur Verfügung. Schwieriger gestaltet sich der multivariate Fall, insbesondere bei unterschiedlicher Skalierung der Variablen, etwa stetig/ordinal oder stetig/nominal. Man beachte, dass die Verwendung der multivariaten Normalverteilung in diesem Kontext als nicht geeignet bezeichnet werden muß. Eine Möglichkeit besteht in der Faktorisierung der gemeinsamen Verteilung in bedingte Verteilungen und Randverteilungen. Als Beispiel sei der Ansatz von Fitzmaurice and Laird (1995) genannt, der sich allerdings nur mit vollständigen Daten beschäftigt. Es entsteht also das Problem der Wahl der Faktorisierung. Zur Veranschaulichung möge das folgende Beispiel dienen. Betrachten wir zwei Variablen unseres Anfangsbeispiels. Die binäre Variable `depot` und die stetige Variable `ek` sind ein Beispiel für einen gemischten Verteilungstyp, bei dem nicht die bivariate Normalverteilung im herkömmlichen Sinne zur Anwendung kommen kann. Prinzipiell stehen hier zwei Möglichkeiten zur Faktorisierung der gemeinsamen Verteilung $f(\texttt{depot, ek})$ zur Verfügung. Die erste Möglichkeit verwendet die Faktorisierung

$$f(\texttt{depot, ek}) = f(\texttt{ek} \mid \texttt{depot}) \cdot f(\texttt{depot})\,, \tag{3.15}$$

die zweite verwendet

$$f(\texttt{depot, ek}) = f(\texttt{depot} \mid \texttt{ek}) \cdot f(\texttt{ek})\,. \tag{3.16}$$

Im ersten Fall ist die Randverteilung der Einkommensvariablen eine Mischung aus den beiden Verteilungen $f(\texttt{ek} \mid \texttt{depot} = \text{'Ja'})$

und $f(\texttt{ek} \mid \texttt{depot} =\text{'Nein'})$. Nimmt man für diese beispielsweise Normalverteilungen an, so ergibt sich eine Mischung aus zwei Normalverteilungen. Die Randverteilung von depot ist eine Bernoulliverteilung. Im zweiten Fall dagegen ist die Randverteilung der Einkommensvariablen eine Normalverteilung. Für die bedingte Verteilung von $f(\texttt{depot} \mid \texttt{ek})$ kann etwa ein Logitmodell angesetzt werden. Wenngleich eine mathematische Äquivalenz zwischen den Gleichungen (3.15) und (3.16) vorliegt, ist zu beachten, dass statistisch zwei unterschiedliche Modelle gebildet werden, sobald die Verteilungen spezifiziert werden.

3.6 EM Algorithmus

Für Probleme mit fehlenden Daten kann ganz allgemein der sog. Expectation–Maximization Algorithmus nach Dempster et al. (1977) zur Gewinnung von Maximum–Likelihood Schätzungen verwendet werden. Es handelt sich dabei um einen iterativen Algorithmus, der nach Gewinnung einer Anfangsschätzung $\theta^{(0)}$ für θ sukzessive eine Folge von Schätzungen $\theta^{(t)}$, $t = 1, 2, \dots$ gewinnt, welche unter gewissen Bedingungen gegen das Maximum $\hat{\theta}$—der Maximum–Likelihood Schätzung—konvergiert. Wir betrachten hier nur den Fall von ignorierbarem Nonresponse (MAR). Ausgangspunkt ist dann die Likelihood der beobachteten Daten in der unter MAR vereinfachten Form (3.4), bei der es genügt,

$$f(y_{\text{beob}} \mid \theta) = \int f(y_{\text{beob}}, y_{\text{fehl}} \mid \theta) dy_{\text{fehl}}$$

bezüglich θ zu maximieren. Ist das (möglicherweise mehrdimensionale) Integral berechenbar, so kann eine Maximum–Likelihood Schätzung durch ein Newton–Raphson oder Fisher–Scoring Verfahren gewonnen werden. Ist dies sehr oder zu komplex, so bietet sich der EM–Algorithmus als Alternative an. Die Grundidee des EM Algorithmus ist, das Schätzproblem immer wieder auf die Maximierung vollständiger Daten zurückzuführen. Im Folgenden bezeichnet $l(\theta \mid y)$ die logarithmierte Likelihood. Ausgangspunkt ist die Faktorisierung

$$f(y_{\text{beob}}, y_{\text{fehl}} \mid \theta) = f(y_{\text{beob}} \mid \theta) f(y_{\text{fehl}} \mid y_{\text{beob}}; \theta) \ . \tag{3.17}$$

Logarithmiert man auf beiden Seiten dieser Gleichung, erhält man

$$l(y_{\text{beob}}, y_{\text{fehl}} \mid \theta) = l(y_{\text{beob}} \mid \theta) + \log f(y_{\text{fehl}} \mid y_{\text{beob}}; \theta) \ . \tag{3.18}$$

Umordnen dieser Gleichung liefert

$$l(y_{\text{beob}} \mid \theta) = l(y_{\text{beob}}, y_{\text{fehl}} \mid \theta) - \log f(y_{\text{fehl}} \mid y_{\text{beob}}; \theta) \ . \tag{3.19}$$

Die Idee ist jetzt, in Gleichung (3.19) auf beiden Seiten den Erwartungswert bezüglich der Verteilung der fehlenden, gegeben die beobachteten Daten an einer aktuellen Schätzung $\theta^{(t)}$, also $f(y_{\text{fehl}} \mid y_{\text{beob}}; \theta^{(t)})$, zu bilden:

$$l(y_{\text{beob}} \mid \theta) = M(\theta \mid \theta^{(t)}) - H(\theta \mid \theta^{(t)}) \ , \tag{3.20}$$

mit

$$M(\theta \mid \theta^{(t)}) = \int l(y_{\text{beob}}, y_{\text{fehl}} \mid \theta) f(y_{\text{fehl}} \mid y_{\text{beob}}; \theta^{(t)}) dy_{\text{fehl}} \ , \tag{3.21}$$

$$H(\theta|\theta^{(t)}) = \int f(y_{\text{fehl}} \mid y_{\text{beob}};\theta) f(y_{\text{fehl}} \mid y_{\text{beob}};\theta^{(t)}) dy_{\text{fehl}} \ , \tag{3.22}$$

und weil

$$\begin{aligned}
\int l(y_{\text{beob}} \mid \theta) \quad &\cdot \quad f(y_{\text{fehl}} \mid y_{\text{beob}};\theta^{(t)}) dy_{\text{fehl}} = l(y_{\text{beob}} \mid \theta) \\
&\cdot \int f(y_{\text{fehl}} \mid y_{\text{beob}};\theta^{(t)}) dy_{\text{fehl}} \\
&= \quad l(y_{\text{beob}} \mid \theta) \cdot 1 \\
&= \quad l(y_{\text{beob}} \mid \theta) \ .
\end{aligned}$$

Der Algorithmus stützt sich ausschließlich auf den Term $M(\theta \mid \theta^{(t)})$: *(E)rwartungs–Schritt:* Bestimme den Erwartungswert $M(\theta \mid \theta^{(t)})$, gegeben durch Gleichung (3.21). Idealerweise ist dieser Schritt analytisch möglich. *(M)aximierungs–Schritt:* Bestimme die nächste Schätzung $\theta^{(t+1)}$ durch Maximieren von $M(\theta \mid \theta^{(t)})$ bezüglich θ. Idealerweise läßt sich dieser Schritt mit einem vorhandenen, implementierten Verfahren für vollständige Daten lösen. Beachtet man nun, dass für die Likelihood an der Stelle $\theta^{(t)}$ nach (3.20) die Gleichung

$$l(y_{\text{beob}} \mid \theta^{(t)}) = M(\theta^{(t)} \mid \theta^{(t)}) - H(\theta^{(t)} \mid \theta^{(t)})$$

gilt und

- $M(\theta^{(t+1)} \mid \theta^{(t)}) \geq M(\theta^{(t)} \mid \theta^{(t)})$, nach Definition, da die linke Seite durch Maximieren im M–Schritt entstanden ist,
- $H(\theta^{(t+1)} \mid \theta^{(t)}) \leq H(\theta^{(t)} \mid \theta^{(t)})$, was hier ohne Beweis bleiben soll (Jensen'sche Ungleichung),

so erhält man:

$$l(y_{\text{beob}} \mid \theta^{(t+1)}) \geq l(y_{\text{beob}} \mid \theta^{(t)}) \ , \qquad (3.23)$$

d.h. mit jedem Iterationsschritt wächst die Likelihood. Zur Konvergenz des Verfahrens finden sich einige Literaturhinweise in Fahrmeir and Tutz (2001). Soviel sei hier gesagt: ist die log–Likelihood multimodal, oder besitzt sie einen Sattelpunkt, so hängt es vom Startpunkt ab, wohin die Sequenz $l(\theta^{(t)})$ konvergiert. Oder anders ausgedrückt: dann kann eine Konvergenz gegen ein globales Maximum nicht garantiert werden. Die Konvergenzgeschwindigkeit hängt gemäß dem sogenannten *missing information principle* (Little and Rubin, 1987) vom Anteil der fehlenden Werte ab: je höher dieser Anteil, desto langsamer die Konvergenz. Ein Nachteil des EM–Algorithmus ist die schwierige Gewinnung von Varianzschätzungen. Im Newton–Raphson–Verfahren, welches die beobachtete Informationsmatrix verwendet (also im wesentlichen die zweiten partiellen Ableitungen der log–Likelihood nach θ), fallen diese quasi als Abfallprodukt im letzten Iterationsschritt an. Allerdings bietet die Formel von Louis (1982) eine Möglichkeit, solche Schätzungen zu gewinnen.

EM–Algorithmus für einfache Exponentialfamilien. Die Dichte einer einfachen Exponentialfamilie läßt sich darstellen als

$$f(Y \mid \theta) = b(Y)\frac{\exp\left(S(Y)\theta\right)}{a(\theta)} \ .$$

Dabei sind a und b jeweils Funktionen ihrer Argumente, $S(Y)$ ist die suffiziente Statistik und $Y = (Y_{\text{beob}}, Y_{\text{fehl}})$. Der E–Schritt vereinfacht sich dann erheblich:

$$l(\theta \mid Y) = \log(b(Y)) - \log(a(\theta)) + S(Y)\theta$$

Nur der Term $S(Y)\theta$ hängt von Y_{fehl} ab. Der E–Schritt ist einfach die Schätzung der complete–data suffizienten Statistik $S(Y)$ durch

$$S^{(t+1)} = E(S(Y) \mid Y_{\text{beob}}, \theta^{(t)}).$$

Dies ist einfach, wenn $l(\theta \mid Y)$ linear in $S(Y)$ ist. Der M–Schritt $\theta^{(t+1)}$ ist dann die Lösung der Likelihood–Gleichungen

$$E(S(Y) \mid \theta) = S^{(t+1)}.$$

Der M–Schritt wird wie bei vollständigen Daten durchgeführt und vorhandene Software ist für den M–Schritt nutzbar. Im Allgemeinen werden im E–Schritt nicht direkt die fehlenden Werte ersetzt, sondern der Anteil der fehlenden Werte an der (complete data) suffizienten Statistik. Folgendes einfaches Beispiel soll die Schritte demonstrieren.

*B*eispiel: Univariate, normalverteilte Daten, Stichprobe vom Umfang n, $n-m$ beobachtet, m fehlend:

$$Y_i \text{ i.i.d. } N(\mu, \sigma^2),\ i = 1, \ldots, n$$

$$\begin{aligned} E(Y_i) &= \mu \\ E(Y_i^2) &= Var(Y_i) + (E(Y_i))^2 \\ &= \sigma^2 + \mu^2 \end{aligned}$$

Die complete data Log–Likelihood ist also linear in den suffizienten Statistiken $\sum_{i=1}^{n-m} Y_i$ und $\sum_{i=1}^{n} Y_i^2$.

E–Schritt:

a)

$$
\begin{aligned}
E\left(\sum_{i=1}^{n} Y_i \mid Y_{\text{beob}}, \mu^{(t)}, \sigma^{2(t)}\right) &= E\left(\sum_{i=1}^{n-m} Y_i \mid Y_{\text{beob}}, \mu^{(t)}, \sigma^{2(t)}\right) \\
&\quad + E\left(\sum_{i=n-m+1}^{n} Y_i \mid Y_{\text{beob}}, \mu^{(t)}, \sigma^{2(t)}\right) \\
&= \sum_{i=1}^{n-m} Y_i + m\mu^{(t)}
\end{aligned}
$$

b)

$$
\begin{aligned}
E\left(\sum_{i=1}^{n} Y_i^2 \mid Y_{\text{beob}}, \mu^{(t)}, \sigma^{2(t)}\right) &= E\left(\sum_{i=1}^{n-m} Y_i^2 \mid Y_{\text{beob}}, \mu^{(t)}, \sigma^{2(t)}\right) \\
&\quad + E\left(\sum_{i=n-m+1}^{n} Y_i^2 \mid Y_{\text{beob}}, \mu^{(t)}, \sigma^{2(t)}\right) \\
&= \sum_{i=1}^{n-m} Y_i^2 + m((\mu^{(t)})^2 + \sigma^{2(t)})
\end{aligned}
$$

M–Schritt:

$$
\begin{aligned}
\mu^{(t+1)} &= \frac{1}{n}\left[\sum_{i=1}^{n-m} Y_i + m\mu^{(t)}\right] \\
\sigma^{2(t+1)} &= \frac{1}{n}\left[\sum_{i=1}^{n-m} Y_i^2 + m((\mu^{(t)})^2 + \sigma^{2(t)})\right] - (\mu^{(t)})^2
\end{aligned}
$$

Im Maximum gilt:

$$
\begin{aligned}
\mu^{(t+1)} &= \mu^{(t)} = \hat{\mu} \\
\sigma^{2(t+1)} &= \sigma^{2(t)} = \hat{\sigma}^2
\end{aligned}
$$

1. Schätzung für μ:

$$\begin{aligned}
\hat{\mu} &= \frac{1}{n}\left[\sum_{i=1}^{n-m} Y_i + m\hat{\mu}\right] \\
&= \frac{1}{n}\sum_{i=1}^{n-m} Y_i + \frac{m}{n}\hat{\mu} \\
n\hat{\mu} &= \sum_{i=1}^{n-m} Y_i + m\hat{\mu} \\
(n-m)\hat{\mu} &= \sum_{i=1}^{n-m} Y_i \\
\Rightarrow\ \hat{\mu} &= \frac{1}{n-m}\sum_{i=1}^{n-m} Y_i = \bar{Y}_{\text{beob}}
\end{aligned}$$

2. Schätzung für σ^2:

$$\begin{aligned}
\hat{\sigma}^2 &= \frac{1}{n}\left[\sum_{i=1}^{n-m} Y_i^2 + m(\hat{\mu}^2 + \hat{\sigma}^2)\right] - \hat{\mu}^2 \\
n\hat{\sigma}^2 &= \sum_{i=1}^{n-m} Y_i^2 + m\hat{\mu}^2 + m\hat{\sigma}^2 - n\hat{\mu}^2 \\
&\vdots \\
\Rightarrow\ \hat{\sigma}^2 &= \left(\frac{1}{n-m}\sum_{i=1}^{n-m} Y_i^2\right) - \hat{\mu}^2
\end{aligned}$$

In diesem Fall erhält man das bekannte Ergebnis, dass die Parameter aus den beobachteten Daten geschätzt werden können. Ein iterativer Algorithmus ist hier also nicht notwendig, man erhält das Ergebnis durch explizite Formeln.

3.7 Bayes Inferenz

Mit der Verfügbarkeit immer schnellerer Computer, der Bekanntmachung der sog. MCMC (Markov Chain Monte Carlo) Algorithmen für die Statistiker durch Gelman and Rubin (1992) und der anschließend explosionsartigen Vermehrung entsprechender Arbeiten und Algorithmen in der Literatur, steht heute ein breites Spektrum an Algorithmen zur Verfügung, um Bayes Inferenz im Allgemeinen und im Besonderen im Fall fehlender Daten betreiben zu können. Eine detaillierte Behandlung würde den Rahmen dieses Artikels sicherlich sprengen. Schafer (1997) bietet einen guten Einstieg für die multivariate Analyse mit fehlenden Daten, wobei die symmetrische Sichtweise auf die erhobenen Variablen betont wird: multivariate Normalverteilung für stetige Daten, loglineare Modelle für rein qualitative Daten und ein Modell für gemischte Variablen, das sogenannte *general location model*, bei dem eine Faktorisierung der gemeinsamen Verteilung der stetigen und diskreten Variablen in der Form $f(\texttt{stetige} \mid \texttt{diskrete})f(\texttt{diskrete})$ vorgenommen wird. Im Folgenden wollen wir nur kurz den MAR Fall beschreiben. Bei der Bayes Inferenz wird der Parameter θ als zufälliger Parameter aufgefaßt, dem zunächst eine a priori Verteilung $\pi(\theta)$ zugeordnet wird. Interessiert ist man dann an der a posteriori Verteilung von θ, $\pi(\theta \mid y)$, nachdem die Daten y erhoben wurden. Die Grundidee ist die des (sukzessiven) Lernens über θ durch (sukzessive) Beobachtung von Daten. Im MAR–Fall kann man zeigen (Schafer, 1997), dass das Problem auf die Berechnung der a posteriori Verteilung gegeben die beobachteten Daten zurückgeführt werden kann, ähnlich wie bei der Maximum–Likelihood Schätzung im MAR–Fall nur die beobachtete Likelihood betrachtet werden muß. Der Zusammenhang wird deutlich durch folgende Beziehung:

$$\pi(\theta \mid y_{\text{beob}}) \propto l(\theta \mid y_{\text{beob}})\pi(\theta) \ .$$

Dabei wird $\pi(\theta \mid y_{\text{beob}})$ entsprechend als beobachtete a posteriori Verteilung bezeichnet. Ist diese Verteilung *explizit* berechenbar und von bekannter Form, erhält man unmittelbar interessierende Größen, wie den a posteriori Erwartungswert von θ. Oftmals ist eine explizite Berechnung aber kompliziert und man stützt sich auf Monte Carlo Methoden zur approximativen Gewinnung dieser Größen. In jüngster Zeit werden dazu häufig die bereits oben genannten MCMC Methoden benutzt, welche (statistisch abhängige) Ziehungen aus der a posteriori Verteilung generieren, woraus sich wiederum z.B. der a posteriori Erwartungswert einfach als Mittelwert der Ziehungen schätzen läßt. In niedrigdimensionalen Problemen, d.h. wenn θ nur aus wenigen Komponenten besteht, können auch Alternativen wie das *importance sampling* sinnvoll sein, welches auch in Schafer (1997) beschrieben ist. In hochdimensionalen Problemen allerdings scheinen unseres Wissens nach kaum Alternativen zu MCMC Verfahren zu bestehen. Ein Nachteil der Bayes Inferenz ist sicherlich die Notwendigkeit der Wahl einer geeigneten a priori Verteilung. Dies wird auch meist als Hauptargument der Kritiker verwendet. Eine Sensitivitätsanalyse der Ergebnisse bei Wahl unterschiedlicher prioris kann hier nützlich sein, ist in der Praxis aber schwierig. Bei geringen Stichprobenumfängen kann es sein, dass der Maximum–Likelihood Schätzer nicht existiert (Beispiel: Kontingenztafeln mit Nullbesetzungen). Hier kann die Wahl einer informativen priori Verteilung den Vorteil einer regularisierenden Wirkung haben, ähnlich der *Ridge*–Schätzung (Rao and Toutenburg (1999)) im linearen Modell. Auf eine weitere Diskussion der Vorteile und Nachteile sei hier aus Platzgründen verzichtet.

3.8 Hinweise und weiterführende Litaratur

Die Idee, Grenzen zu berechnen, um die Problematik der Sensitivität von Modellen für Fehlendmechanismen zu umgehen—um also im Fall, dass keinerlei Vorwissen über den Fehlendmechanismus existiert, die durch die fehlenden Daten induzierte Unsicherheit auszuloten—wird nur von wenigen Autoren verfolgt, siehe Horowitz and Manski (2000), Horowitz and Manski (2001), Vansteelandt and Goetghebeur (2001), Horowitz and Manski (2002), Zaffalon (2002), **?** und Heumann (2003). Dies liegt zum einen vermutlich an der Tatsache, dass es sich um eventuell komplexe Optimierungsprobleme handelt, zum anderen daran, dass die berechneten Grenzen sehr weit werden können und man sich damit eingestehen müßte, dass die erhobenen Daten für eine bestimmte gewünschte Analyse unbrauchbar sind. Pattern Mixture Modelle und Selection Modelle werden meist im Zusammenhang mit speziellen statistischen Datenmodellen diskutiert. Neben Little (1993) seien als Beispiele ohne Anspruch auf Vollständigkeit oder Systematik aufgeführt: Little (1994) für einen Pattern Mixture Ansatz bei multivariat (bivariat) normalverteilten Daten, Diggle and Kenward (1994) für Drop–Out im multivariaten Normalverteilungsmodell, wobei sich die Varianz–Kovarianzstruktur additiv aus verschiedenen Komponenten (unabhängige Fehler, individuenspezifische Zufallseffekte und serielle Korrelation) zusammensetzt, Little and Wang (1996) für den allgemeinen Fall (enthält auch ein Kontingenztafelbeispiel) und Pattern Mixture, Hedeker and Gibbons (1997) für longitudinale Daten mit fehlenden Werten und einem sog. Random Effects Pattern Mixture Modell, Ekholm and Skinner (1998) für eine Anwendung des Pattern Mixture Ansatzes auf Longitudinaldaten mit fehlendem binären Response, Molenberghs, Michiels and Kenward (1998) für einen Pseudo–Likelihood–Ansatz für Pattern Mixture und Selection Modelle, Molenberghs, Michiels, Kenward and Diggle (1998) für Pattern Mixture Modelle, wenn das Fehlendmuster monoton ist, Kenward (1998) zur Sensitivität von Selection Modellen, Troxel, Lipsitz and Harrington (1998) für den Spezialfall der marginalen Modelle (Stichwort: GEE, *generalized estimating equations*), Toledano and Gatsonis (1999) für GEE bei ordinalen Daten und fehlendem Response und fehlenden Werten in einer Kovariablen, Michiels, Molenberghs and Lipsitz (1999), Daniels and Hogan (2000) für eine Sensitivitätsanalyse bei informativem (also nichtignorierbarem) Drop–Out, also bei monotonem Fehlendmuster, Park and Lee (1999) bei longitudinalen Daten, Pattern Mixture Modell, Ibrahim (2001) für Maximum–Likelihood Schätzung von Selection Modellen im generalisierten linearen Modell mit Zufallseffekten, Michiels, Molenberghs, Bijnens, Vangeneugden and Thijs (2002) für Selection und Pattern Mixture Modellen bei Longitudinaldaten mit Drop–out, Heumann (2003) für Bayes–Schätzung von Selection Modellen mittels Monte Carlo Methoden im generalisierten linearen

Modell mit Zufallseffekten. Im Zusammenhang mit dem EM–Algorithmus wurden einige Modifikationen erarbeitet. Ist beispielsweise der M–Schritt selbst nur iterativ lösbar (Beispiel: loglineare Modelle), d.h. man hat ineinander geschachtelte iterative Verfahren, so kann man auf eine vollständige Maximierung im M–Schritt (die ja dann sowieso noch nicht das gewünschte Maximum liefert) verzichten und nur einen Schritt des Maximierungsalgorithmus im M–Schritt anwenden (oder eine wachsende Anzahl mit zunehmender Anzahl an EM–Schritten). Dieser Algorithmus wurde als ECM–Algorithmus (Expectation Conditional Maximization) von van Dyk, Meng and Rubin (1995) eingeführt und soll das Verfahren insgesamt beschleunigen. Ist der E–Schritt analytisch schwierig zu berechnen, so steht als Alternative der MCEM–Algorithmus (Monte Carlo Expectation Maximization) nach Tanner (1991) zur Verfügung. Der Erwartungswert im E–Schritt wird hier durch eine Monte Carlo Simulation berechnet. Der EM–Algorithmus wird auch in anderen statistischen Fragestellungen eingesetzt, beispielsweise zur Schätzung von Modellen mit Zufallseffekten oder bei Mischverteilungsmodellen, siehe auch McLachlan and Krishnan (1997).

4 Modelle vom Regressionstyp

Im Folgenden wollen wir eine kleine Übersicht geben über Methoden zur Behandlung von Modellen vom Regressionstyp (vgl. Toutenburg (2002)). Diese zeichnen sich gerade dadurch aus, dass a priori die Variablen eingeteilt werden in Einflußgrößen (Regressoren) und in von den Regressoren beeinflußte, abhängige Variablen (Response). Der Vorteil von Regressionsmodellen liegt darin, dass keine Annahme über die Verteilung der Regressoren getroffen werden muß, zumindest wenn die Regressoren vollständig sind. Wir beschränken uns hier auf eine Darstellung des linearen Modells, wie sie in umfangreicherer Form auch in Rao and Toutenburg (1999) zu finden ist, sowie eine kurze Einführung in nicht– und semiparametrische Modelle. Komplexere Modelle, wie etwa die generalisierten linearen Modelle oder Modelle für Longitudinal– und Clusterdaten werden lediglich im Abschnitt über weiterführende Literatur angeführt.

4.1 Lineares Modell

4.1.1 Fehlende Daten im Response

Bei kontrollierten Experimenten wie klinischen Studien in der Pharmakologie oder technischen Laboruntersuchungen wird die X–Matrix durch gezielte Versuchsplanung festgelegt und ein Response Y beobachtet. Die Auswertung erfolgt mit Standardverfahren wie z.B. der Varianzanalyse oder dem üblichen linearen Modell und den zugehörigen Testverfahren. Bei dieser Versuchsanlage kann man davon ausgehen, dass fehlende Werte eher im Response als im Versuchsplan auftreten. Damit wird die Balanziertheit gestört. Selbst wenn für die Daten die MCAR–Annahme gilt, ist es vorteilhafter, mit einem aufgefüllten Y–Vektor die Standardanalyse balanzierter Modelle durchzuführen als mit dem kleineren complete case Datensatz zu arbeiten.

Falls der Versuchsplan z.B. vollständig gekreuzt ist, würde die Beschränkung auf den complete case Datensatz zu Schwierigkeiten bei der Interpretation führen. Im Folgenden nehmen wir an, dass die Fehlend–Wahrscheinlichkeit für eine Beobachtung y nicht von y abhängt.

KQ–Schätzung bei vollständigem Datensatz Sei Y die Responsevariable und X die (n, K)–Designmatrix, so gelte für die Realisierung y von Y das lineare Modell

$$y = X\beta + \epsilon, \quad \epsilon \sim N(0, \sigma^2 I) \ . \tag{4.1}$$

Die KQ–Schätzung von β ist $b = (X'X)^{-1}X'y$ und die beste erwartungstreue Schätzung von σ^2 ist

$$\begin{aligned} s^2 &= (y - Xb)'(y - Xb)(n - K)^{-1} \\ &= \frac{\sum_{t=1}^{T}(y_t - \hat{y}_t)^2}{n - K}. \end{aligned} \tag{4.2}$$

Zum Prüfen linearer Hypothesen $R\beta = 0$ (R eine $J \times K$–Matrix vom Rang J) wird die Teststatistik

$$F_{J,n-K} = \frac{(Rb)'(R(X'X)^{-1}R')^{-1}(Rb)}{Js^2} \tag{4.3}$$

eingesetzt.

KQ–Schätzung nach Auffüllen fehlender Werte Yates (1933) schlug folgende Methode vor. Falls m Responsewerte in

y nicht beobachtet wurden, organisiert man den Datensatz um (c: complete):

$$\begin{pmatrix} y_{\text{beob}} \\ y_{\text{fehl}} \end{pmatrix} = \begin{pmatrix} X_c \\ X_* \end{pmatrix} \beta + \begin{pmatrix} \epsilon_c \\ \epsilon_* \end{pmatrix}, \tag{4.4}$$

schätzt β zunächst aus dem vollständigen Submodell gemäß

$$b_c = (X_c' X_c)^{-1} X_c' y_{\text{beob}} \tag{4.5}$$

$(X_c : (n-m) \times K)$ und schätzt den m–Vektor y_{fehl} durch die klassische Vorhersage

$$\hat{y}_{\text{fehl}} = X_* b_c. \tag{4.6}$$

Diese Schätzung wird in (4.4) eingesetzt und danach wird die KQ–Schätzung von β im aufgefüllten Modell berechnet. Die KQ–Schätzung von β im aufgefüllten Modell ist Lösung des Optimierungsproblems

$$\begin{aligned} S(\beta) &= \left\{ \begin{pmatrix} y_{\text{beob}} \\ \hat{y}_{\text{fehl}} \end{pmatrix} - \begin{pmatrix} X_c \\ X_* \end{pmatrix} \beta \right\}' \left\{ \begin{pmatrix} y_{\text{beob}} \\ \hat{y}_{\text{fehl}} \end{pmatrix} - \begin{pmatrix} X_c \\ X_* \end{pmatrix} \beta \right\} \\ &= \sum_{i=1}^{n-m} (y_i - x_i'\beta)^2 + \sum_{i=n-m+1}^{n} (\hat{y}_i - x_i'\beta)^2. \end{aligned} \tag{4.7}$$

Der erste Summand wird minimal für b_c (4.5). Setzt man diesen Wert für β in den zweiten Summanden ein, so wird dieser Ausdruck gemäß (4.6) gleich Null, nimmt also sein absolutes Minimum an. Damit liefert b_c das Minimum der Fehlerquadratsumme $S(\beta)$ (4.7), d.h. b_c ist KQ–Schätzer im aufgefüllten Modell.

Schätzung von σ^2

1. Falls keine Werte fehlen, ist $s^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2/(n-K)$ die korrekte Schätzung.

2. Falls m Daten (y_{fehl} in (4.4)) fehlen, wäre

$$\hat{\sigma}^2_{\text{fehl}} = \sum_{i=1}^{n-m} (y_i - \hat{y}_i)^2/(n-m-K) \tag{4.8}$$

die korrekte Schätzung von σ^2.

3. Die Auffüllmethode von Yates liefert automatisch folgende Schätzung

$$\begin{aligned}\hat{\sigma}^2_{\text{Yates}} &= \left\{ \sum_{i=1}^{n-m} (y_i - \hat{y}_i)^2 + \sum_{i=n-m+1}^{n} (\hat{y}_i - \hat{y}_i)^2 \right\} /(n-K) \\ &= \sum_{i=1}^{n-m} (y_i - \hat{y}_i)^2 /(n-K) \; . \end{aligned} \tag{4.9}$$

Damit gilt

$$\hat{\sigma}^2_{\text{Yates}} = \hat{\sigma}^2_{\text{fehl}} \cdot \frac{n-m-K}{n-K} < \hat{\sigma}^2_{\text{fehl}} \, , \tag{4.10}$$

so dass Yates' Methode zu einer Unterschätzung der Varianz führt. Damit werden Konfidenzintervalle zu klein und die Teststatistiken (vgl. (4.3)) zu groß, so dass Nullhypothesen schneller abgelehnt werden können. Um eine korrekte Analyse zu gewährleisten, müssten also die Schätzung der Varianz und damit alle nachfolgenden Statistiken mit dem Faktor $\frac{n-K}{n-m-K}$ korrigiert werden.

Bartlett's Kovarianzanalyse Bartlett (1937) schlug eine Verbesserung von Yates' ANOVA vor, die als Bartlett's ANCOVA (analysis of covariance) bekannt wurde. Die Methode läuft in folgenden Schritten ab:

1. jeder fehlende Wert wird durch eine beliebige Ersetzung (guess) aufgefüllt: $y_{\text{fehl}} \Rightarrow \hat{y}_{\text{fehl}}$,
2. es wird eine Indikatormatrix $Z \ (n \times m)$ als Kovariable eingeführt und zwar durch die Festlegung

$$Z = \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 0 \\ 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix} . \tag{4.11}$$

   Die $n - m$ Nullvektoren deuten auf no–missing und die m Vektoren $e_i'(e_i$ ist ein Vektor mit Nullen und einer 1 an der i–ten Stelle) auf missing hin. Über diese Kovariablen wird ein zusätzlicher Parameter $\gamma \ (m \times 1)$ in das Modell eingeführt und mitgeschätzt:

$$\begin{aligned} \begin{pmatrix} y_{\text{beob}} \\ \hat{y}_{\text{fehl}} \end{pmatrix} &= X\beta + Z\gamma + \epsilon \\ &= (X, Z) \begin{pmatrix} \beta \\ \gamma \end{pmatrix} + \epsilon. \end{aligned} \tag{4.12}$$

Die KQ–Schätzung von $(\beta\gamma)'$ erhält man durch Minimierung der Fehlerquadratsumme

$$S(\beta,\gamma)=\sum_{i=1}^{n-m}(y_i-x_i'\beta-0'\gamma)^2+\sum_{i=n-m+1}^{n}(\hat{y}_i-x_i'\beta-e_i'\gamma)^2\,. \tag{4.13}$$

Der erste Summand wird minimal für $\hat{\beta}=b_c$ (4.5), der zweite Summand wird minimal (und zwar gleich Null) für $\hat{\gamma}=\hat{y}_{\text{fehl}}-X_*b_c$. Damit ist die Gesamtsumme minimal für $(b_c,\hat{\gamma})$, d.h.

$$\begin{pmatrix} b_c \\ \hat{y}_{\text{fehl}}-X_*b_c \end{pmatrix} \tag{4.14}$$

ist KQ–Schätzung von $\begin{pmatrix}\beta\\ \gamma\end{pmatrix}$ im Modell (4.12). Wählt man als Ersetzung speziell $\hat{y}_{\text{fehl}}=X_*b_c$ (wie bei Yates' Methode), so wird $\hat{\gamma}=0$. Beide Methoden liefern also als Schätzung von β die complete case KQ–Schätzung b_c. Die Einführung des zusätzlichen Paramters γ, an dessen Wert man gar nicht interessiert ist, bietet jedoch einen entscheidenden Vorteil: die Freiheitsgradzahl bei der Schätzung von σ^2 im Modell (4.12) ist gleich T minus Anzahl der geschätzten Parameter, also $n-K-m=n-m-K$ und damit korrekt, d.h. bei Bartlett's ANCOVA erhalten wir $\hat{\sigma}^2=\hat{\sigma}^2_{\text{fehl}}$ (vgl. (4.8)) und damit eine unverzerrte Schätzung von σ^2.

4.1.2 Fehlende Werte in der X–Matrix

Wenn wir die Standardsituation in der mehr ökonometrisch orientierten Regressionsanalyse betrachten, so ist X häufig kein fester Versuchsplan wie in der Biometrie, sondern das Ergebnis von Beobachtungen exogener Variablen. Damit ist X häufig eine Matrix aus zufälligen Variablen, so dass auch in X Beobachtungen fehlen können. Wir können deshalb folgende Struktur antreffen,

$$\begin{pmatrix} y_{\text{beob}} \\ y_{\text{fehl}} \\ y_{\text{beob}} \end{pmatrix} = \begin{pmatrix} X_{\text{beob}} \\ X_{\text{beob}} \\ X_{\text{fehl}} \end{pmatrix} \beta + \epsilon. \tag{4.15}$$

Die Schätzung von y_{fehl} stellt das Vorhersageproblem dar, das wir bereits ausführlich beschrieben haben. Dabei entspricht die klassische Vorhersage der Methode von Yates. Wir können uns deshalb auf die Substruktur

$$y_{\text{beob}} = \begin{pmatrix} X_{\text{beob}} \\ X_{\text{fehl}} \end{pmatrix} \beta + \epsilon \tag{4.16}$$

von (4.15) beschränken und führen folgende Bezeichnungsweise ein:

$$\begin{pmatrix} y_c \\ y_* \end{pmatrix} = \begin{pmatrix} X_c \\ X_* \end{pmatrix} \beta + \begin{pmatrix} \epsilon_c \\ \epsilon_* \end{pmatrix}, \quad \begin{pmatrix} \epsilon_c \\ \epsilon_* \end{pmatrix} \sim (0, \sigma^2 I). \tag{4.17}$$

Das Submodell

$$y_c = X_c \beta + \epsilon_c \tag{4.18}$$

bezeichnet den vollständig beobachteten Datensatz (c: complete), wobei $y_c : (n-m) \times 1$, $X_c : (n-m) \times K$ und Rang $(X_c) = K$ gelten. Wir beschränken uns auf X nichtstochastisch. Bei zufälligem X würden wir mit bedingten Erwartungswerten arbeiten. Das andere Submodell

$$y_* = X_* \beta + \epsilon_* \tag{4.19}$$

hat die Dimension $m = J$. Dabei ist y_* vollständig beobachtet. In der Matrix X_* fehlen Beobachtungen, wobei Einzelwerte oder ganze Spalten oder Zeilen fehlen können. Zur Unterscheidung von der Schreibweise X_{fehl}, die auf vollständiges Fehlen hindeutet, wählen wir die Notation X_* (partially missing). Die Kombination der beiden Submodelle im Modell (4.17) entspricht dem mixed Modell. Es ist deshalb naheliegend, dass wir die Methode der mixed Schätzung zur Behandlung fehlender Werte einsetzen werden. Die optimale, wegen X_* partiell unbekannt aber nicht operationale Schätzung von β im Modell (4.17) ist durch den mixed Schätzer

$$\begin{aligned} \hat{\beta}(X_*) &= (X_c'X_c + X_*'X_*)^{-1}(X_c'y_c + X_*'y_*) \\ &= b_c + S_c^{-1}X_*'(I_J + X_*S_c^{-1}X_*')^{-1}(y_* - X_*b_c) \end{aligned} \tag{4.20}$$

gegeben, wobei

$$b_c = (X_c'X_c)^{-1}X_c'y_c \tag{4.21}$$

der KQ–Schätzer im complete case Submodell (4.18) und

$$S_c = X_c'X_c \tag{4.22}$$

ist.

Die Kovarianzmatrix von $\hat{\beta}(X_*)$ ist

$$\mathrm{V}(\hat{\beta}(X_*)) = \sigma^2 (S_c + S_*)^{-1} \tag{4.23}$$

mit

$$S_* = X_*' X_* \,. \tag{4.24}$$

4.2 Nicht– und semiparametrische Modelle

Eines der Standardwerke in diesem Zusammenhang ist das Buch von Hastie and Tibshirani (1990). Nicht– und semiparametrische Regressionsmodelle bilden eine äußerst flexible Modellklasse, die a priori keinen Funktionstyp für den Zusammenhang zwischen y und X postuliert. Lediglich einige mathematische Eigenschaften, wie etwa Differenzierbarkeit und Stetigkeit der unbekannten Funktion werden vorausgesetzt. Lineare bzw. polynomiale Regressionsmodelle setzen die Kenntnis über den funktionalen Zusammenhang zwischen y und X voraus, was deren Flexibilität und Aussagekraft oftmals deutlich einschränkt. Die Flexibilität nicht– bzw. semiparametrischer Modelle wird allerdings durch komplexere Schätzverfahren 'erkauft'. Ein nichtparametrisches Modell zur Modellierung eines Zusammenhangs zwischen einem Responsevektor y und einer unabhängigen Variablen X wird ganz allgemein geschrieben als

$$y = f(X) + \epsilon \,. \tag{4.25}$$

In diesem einfachen Fall entspricht die Fragestellung dem einfachen Scatterplot–Smoother. Von einem semiparametrischen Modell in diesem Kontext spricht man, wenn etwa ein weiterer linearer Term als auf den Response wirkend angenommen wird, etwa durch

$$y = g(X) + \beta_0 + \ \beta_1 Z + \epsilon\,. \tag{4.26}$$

Ein möglicher Ansatz zur Schätzung der Funktion f sind sogenannte Smoothing Splines, die ihrerseits Lösung des Minimierungsproblems

$$\sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda \int_{-\infty}^{\infty}(f''(u))^2 du \tag{4.27}$$

sind. In diesem Zusammenhang wird häufig von einem Varianz–Bias trade–off gesprochen, d.h. eine geringere Verzerrung (Nähe zu den Daten) kann nur mit einer größeren ‘Rauheit’ der geschätzten Funktion ‘erkauft’ werden. und tragen dem trade–off zwischen der Nähe zu den Daten und Glattheit der geschätzten Funktion Rechnung. Diese Wechselbeziehung erkennt man in Problem (4.27): der linke Term ist ein Maß für die Nähe an den Daten, also den Bias, der rechte Term ein Maß für die Glattheit der Funktion, was dem Parameter λ auch den Namen Glättungsparameter verliehen hat. Zu den Voraussetzungen von f gehören stetige erste und zweite Ableitungen f' und f'', f'' muß quadratisch integrierbar sein. Die Lösung $\hat{f}_\lambda$ von (4.27) wird als natürlicher kubischer Glättungsspline (Beweis siehe etwa Green and Silverman (1994), Kap. 2.5) bezeichnet. Die Problematik fehlender Werte hat allgemeine Grundzüge wie auch modellspezifische Eigenheiten. Während man etwa bei der Imputation über nächste Nachbarn bei Ansatz eines linearen Modells aufgrund der strengen Monotonie der geschätzten Funktion—hier ja einer Geraden—mit der NNI eindeutige Werte über die Spezifikation der Nachbarschaft erhält, ist dies bei glatten Funktionen nicht notwendig mehr der Fall. Folgen die Daten zum Beispiel einem Polynom dritten Grades, so kann der Fall eintreten, dass aufgrund der wechselnden Monotonieeigenschaft mehrere, äußerst unterschiedliche, nächste Nachbarn für die Imputation in Frage kommen. Auf dieses Problem sei lediglich verwiesen, denn den Autoren ist bis dato kein wissenschaftlich anerkanntes Verfahren explizit bekannt. Es sei lediglich erwähnt, dass im Rahmen von Nittner (2003a) erste Ergebnisse durch Simulationsstudien erzielt wurden. Als Fazit kann behauptet werden, dass sich die NNI im Kontext glatter Funktionen durchaus als Alternative, unter gewissen Voraussetzungen gar als bessere Alternative zur Analyse der vollständigen Fälle herausgestellt hat. Selbstverständlich können bekannte und damit auch die hier behandelten Verfahren verwendet werden, zumal bei statistischer Standardsoftware nur begrenzte Möglichkeiten bestehen. Allerdings sollte es uns hier zumindest gelungen sein, dass Bewusstsein in dem Sinne zu schärfen, zumindest auf die Tatsache der Imputation hinzuweisen und sein statistisches Wissen dahingehend zu nutzen, Unterschiede zwischen vollständigen und vervollständigten Daten zu charakterisieren.

4.3 Hinweise und weiterführende Litaratur

Die Literatur zu linearen Modellen ist, gewissermaßen naturgemäß wegen der häufigen anwendung in der Praxis, sehr umfangreich. Wir konnten hier nur einen kleinen Ausschnitt bieten. Ausführlich, mit einigen interessanten Ideen, wie ganz allgemein auf nicht MCAR Prozesse mittels Anwendung von aus den linearen Modellen bekannten Diagnosemaßen —siehe Rao and Toutenburg (1999)—getestet werden kann, beschäftigt sich Fieger (2000). Generalisierte lineare Modelle (McCullagh and Nelder, 1989) mit fehlenden Kovariablen betrachten Ibrahim (1990) und Ibrahim and Weisberg (1992) im ignorierbaren Fall, Ibrahim, Lipsitz and Chen (1999) im nichtignorierbaren Fall. Die Auswahl der Literatur zu nicht– und semiparametrischen Ansätzen und deren Schätzung ist groß. Wir empfehlen neben Hastie and Tibshirani (1990) insbesondere Fahrmeir and Tutz (2001) oder Green and Silverman (1994). Diese enthalten auch die wesentliche Grundzüge über die Wahl der Glättungsparameter, zum Beispiel über *generalized corss validation* oder das *Akaike information criterion.* Ebenso empfehlenswert für alternative Ansätze, etwa über Basisfunktionen, zum Beispiel sogenannte B–Splines, ist die Arbeit von Eilers and Marx (1996). Ein weiteres Konzept basiert auf lokaler Regression deren Schätzung. Dabei werden z.B. Gewichte verwendet, die die Eigenschaften einer Dichtefunktion haben; eine ausführliche Darstellung dieser Methoden ist zum Beispiel in Härdle (1990) zu finden. Eine Auswahl an Literatur zu marginalen Modellen und Modellen mit Zufallseffekten wurde bereits im Abschnitt 3.8 gegeben.

Eine umfangreiche Arbeit von Kastner (2000) beschäftigt sich mit marginalen Modellen für korrelierten Response mit fehlenden Werten im Response. Lipsitz, Ibrahim and Fitzmaurice (1999) betrachten binäre Longitudinaldaten mit fehlenden Werten in Response *und* Kovariablen. Lineare Modelle mit Zufallseffekten und fehlenden Werten im Response werden ausführlich in Verbeke and Molenberghs (2000) behandelt. Generalisierte lineare Modelle mit Zufallseffekten und fehlenden Werten im Response werden nicht bayesianisch in Ibrahim (2001) und bayesianisch in Heumann (2003) behandelt. Daneben gibt es natürlich noch sehr spezielle Methoden für sehr spezielle praktische Probleme. Beispielsweise beschäftigt sich Rässler (2002) mit dem Problem des Zusammenfügens (*merging*) von Datensätzen, wobei sich automatisch fehlende Werte ergeben, wenn gewisse Variablen nicht in allen Datensätzen erhoben wurden. Ein ähnlicher Fall ergibt sich beispielsweise auch, wenn eine sehr teuer zu erhebende Variable nur an wenigen Subjekten erhoben wird, bei der Masse der Subjekte jedoch nur eine Hilfsvariable, welche mit der teuer zu erhebenden Variablen dann eine möglichst große Korrelation haben sollte, erhoben wird.

References

Afifi, A. A., and Elashoff, R. M. (1966). Missing observations in multivariate statistics: Part I: review of the literature, *Journal of the American Statistical Association* **61**: 595–604.

Agresti, A. (1990). *Categorical Data Analysis*, Wiley, New York.

Amemiya, T. (1985). *Advanced Econometrics*, Blackwell, Oxford.

Bartlett, M. S. (1937). Some examples of statistical methods of research in agriculture and applied botany, *Journal of the Royal Statistical Society, Series B* **4**: 137–170.

Buck, S. F. (1960). A method of estimation of missing values in multivariate data suitable for use with an electronic computer, *Journal of the Royal Statistical Society, Series B* **22**: 302–307.

Chen, J., and Shao, J. (2000). Nearest Neighbor imputation for Survey Data, *Journal of Official Statistics* **16**(2): 113–132.

Chen, J., and Shao, J. (2001). Jackknife variance estimation for nearest–neighbor imputation, *Journal of the American Statistical Association* **96**(453): 260–269.

Cohen, J., and Cohen, P. (1983). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*, Lawrence Erlbaum, Hillsdale, NJ.

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

Daniels, M. J., and Hogan, J. W. (2000). Reparameterizing the pattern mixture model for sensitivity analyses under informative dropout, *Biometrics* **56**(4): 1241–1248.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society, Series B* **43**: 1–22.

Diggle, P. J., and Kenward, M. G. (1994). Informative dropout in longitudinal data analysis, *Applied Statistics* **43**: 49–94.

Eilers, P. H. C., and Marx, B. D. (1996). Flexible Smoothing with B-splines and Penalties., *Statistical Science* **11**: 89–121.

Ekholm, A., and Skinner, C. J. (1998). The Muscatine children's obesity data reanalysed using pattern mixture models, *Applied Statistics* **47**(2): 251–263.

Fahrmeir, L., and Tutz, G. (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models*, 2 edn, Springer–Verlag, New York.

Fay, R. E. (1999). Theory and application of nearest–neighbor imputation in census 2000, *Proceedings of the Survey Research Section, American Statistical Association* pp. 112–121.

Fieger, A. (2000). *Fehlende Kovariablenwerte bei Linearen Regressionsmodellen*, Dissertation, Ludwig-Maximilians-Universität München.

Fitzmaurice, G. M., and Laird, N. M. (1995). Regression models for a bivariate discrete and continuous outcome with clustering, *Journal of the American Statistical Association* **90**(431): 845–852.

Gelman, A., and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences, *Statistical Science* **7**: 457–511.

Green, P., and Silverman, B. (1994). *Nonparametric Regression and Generalized Linear Models*, Chapman and Hall, London.

Groves, R. M., Dillman, D. A., Eltinge, J. L., and Little, R. J. A. (2002). *Survey Nonresponse*, Wiley, New York.

Hastie, T., and Tibshirani, R. J. (1990). *Generalized Additive Models*, Chapman and Hall, London.

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

Hedeker, D., and Gibbons, R. D. (1997). Application of random-effects pattern-mixture models for missing data in longitudinal studies, *Psychological Methods* **2**: 64–78.

Heumann, C. (2003). *Monte Carlo methods for missing data in generalized linear and generalized mixed models*, Ludwig-Maximilians-Universität München.

Horowitz, J. L., and Manski, C. F. (2000). Nonparametric Analysis of Randomized Experiments with Missing Covariate and Outcome Data, *Journal of the American Statistical Association* **95**(449): 77–88.

Horowitz, J., and Manski, C. (2001). Imprecise identifcation from incomplete data, *Proceedings of the Second International Symposium on Imprecise Probabilities and Their Applications*, http://ippserv.rug.ac.be/ isipta01/proceddings/index.html.

Horowitz, J., and Manski, C. (2002). Identification and estimation of statistical functionals using incomplete data.

Härdle, W. (1990). *Applied Nonparametric Regression*, Cambridge University Press, Cambridge.

Ibrahim, J. G. (1990). Incomplete data in generalized linear models, *Journal of the American Statistical Association* **85**(411): 765–769.

Ibrahim, J. G. (2001). Missing responses in generalised linear mixed models when the missing data mechanism is nonignorable, *Biometrika* **88**(2): 551–564.

Ibrahim, J. G., Lipsitz, S. R., and Chen, M.-H. (1999). Missing covariates in generalized linear models when the missing data mechanism is non-ignorable, *Journal of the Royal Statistical Society, Series B* **61**(1): 173–190.

Ibrahim, J. G., and Weisberg, S. (1992). Incomplete data in generalized linear models with continuous covariates, *Australian Journal of Statistics* **34**(3): 461–470.

Kastner, C. (2000). *Fehlende Werte bei korrelierten Beobachtungen*, Dissertation, Ludwig-Maximilians-Universität München.

Kenward, M. G. (1998). Selection models for repeated measurements with non-random dropout: An illustration of sensitivity, *Statistics in Medicine* **17**: 2723–2732.

Liang, K.-Y., and Zeger, S. L. (1989). A class of logistic regression models for multivariate binary time series, *Journal of the American Statistical Association* **84**(406): 447–451.

Lipsitz, S. R., Ibrahim, J. G., and Fitzmaurice, G. M. (1999). Likelihood methods for incomplete longitudinal binary responses with incomplete categorical covariates, *Biometrics* **55**(1): 214–223.

Little, R. J. A. (1992). Regression with missing X's: A review, *Journal of the American Statistical Association* **87**(420): 1227–1237.

Little, R. J. A. (1993). Pattern-mixture models for multivariate incomplete data, *Journal of the American Statistical Association* **88**(421): 125–133.

Little, R. J. A. (1994). A class of pattern-mixture models for normal incomplete data, *Biometrika* **81**: 471–483.

Little, R. J. A., and Rubin, D. B. (1987). *Statistical Analysis with Missing Data*, Wiley, New York.

Little, R. J. A., and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*, 2 edn, Wiley, New York.

Little, R. J. A., and Wang, Y.-X. (1996). Pattern-mixture models for multivariate incomplete data with covariates, *Biometrics* **52**: 98–111.

Louis, T. A. (1982). Finding the observed information matrix ewhen using the em algorithm, *Journal of the Royal Statistical Society, Series B* **44**(2): 226–233.

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979). *Multivariate Analysis*, Academic Press, London.

McCullagh, P., and Nelder, J. A. (1989). *Generalized Linear Models*, Chapman and Hall, London.

McLachlan, G. J., and Krishnan, T. (1997). *The EM Algorithm and Extensions*, Wiley, New York.

Michiels, B., Molenberghs, G., Bijnens, L., Vangeneugden, T., and Thijs, H. (2002). Selection models and pattern-mixture models to analyse longitudinal quality of life data subject to drop-out, *Statistics in Medicine* **21**(8): 1023–1041.

Michiels, B., Molenberghs, G., and Lipsitz, S. R. (1999). Selection models and pattern-mixture models for incomplete data with covariates, *Biometrics* **55**: 978–983.

Molenberghs, G., Michiels, B., and Kenward, M. G. (1998). Pseudo-Likelihood for Combined Selection and Pattern-Mixture Models for Incomplete Data, *Biometrical Journal* **40**(5): 557–572.

Molenberghs, G., Michiels, B., Kenward, M. G., and Diggle, P. J. (1998). Monotone missing data and pattern-mixture models, *Statistica Neerlandica* **52**(2): 153–161.

Nielsen, S. F. (2001). Nonparametric conditional mean imputation, *Journal of Statistical Planning and Inference* **99**(2): 129–150.

Nittner, T. (2003a). *Fehlende Daten in Additiven Modellen*, Peter Lang Europäischer Verlag der Wissenschaften, Frankfurt am Main.

Nittner, T. (2003b). Missing at random (MAR) in nonparametric regression - a simulation experiment, *Statistical Methods & Applications* **12**: 195–210.

Nittner, T., and Toutenburg, H. (2004). Identifying Missing Data Mechanisms in (2×2) – Contingency Tables, *SFB386 – Discussion paper 373*, Ludwig-Maximilians-Universität München.

Park, T., and Lee, S.-Y. (1999). Simple pattern-mixture models for longitudinal data with missing observations: Analysis of urinary incontinence data, *Statistics in Medicine* **18**: 2933–2941.

Rao, C. R., and Toutenburg, H. (1999). *Linear Models: Least Squares and Alternatives*, 2 edn, Springer–Verlag, New York.

Rässler, S. (2002). *Statistical matching: a frequentist theory, practical applications, and alternative Bayesian ap-*

proaches, Lecture notes in statistics; 168, Springer, New York.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Sample Surveys*, Wiley, New York.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*, Chapman and Hall, London.

Tanner, M. A. (1991). *Tools for Statistical Inference*, Springer–Verlag, New York.

Toledano, A. Y., and Gatsonis, C. (1999). GEEs for ordinal categorical data: Arbitrary patterns of missing responses and missingness in a key covariate, *Biometrics* **55**: 488–496.

Toutenburg, H. (2002). *Statistical Analysis of Designed Experiments*, 2 edn, Springer–Verlag, New York.

Toutenburg, H., Fieger, A., and Srivastava, V. K. (1999). Weighted modified first order regression procedures for estimation in linear models with missing X-observations, *Statistical Papers* **40**: 351–361.

Toutenburg, H., and Nittner, T. (2002). Linear Regression Models with Incomplete Categorical Covariates, *Computational Statistic* **17**(2): 215–232.

Troxel, A. B., Lipsitz, S. R., and Harrington, D. P. (1998). Marginal models for the analysis of longitudinal measurements with nonignorable non-monotone missing data, *Biometrika* **85**(3): 661–672.

van Dyk, D. A., Meng, X.-L., and Rubin, D. B. (1995). Maximum likelihood estimation via the ECM algorithm: Computing the asymptotic variance, *Statistica Sinica* **5**: 55–75.

Vansteelandt, S., and Goetghebeur, E. (2001). Analyzing the sensitivity of generalized linear models to incomplete outcomes via the ide algorithm, *Journal of Computational and Graphical Statistics* **10**: 656–672.

Verbeke, G., and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*, Springer, New York.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*, Wiley, Chichester.

Wilks, S. S. (1932). Moments and distributions of estimates of population parameters from fragmentary samples, *Annals of Mathematical Statistics* **3**: 163–195.

Yates, F. (1933). The analysis of replicated experiments when the field results are incomplete, *Empire Journal of Experimental Agriculture* **1**: 129–142.

Zaffalon, M. (2002). Exact credal treatment of missing data, *Journal of Statistical Planning and Inference* **105**(1): 105–122.