Seminararbeit:

Übersicht zu fehlenden Daten

Autor: Alexander Pokatilo

Betreuer: Eva Endres

25. Januar 2015

Inhaltsverzeichnis

Abbildungsverzeichnis

1 Einleitung

In der statistischen Praxis muss man oft mit unvollständigen Datensätzen arbeiten. Wenn der Anteil solcher Werte relativ gering ist, werden sie oft ohne Bedenken bei der Analyse ignoriert. Was macht man aber bei einem großen Datenausfall? Was kann ein solcher Ausfall verursachen? Welche Auswirkung kann einfaches Ignorieren von fehlenden Werten auf die Analyseergebnisse haben? Und welche Strategien gibt es zur Behandlung dieses Problems? In Rahmen dieser Seminararbeit wird ein Versuch unternommen, auf alle diese Fragen einzugehen.

1.1 Definition und mögliche Ursachen

Unter fehlenden Werten werden im Sinne der Definition von Spieß [2008] als existierend angenommene und als Reaktion auf einen Reiz hervorrufbare Werte verstanden, deren Beobachtung intendiert ist, die nicht beobachtet werden und nicht ohne Unsicherheit aus anderen Informationen ableitbar sind. Somit werden fehlende Altersangaben in einem Datensatz nicht als fehlend bezeichnet, wenn die Geburtsdaten vollständig erhoben sind. In diesem Fall kann man fehlende Altersangaben der Befragten ohne Unsicherheit aus der Geburtsdaten ableiten. Auch die Daten, die nicht erhoben wurden und somit missing-by-design sind oder bei denen die Existenz eines gültigen Wertes nur angenommen wird, gelten im Sinne obiger Definition nicht als fehlende.
Die Ursachen fehlender Werte sind vielfältig und können auf allen Etappen der Datenerhebung und Aufbereitung zum Datenausfall führen. Hier werden die wichtigsten Ursachen skizziert, eine ausführliche Übersicht findet man zum Beispiel in der Arbeit von Schnell [1986]:

- fehlerhaftes Untersuchungsdesign
- befragtenbedingter Ausfall
- interviewerbedingter Ausfall

- Unvolständigkeit der Sekundärdaten
- Datenaufbereitungsfehler

Unter dem fehlerhaften Untersuchungsdesign versteht man missverständliche Fragen oder nicht erschöpfende Antwortmöglichkeiten. Als Beispiel kann man die Frage zum Alter des Kindes bei einer demografischen Erhebung nennen. Wenn die Antwortkategorie "Habe keine Kinder"nicht gegeben ist, dann tritt bei kinderlosen oder schwangeren Befragten an dieser Stelle zwangsläufig ein fehlender Wert auf.
Bei sensiblen und persönlichen Fragen muss man auf die Formulierung achten und den Befragten Anonymität sichern. Bei besonders sensiblen Fragen können auch spezielle Verfahren (zum Beispiel Randomized-Response-Technik) angewandt werden. Um den durch den Befragten bedingten Datenausfall zu minimieren, soll man sicherstellen, dass die ausgewählten Personen sowohl über die notwendige Kompetenz als auch über die ausreichende Motivation verfügen, um an der Befragung teilzunehmen. Mögliche Probleme können bei den Pre-Tests aufgedeckt und behoben werden. Aber selbst dann ist es nicht sichergestellt, dass einige Teilnehmer manche Fragen übersehen oder die Antwort bewusst verweigern werden.
Wenn die für die Analyse notwendigen Daten in Rahmen des Interviews erhoben werden, dann stellen die Interviewer eine weitere Quelle fehlender Werte dar. Zum einen kann eine Frage aus Versehen oder bewusst dem Befragten nicht gestellt werden (falsche Führung durch Filterfragen), zum anderen kann der Interviewer eine Antwort (bewusst oder unbewusst) falsch als fehlend kodieren. Um solche Fälle zu vermeiden, sollte für jeden Interviewer klar sein, dass sie Fälschungen, Fehlcodierungen der Daten und sachfremde Beeinflussungen der Befragten vermeiden sollen.
Fehlende werte können auch bei dem Einlesen von schriftlichen Fragebögen durch Programmierfehler oder Verlust bzw. Zerstrung der Informationsträger entstehen. Dabei sind Aufmerksamkeit und Vorsicht im Umgang mit Daten geboten.

1.2 Auswirkung

Die meisten Auswertungsmethoden, die in statistischen Software-Paketen implementiert sind, basieren auf einer vollständigen Datenmatrix, bei der alle Untersuchungseinheiten bei allen Variablen einen gültigen Wert aufweisen. Fehlt bei einer Untersuchungseinheit der Wert für mindestens eine Variable, so wird die ganze Einheit von der Analyse automatisch ausgeschlossen. Dieses Verfahren wird später als complete case Analyse nahe

vorgestellt.
Ferner lassen sich die korrekten empirischen Schätzer nur auf dem Basis vollständiger Daten berechnen. Fehlende Werte stellen somit eine Gefahr für eine unverzerrte Schätzung. Eine Untersuchung fehlender Werte und eine Ermittlung möglicher Ursachen spielen bei der Datenanalyse eine wichtige Rolle, da sie über die Anwendbarkeit bestimmter Ersetzungsmethoden entscheiden. Die Wahl ungeeigneter Ersetzungsmethode kann dazu führen, dass die auf Basis des vervollständigten Datensatzes berechnete Schätzer verzerrt werden.
Fehlende Werte sind immer mit einem Verlust an Informationen verbunden, was die Fähigkeit des Datensatzes reduziert, über eine bestimmte Größe Auskunft zu geben. Alle Ergebnisse werden somit unter einem bestimmten Anteil an Unsicherheit ermittelt und stellen eine Einschränkung der Aussagekraft der Studie dar.
Ein Ziel bei der Analyse fehlender Werte besteht somit darin, die Natur des Datenausfalls zu verstehen und eine geeignete Behandlungsmethode zu wählen.

2 Muster fehlender Daten

Das Muster fehlender Werte oder Missing-data pattern gibt einen Überblick über tatsächlich beobachtete und fehlende Werte im Datensatz bzw. beschreibt die globale Struktur fehlender Werte im Datensatz. Einige Analyseverfahren, wie zum Beispiel Likelihood Methoden, benötigen für deren Ausführung bestimmte Datenstrukturen, andere Verfahren können bei beliebigem Muster angewandt werden.
Wenn man den Datensatz $Y = (y_{ij})$ in Form einer (nxp)-Matrix auffasst, wobei n Zeilen die Beobachtungen (Fälle, Personen) und p Spalten die Variablen repräsentieren, so, dass y_{ij} der Wert der Variablen Yj für i-te Beobachtung ist, dann kann man für diese Datenmatrix laut Little and Rubin [2002] eine Indikatormatrix $M = (m_{ij})$ bilden. Die Indikatormatrix M nimmt den Wert $m_{ij} = 1$ ein, wenn y_{ij} einen fehlenden Wert aufweist, und $m_{ij} = 0$ bei einem vorhandenen y_{ij} Wert. Diese Matrix M bestimmt dann das Muster fehlender Werte. Little and Rubin [2002] schlagen auch vor, die Zeilen und Spalten der Datenmatrix so umzustellen, dass sich einige Muster ergeben. Im Hinblick auf die möglichen Ersetzungsmethoden sind für die Analyse vor allem monotones und zufälliges oder allgemeines Muster sowie uni- und multivariater Datenausfall von Bedeutung.

Univariater Datenausfall (item-nonresponse)
Wenn der Datensatz nur bei einer einzelner Variable fehlende Werte aufweist, dann spricht man von univariatem Ausfall oder Item-Nonresponse.

$$M = \begin{pmatrix} 0 & \dots & \dots & 0 & 0 \\ \vdots & \ddots & & \vdots & 0 \\ \vdots & & \ddots & \vdots & 1 \\ 0 & \dots & \dots & 0 & 1 \end{pmatrix}$$

Diese Situation ist zum Beispiel gegeben, wenn einige Teilnehmer einer sozio-ökonomischer Befragung Auskunft zur ihrem Einkommen verweigern oder wenn bei einem Expe-

riment Endergebnise für einige Einheiten aus technischen Gründen nicht aufgenommen wurden. Als Resultat bekommt man ein Muster bei dem die Variable Y_P unvollständig ist und alle andere $Y_1, ...Y_{P-1}$ Variablen vollständig beobachtet sind. In der Praxis es kommt aber auch vor, dass bestimmte Variablen, deren Erfassung schwierig oder mit hohen Kosten verbunden ist, nur bei einem Teil der Untersuchungseinheiten erhoben wird.

Multivariater Datenausfall (unit-nonresponse)

Der Begriff Unit-nonresponse hat in der Literatur zwei Bedeutungen. Little and Rubin [2002] bezeichnen damit eine Situation, bei der fehlende Werte nicht bei einer einzelner, sondern bei mehreren Variablen vorhanden sind, zum Beispiel, wenn eine Person alle Fragen im Einkommensblock (Personaleinkommen, Familieneinkommen, soziale Hilfen) ohne Antwort gelassen hat.

$$M = \begin{pmatrix} 0 & \dots & \dots & 0 & 0 & 0 & 0 \\ \vdots & \ddots & & \vdots & 0 & 0 & 0 \\ \vdots & & \ddots & \vdots & 1 & 1 & 1 \\ 0 & \dots & \dots & 0 & 1 & 1 & 1 \end{pmatrix}$$

Yan and Curtin [2010] verstehen unter Unit-Nonresponse eine komplette Verweigerung der Befragung von einem Teilnehmer. In diesem Fall kann man abhängig vom Untersuchungsdesign entweder neue Teilnehmer auswählen, oder die vorhandenen Teilnehmer im Datensatz lassen. Manchmal kennt man schon bei der Stichprobenziehung einige Angaben der befragten Person aus externer Quellen, auch wenn sie die Teilnahme verweigert hat (Man stellt sich zum Beispiel eine Umfrage unter der Studierenden der LMU vor, bei der im Vorfeld einige Informationen über die zu befragenden Personen etwa Geschlecht, Alter, Studienfach bekannt sind), und kann diese Information bei der Analyse verwenden.

Einen interessanten Fall stellt in diesem Zusammenhang der deutsche Mikrozensus dar, bei dem die Auskunftspflicht der Haushalte per §7 des Mikrozensusgesetzes in Verbindung mit §15 Bundesstatistikgesetz festgelegt ist.

Monotones Muster

Wenn die Variablen in einem Datensatz so geordnet werden können, dass für alle Beobachtungen, bei denen eine Variable Y_j einen fehlenden Wert aufweist, auch alle folgenden

$Y_{j+1}, ..., Y_P$ $J = 1, ..., P1$ Variablen fehlende Werte aufweisen, dann liegt ein monotones Muster vor.

$$M = \begin{pmatrix} 0 & \dots & \dots & \dots & 0 \\ \vdots & & & \iddots & 1 \\ \vdots & & \iddots & \iddots & \vdots \\ \vdots & \iddots & \iddots & & \vdots \\ 0 & 1 & \dots & \dots & 1 \end{pmatrix}$$

Monotones Ausfall-Muster kann man vor allem bei den Longitudinalstudien beobachten, dann wird es auch Panelattrition oder Panelmortalität genannt. Attrition bezeichnet eine Situation, bei der einige Untersuchungseinheiten aus der Studie aus verschiedensten Gründen (wie zum Beispiel Umzung, gesundheitliche Probleme, Tod) vorzeitig ausscheiden und nicht mehr zurückkehren, so dass ab einem Zeitpunkt alle Antworten zu den späteren Fragen fehlen. Monotones Muster kann man auch oft bei der Auswertung langer Fragebögen beobachten, wenn manchen Personen am Ende Zeit oder/und Motivation fehlen die restlichen Fragen zu beantworten.
Little and Rubin [2002] weisen darauf hin, dass in der Praxis Muster fehlender Werte selten monoton sind, aber oft nahezu monoton. Methoden zur Behandlung monoton fehlender Werte sind laut Forscher auch einfacher, als die Verfahren, die man bei einem zufälligen Muster anwenden kann.

Zufälliges Muster

Wenn man die Variablen nicht so anordnen kann, dass ein spezielles Muster sichtbar wird, dann weist ein solcher Datensatz ein zufälliges Muster fehlender Werte auf. Dies bedeutet aber nicht, dass es zwischen fehlenden und beobachteten Werten keine statistischen oder kausalen Zusammenhänge gibt.

$$M = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \end{pmatrix}$$

Ein interessantes Muster fehlender Werte entsteht, wenn man versucht, zwei oder mehrere Datensätze aus unterschiedlichen Quellen mit den Variablen, die nie zusammen beobachtet wurden, zusammenzufügen, vgl. Little and Rubin [2002]:

$$M = \begin{pmatrix} 0 & \dots & \dots & 0 & 0 & 0 & 0 & 1 & 1 \\ \vdots & \ddots & & \vdots & 0 & 0 & 0 & 1 & 1 \\ \vdots & & \ddots & \vdots & 1 & 1 & 1 & 0 & 0 \\ \underbrace{0 & \dots & \dots & 0}_{Y_A} & \underbrace{1 & 1 & 1}_{Y_B} & \underbrace{0 & 0}_{Y_C} \end{pmatrix}$$

In diesem Beispielmuster wurden die Daten aus zwei unterschiedlichen Quellen zusammengefügt. Der Variablensatz Y_A wurde in beiden zusammengefügten Datensätzen erhoben, die Variablen Y_B nur für die erste Datenquelle beobachtet und die Variablen Y_C nur für die zweite. Ein zusätzliches Problem stellt die Tatsache, dass es sich bei den beiden Datensätzen nicht um dieselben Objekte handelt, dar. Das Zusammenfügen und die Behandlung solcher Datensätze ist auch als *statistisches Matching* bekannt.

3 Mechanismen fehlender Daten

Neben dem Muster ist bei der Analyse der fehlenden Werte auch von Interesse, welcher globaler Prozess den Ausfall verursacht. Die Art von diesem fehlende Werte erzeugenden Prozess oder Mechanismus hat eine zentrale Bedeutung für die Lösung des Problems, da er die Anwendung von Behandlungsmethoden einschränken kann. Das Konzept der Mechanismen fehlender Daten wurde von Rubin [1976] vorgeschlagen und später von Little and Rubin [2002] erweitert. Laut Rubin kann der Datenausfall als stochastisches Phänomen gesehen werden, wobei die Indikatoren der fehlenden Werte Zufallsvariablen darstellen. Wie im vorherigen Abschnitt definiert $Y = (y_{ij})$ die vollständige Datenmatrix, wobei Y_{obs} die beobachteten und Y_{mis} die fehlenden Elemente der Matrix bezeichnen. So ergibt sich folgende Formel für den kompletten Datensatz:

$$Y = Y_{obs} + Y_{mis} \tag{3.1}$$

Die Indikator-Matrix für die fehlenden Werte wird durch $M = (m_{ij})$ definiert und stellt eine Vereinigung der Zufallsvariablen dar, die durch eine gemeinsame Verteilung charakterisiert werden können. Und auch wenn man die Verteilung vom M nicht genau spezifizieren muss, so muss man zustimmen, dass diese existiert, vgl. Schafer and Graham [2002]. Die Autoren betonen jedoch, dass es fast unmöglich ist, alle Ursachen des Datenausfalls mit einem statistischen Model zu beschreiben. So betrachtet man die Verteilung von M mehr als ein mathematisches Instrument, mit dessen Hilfe man Anteile und Muster fehlender Daten beschreiben und mögliche Zusammenhänge zwischen dem Fehlen und den Werten fehlender Elemente grob erfassen kann, vgl. Schafer and Graham [2002]. Der Mechanismus der fehlenden Werte wird durch eine durch Y bedingte Verteilung von M mit unbekannten Parametern ϕ beschrieben $f(M|Y, \phi)$.
Laut mittlerweile klassischer Anordnung von Little and Rubin [2002] können fehlende Werte in Abhängigkeit von dem zugrundeligenden Mechanismus als *"missing completely at random"* (MCAR), *"missing at random"* (MAR) oder *"not missing at random"* (NMAR) klassifiziert werden.

3.1 Missing Completely at Random

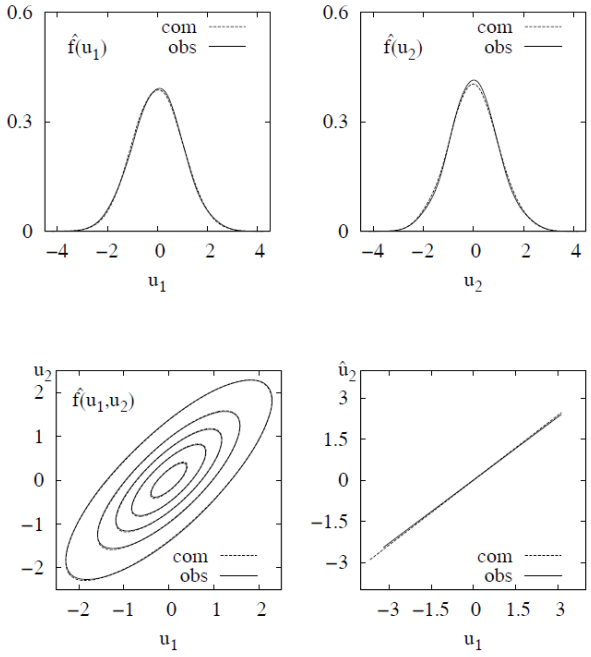
Wir nehmen an, dass die Daten Missing-Completely-at-Random-Bedingung erfüllen und das Fehlen von Werten somit vollkommen zufällig ist, wenn die Wahrscheinlichkeit des Fehlens einzelner Werte in keinerlei Zusammenhang mit den beobachteten und nicht beobachteten Werten in Y steht:

$$f(M|Y_{obs}, Y_{mis}, \phi) = f(M|\phi) \quad \text{für alle } Y_{obs}, Y_{mis}, \phi \tag{3.2}$$

MCAR ist der strengste von drei betrachtenden Mechanismen bzw. Annahmen. Man geht dabei davon aus, dass die Wahrscheinlichkeit für die Beobachtung y_{ij} zu fehlen in keinem Zusammenhang sowohl zu dem wahren Wert von y_{ij}, als auch zu den Werten von anderen Variablen steht. Betrachtet man zum Beispiel einen Datensatz mit den Variablen Alter und Einkommen mit fehlenden Einkommenseingaben für manche befragte Personen, so wäre die MCAR-Bedingung erfüllt, wenn die Wahrscheinlichkeit für das Fehlen der Daten weder vor Einkommensgrösse noch vom Alter der Befragten abhängen würde. Wenn ein Patient wegen Stau nicht zur Kontrolle erscheint, wenn ein Teil der Materialproben im Labor zufällig beschädigt wird, so dass man diese nicht mehr untersuchen kann, oder wenn wegen eines Systemfehlers einige Angaben in der Computerbefragung nicht erfasst werden, kann man annehmen, dass die fehlenden Werte MCAR sind.
In allen solchen Fällen kann man die beobachteten Werte als Resultat einer einfachen Zufallsstichprobe aus der ursprünglich vollständigen Stichprobe auffassen, vgl. Spieß [2008]. Für die Praxis bedeutet dies, dass unter MCAR die Auswertungsergebnisse, die nur auf den vollständigen Beobachtungen basieren, valide sind, wobei manche Informationen verloren gehen.
Anschaulich wird die Auswirkung von MCAR-Daten auf die Analyse im Rahmen einer Simulationsstudie von Spieß [2008]. Dabei wurden 2000 Wertepaare (u_1, u_2) unabhängig voneinander entsprechend einer bivariaten Standardnormalverteilung mit einer Korrelation von 0,8 erzeugt. Danach wurde etwa die Hälfte der Beobachtungen der ersten Variable u_1 völlig zufällig gelöscht. Die fehlenden Werte erfüllen somit die MCAR-Annahme. Auf der Abbildung 3.1 sind oben links und rechts die geglätteten Verteilungen der Variablen u_1 und u_2 vor (mit com für complete bezeichnet) und nach dem Löschen von Daten (mit obs für observed) dargestellt. Auf dem Bild unten links befinden sich die Höhenlinien der bivariaten Dichte und unten rechts die Regressionsgeraden für Regression von u_2 auf u_1. Es sind nur kleine Unterschiede zwischen den Grafiken zu erkennen, was

veranschaulicht, dass nach dem Löschen die beobachteten Werte als eine einfache Zufallsstichprobe aus der ursprünglich vollständigen Stichprobe aufgefasst werden können.

Abbildung 3.1: Uni- und bivariate Dichten sowie Stichprobenregression, vor und nach Löschen der Daten. Fehlende Werte sind MCAR. Spieß [2008]



An dieser Stelle wäre die Frage nach der Prüfbarkeit der MCAR-Bedingung von Interesse. Wenn die MCAR-Bedingung erfüllt ist, dann geht man davon aus, dass für jede Variable Y_j die beiden Teilpopulationen mit $M_j = 1$ und $M_j = 0$ bei allen anderen Variablen $Y_1, ..., Y_k$ mit $k \neq j$ dieselbe Verteilung haben. Bei metrischen Variablen kann man sich auf die Erwartungswerte der beiden Teilpopulationen beschränken und einen einfachen t-Test für unabhängige Stichproben durchführen, wobei die Gleichheit der Erwartungswerte als die Nullhypothese angenommen wird. Signifikante Testergebnisse sprechen gegen die MCAR-Bedingung. Lin and Bentler [2012] deuten aber darauf hin, dass multiples Testen in derselben Stichprobe zur Erhöhung der Alpha-Fehler-Wahrscheinlichkeit führt. Um die Probleme mit dem Fehler 1. Art zu vermeiden, wurde von Little [1988] ein Testverfahren vorgeschlagen, das alle vorhandenen Daten berücksichtigt und eine globale Beurteilung der MCAR-Bedingung ermöglicht. Dieses, einem χ^2-Unabhängigkeitstest ähnliches Testverfahren, besteht im Wesentlichen aus drei Schritten:

- Mit der Hilfe von dem Expectation-Maximization Algorithmus schätzt man die erwartete Mittelwerte und Varianz-Kovarianz-Matrix

- Man gruppiert die Fälle anhand des Musters fehlender Werte und berechnet beobachtete Mittelwerte für jede Gruppe
- Man betrachtet die Summe der Differenzen zwischen den beobachteten und geschätzten Mittelwerten, die mit der geschätzten Varianz-Kovarianz-Matrix und der Anzahl der Fälle in jeder Gruppe gewichtet werden. Die resultierende Teststatistik d^2 ist unter der Null-Hypothese (die Daten sind MCAR) asymptotisch χ^2-verteilt mit den $\sum k_j - k$ Freiheitsgraden, wobei k_j die Anzahl der vollständigen Variablen für j-es Muster und k die Gesamtzahl der Variablen bezeichnen:

$$d^2 = \sum_{j=1}^{J} n_j (\bar{y}_{beob.j} - \hat{\mu}_{beob.j})^T \hat{\Sigma}^{-1}_{beob.j} (\bar{y}_{beob.j} - \hat{\mu}_{beob.j}) \tag{3.3}$$

In der Formel 3.3 bezeichnen $\bar{y}_{beob.j}$ beobachtete Mittelwerte für j-es Muster fehlender Werte, $\hat{\mu}_{beob.j}$ erwartete Mittelwerte und $\hat{\Sigma}^{-1}_{beob.j}$ die geschätzte Varianz-Kovarianz-Matrix.

Fehlende Werte unter MCAR stellen den unproblematischsten Fall dar, der in der Praxis aber selten gegeben ist.

3.2 Missing at Random

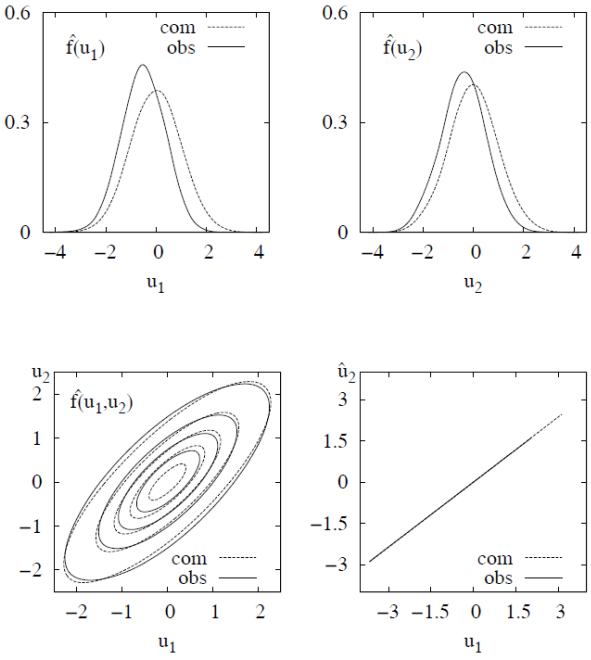
Eine weniger restriktive und damit realistischere Annahme als MCAR ist Missing-at-Random. Dies ist gegeben, wenn das Auftreten von fehlenden Daten von der tatsächlich beobachteten Variablenwerten Y_{obs} von Y abhängt, nicht aber von der fehlenden Werten. Obwohl der von Little and Rubin [2002] vorgeschlagene Name MAR sich durchgesetzt hat, kann er unter Umständen für Missverständnis sorgen und zur Verwechselung mit MCAR führen. Der Name, der der Natur dieses Mechanismus besser entspricht, wäre z.B. missing dependet, abhängig fehlend. Es gilt

$$f(M|Y_{obs}, Y_{mis}, \phi) = f(M|Y_{obs}, \phi) \quad \text{für alle}\, Y_{mis},\, \phi \tag{3.4}$$

Betrachtet man das Beispiel mit den Variablen Einkommen und Alter, so würden die fehlenden Einkommenseingaben die MAR Bedingung erfüllen, wenn das Fehlen vom Alter abhängig wurde (zum Beispiel wenn ältere Personen eher dazu neigen, ihr Einkommen zu verweigern, als jüngere), nicht aber zusätzlich von der Höhe des Einkommens selbst.

Per Definition sind unter MAR alle Informationen über das Auftreten von fehlenden Werten in den beobachteten Werten enthalten. Berücksichtigt man diese Informationen bei der statistischen Analyse, so kann man die fehlenden Werte bei festem Responsepattern als Realisation einer einfachen Zufallsstichprobe betrachtet werden können, vgl. Spieß [2008]. Da unter MAR der zugrundeliegende Mechanismus der fehlenden Werte nicht explizit modelliert werden muss, d.h. ignoriert werden kann, werden solche fehlenden Werte oft als *ignorable nonresponse* bezeichnet (s. z.B. Schafer and Olsen [1998]). Ignorable nonresponse bedeutet aber nicht, dass die fehlenden Werte an sich ignoriert werden dürfen. Ein einfacher Ausschluss der Fälle mit fehlenden Werten führt unter MAR zu verzerrten Analyseergebnissen, was in der folgenden Simulation deutlich gezeigt wird.
Im Rahmen der Simulationsstudie kann man die Auswirkung von MAR-Mechanismus betrachten. Aus dem Bild 3.2 ist es deutlich sichtbar, dass die Analyse nur auf Basis der vollständigen Fälle zu verzerrten Ergebnissen führen kann. Wichtig ist aber die Tatsache, dass die Regressionsbeziehung der beiden Variablen unverzerrt bleibt. Diese Eigenschaft spielt unter anderem bei der Ersetzung der fehlenden Werte mit Hilfe der Methoden der multiplen Imputation wichtige Rolle.

Abbildung 3.2: Uni- und bivariate Dichten sowie Stichprobenregression, vor und nach Löschen der Daten. Fehlende Werte sind MAR. Spieß [2008]



Fehlende Werte unter MAR werden manchmal auch als *ignorable nonresponse* bezeichnet (s. z.B.Schafer and Olsen [1998]).

Einige Behandlungsmethoden von fehlenden Werten zum Beispiel Imputationsverfahren, die im nächsten Kapitel vorgestellt werden, setzen MAR voraus. Einen Test, der Vorliegen von MAR bestätigt, gibt es bis jetzt nicht.

3.3 Not Missing at Random

Wenn die Wahrscheinlichkeit für das Auftreten von fehlenden Werten auch nach der Kontrolle aller Einflussgrössen von den unbeobachteten Werten abhängt, dann liegt ein - für die Analyse der komplizierteste - Not-Missing-at-Random Mechanismus vor. Die Formel lässt sich nicht mehr vereinfachen:

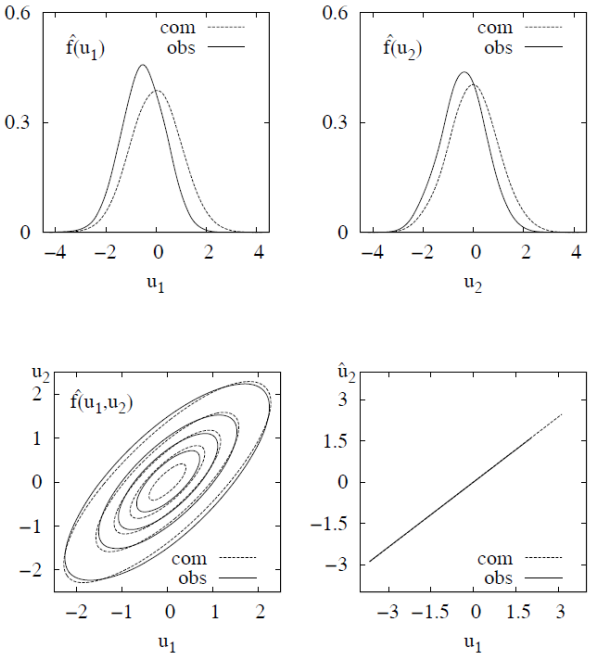
$$f(M|Y,\phi) = f(M|Y_{obs}, Y_{mis}, \phi) \quad \text{für alle}\, Y_{mis},\ \phi \tag{3.5}$$

Bei unbeobachteten Werten kann es sich sowohl um die fehlenden Werte der Variable selbst, als auch um die Werte von anderen nicht erhobenen Variablen handeln. Fehlende Einkommensdaten wären zum Beispiel dann NMAR, wenn die Wahrscheinlichkeit des Fehlens auch nach der Kontrolle auf das Alter von der Höhe des Einkommens selbst abhängen würde. Aus den Simulationen, deren Ergebnisse auf der Abbildung 3.3 dargestellt sind, wird ersichtlich, dass die Daten nach dem Löschen von unvollständigen Fällen deutliche Verzerrungen nicht nur bei uni- und bivariaten Verteilungen, sondern auch bei der Regression aufweisen.
Fehlende Werte unter NMAR werden auch als *nonignorable nonresponse* bezeichnet, da man valide Auswertungsergebnisse nur unter der Betrachtung von dem zugrundeliegenden Mechanismus der fehlenden Werte erzielen kann, Schafer and Graham [2002]. Das Ersetzen von fehlenden Werten alleine auf Basis der in der Stichprobe vorliegenden Informationen ist im Allgemeinen nicht möglich. In diesem Fall muss man die notwendigen Informationen durch externes Wissen ersetzen.
So wie MAR, kann man das Vorliegen von NMAR nicht formal testen, was bedeutet, dass anhand der beobachteten Daten allein nicht beurteilt werden kann, welcher der beiden Mechanismen vorliegt.

3.4 Ignorierbarkeit

Bei der statistischen Analyse der Datensätze mit fehlenden Werten ist der zugrundeliegende Missingmechanismus im Allgemeinen unbekannt. Deswegen stellt sich die Frage,

Abbildung 3.3: Uni- und bivariate Dichten sowie Stichprobenregression, vor und nach Löschen der Daten. Fehlende Werte sind NMAR. Spieß [2008]



ob man diesen unbekannten Mechanismus bei den Auswertungen wie zum Beispiel die Schätzung der Modellparameter oder Stichprobenkenngrößen berücksichtigen soll. Laut Rubin [1976] kann der Mechanismus fehlender Werte ignoriert werden, wenn zwei Bedingungen erfüllt sind, nämlich:

- Fehlende Werte sind MAR
- Die zu schätzenden Parameter θ sind von den Parametern ϕ, die den Missingmechanismus steuern, unabhängig (Distinktheit der Parameter)

Spieß [2008] veranschaulicht die Gültigkeit dieser Regel im Falle von ML-Schätzung der Modellparameter. Als Grundlage für die Schätzung dient die Verteilung der interessierenden Variablen, die auch fehlende Werte aufweisen kann, hier mit $r(Y_{obs}, Y_{mis}; \theta)$ bezeichnet. Um korrekte Schätzung der Parameter zu erzielen, muss man neben der Variablen Y auch die Indikator-Variable M berücksichtigen. Unter MAR-Bedingung 3.4 bekommt man folgende gemeinsame Verteilung:

$$h(Y, M; \theta, \phi) = r(Y_{obs}, Y_{mis}; \theta) f(M|Y_{obs}, Y_{mis}; \phi) = r(Y_{obs}, Y_{mis}; \theta) f(M|Y_{obs}; \phi) \quad (3.6)$$

Da man keine Informationen über Y_{mis} hat, muss man die Formel 3.7 weiter transformieren durch Integration über Y_{mis}:

$$r_{obs}(Y_{obs};\theta)f(M|Y_{obs};\phi) = \int r(Y_{obs},Y_{mis};\theta)f(M|Y_{obs},Y_{mis};\phi)\mathrm{d}Y_{mis} \tag{3.7}$$

Die dazugehörige Log-Likelihood-Funktion besteht aus zwei Summanden:

$$l(\theta,\phi) = \ln r_{mis}(Y_{obs};\theta) + \ln f(M|Y_{obs};\phi) \tag{3.8}$$

Wenn die Parameter θ und ϕ distinkt sind, dann enthält der zweite Summand keine Informationen über θ, und man kann nun zur Schätzung von θ folgende Formel $l(\theta) = \ln r_{obs}(Y_{obs};\theta)$ verwenden.
Auch wenn die Regel von Rubin plausibel erscheint, ist sie in der Praxis nicht so leicht anzuwenden. Wie schon vorher erwähnt, kann man nur anhand der Stichprobeninformationen zwischen MAR und NMAR nicht unterscheiden. Außerdem muss der Mechanismus der fehlenden Werte nicht für alle Einheiten identisch sein. Es ist leicht vorstellbar, dass ein Teil der Befragten die Angaben zum Einkommen, aus welchem Grund auch immer, ganz bewusst verweigert hat. In diesem Fall sind die fehlenden Werte eindeutig NMAR. Ein anderer Teil der Personen könnte die Frage einfach übersehen. Solcher Datenausfall wäre MCAR.

4 Behandlungsmethoden

Zur Behandlung des Problems der fehlenden Werte wurden inzwischen zahlreiche Methoden vorgeschlagen, von relativ einfachen so genannten Ad-hoc-Verfahren über simple Imputation bis aufwändige modellbasierte Methoden. Abhängig von der Art der Behandlung des Datenausfalls kann man alle Methoden in zwei Gruppen aufteilen, nämlich die Ersetzungsmethoden, bei welchen man versucht, fehlende durch plausible Werte zu ersetzen; und die Methoden der Fallreduktion, die fehlende Werte aus der Analyse ausschließen. Klassifikation der Methoden, die in Rahmen dieser Seminararbeit betrachtet werden, ist in der Abbildung 4.1 dargestellt.

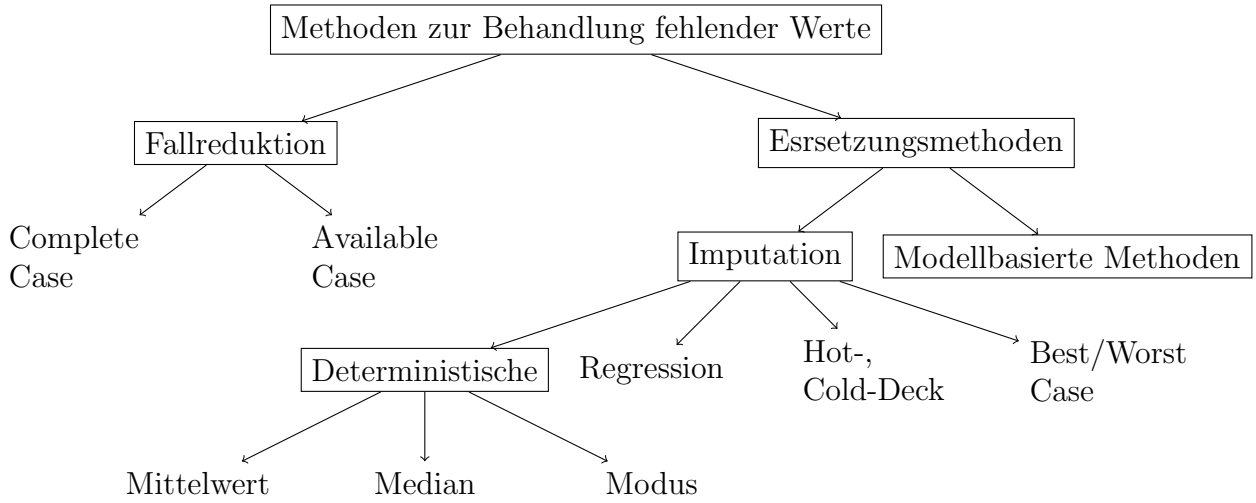


Abbildung 4.1: Klassifikation einiger Verfahren zur Behandlung fehlender Werte

4.1 Fallreduktion

Die einfachste Strategie im Umgang mit der Datensätzen mit fehlenden Werten ist es, die Fälle mit dem Datenausfall bei der Analyse zu ignorieren. Zwei Verfahren, die diese fallreduzierende Vorgehensweise repräsentieren, sind Fallweiser und Paarweiser Ausschluss besser bekannt als Complete Case und Available Case Analyse.

4.1.1 Complete Case Analyse

Complete case Analyse (man findet auch die Bezeichnungen "casewise deletion", "listwise deletion") basiert auf kompletten Fällen, bei denen für jede Variable ein gültiger Wert vorliegt. Alle Einheiten, für die wenigstens ein Wert nicht beobachtet wurde, werden von der Analyse ausgeschlossen. Vorteile dieser Methode sind vor allem die einfache Anwendung, die von allen Standard-Software-Paketen unterstützt wird, und die Vergleichbarkeit univariater Statistiken, da sie alle auf gleicher Stichprobengröße basieren. Ein Nachteil besteht in einem potenziell hohem Informationsverlust vor allem wenn mehrere Variablen im Datensatz fehlende Werte aufweisen. In einem Beispiel zeigen Little and Rubin [2002], dass bei 20 Variablen und 10%-Wahrscheinlichkeit für das Auftreten von fehlenden Werten für jede Variable der erwartete Anteil an komplette Fälle nur $0,9^{20} = 0,12$ beträgt. Dies bedeutet, dass nur 13 % der beobachteten Daten für die Analyse bleibt.

Der Informationsverlust bei complete case Analyse hat zwei Aspekte: zum einen den Präzisionverlust und mögliche Verzerrung der Ergebnisse zu anderem. Deswegen schlagen Little and Rubin [2002] vor, diese Methode nur bei den Daten anzuwenden, die MCAR-Bedingung erfüllen. Da in diesem Fall die beobachteten Daten eine einfache Zufallsstichprobe aus dem ursprünglich vollständigen Datensatz darstellen, ist die Analyse nur der kompletten Fälle unproblematisch. Mittelwerte oder Regressionskoeffiziente, berechnete für reduzierten Datensatz, bleiben unverzerrt. Wenn die MCAR-Annahme verletzt ist, was in der Praxis oft gegeben ist, dann können die entsprechenden Schätzer verzerrt und Inferenz irreführend, vgl. Spieß [2008]. Dies ist vor allem bei NMAR-Daten der Fall. Man stellt sich zum Beispiel vor, dass die Personen mit einem hohen Einkommen eher dazu neigen, dessen Angabe zu verweigern. Unter Anwendung der complete case Analyse wird das mittlere Einkommen der Stichprobe unterschätzt.

Für die Praxis kann es von Interesse sein, den Effizienzverlust beim Fallweiser Ausschluss zu messen. Betrachtet man den Fall von bivariat normalverteilten Daten mit einem monotonem Muster, kann man die Varianz des reduzierten Datensatzes folgendermaßen definieren, vgl. Little and Rubin [2002]:

$$Var(\hat{\theta}_{CC}) = Var(\hat{\theta}_{EFF})(1 + \boldsymbol{\Delta}_{CC}) \tag{4.1}$$

Wobei $\hat{\theta}_{CC}$ die Schätzung für Parameter θ auf Basis der kompletten Fälle, $\hat{\theta}_{EFF}$ die effiziente Schätzung von θ auf Basis aller verfügbaren Informationen und $\boldsymbol{\Delta}_{CC}$ den Effizienzverlust bezeichnen. Ferner ist die Variable Y_1 für alle n Beobachtungen komplett

erhoben und die Variable Y_2 ist nur in r der n Fälle komplett, somit fehlen $n - r$ Beobachtungen. Wenn man die unvollständigen Fälle von der Analyse ausschließt, dann verringert sich die Effizienz bei der Schätzung des Mittelwertes in Y_1:

$$\boldsymbol{\Delta}_{CC} = \frac{n-r}{r} \tag{4.2}$$

So würde sich die Varianz laut 4.1 verdoppeln, wenn die Hälfte der Beobachtungen fehlen würde.
Für den Mittelwert von Y_2 hängt der Effizienzverlust bei complete Case Analyse nicht nur vom Anteil der fehlender Werte, sondern auch von der Korrelation zwischen Y_1 und Y_2:

$$\boldsymbol{\Delta}_{CC} = \frac{(n-r)\rho^2}{n(1-\rho^2) + r\rho^2} \tag{4.3}$$

Die Complete Case Analyse ist dann voll effizient, wenn die unvollständigen Beobachtungen in Y_1 keine Informationen für die Regression von Y_1 und Y_2 erhalten. Dies ist gegeben, wenn die Variablen unkorreliert sind.

Eine mögliche Erweiterung und Verbesserung der Methode stellt die gewichtete complete case Analyse. Dabei versucht man mögliche Verzerrung der Schätzer durch unterschiedliche Gewichtung der vollständiger Fälle anzupassen. Diese z.B. in Little and Rubin [2002] ausführlich diskutierte Methoden sind vor allem bei großen Datensätzen mit beschränkter Kovarianzinformation sinnvoll, wenn mögliche Verzerrung größeres Problem, als Varianz darstellt, da sie über keine Kontrolle über Varianz verfügen.

4.1.2 Available Case Analyse

Für die Analysen, bei denen nur die Randverteilungen betrachtet werden, kann fallweiser Ausschluss zu unnötigen Informationsverlusten führen. Eine Alternative dazu bietet paarweiser Ausschluss oder available case Analyse. Dabei werden alle Einheiten berücksichtigt, bei denen die Werte der für eine spezifische Fragestellung relevanten Variablen beobachtet wurden, Spieß [2008]. Der Vorteil dieser Methode besteht in der Verwendung aller verfügbaren Daten. Der große Nachteil ist dabei die Tatsache, dass die Berechnung der Korrelationen, Mittelwerte, Standardabweichungen jeweils auf unterschiedlichen Fallzahlen basieren. Unter MCAR-Annahme ist eine konsistente Schätzung der Korrelationen zwar möglich, aber es ist auch möglich und nicht selten, dass die be-

rechneten Kovarianz- oder Korrelationmatrizen nicht positiv definit werden, vgl. Howell [2007]. Wenn die Werte nicht MCAR sind, wird diese Vorgehensweise noch problematischer, da die Schätzwerte ohne weitere Modifikationen verzerrt sein können.
Obwohl bei available case Analyse deutlich weniger Daten ungenutzt bleiben, als beim complete case Analyse, gibt es keine eindeutige Meinung darüber, ob paarweiser Ausschluss effizienter als fallweiser Ausschluss ist. Available case liefert effiziente Ergebnisse bei MCAR-Daten und mässigen Korrelationen zwischen Variablen, wie die Simulationen von Kim and Curry [1977] zeigen. Die complete case Methode wird bei großen Korrelationen bevorzugt. Wenn der Anteil der fehlenden Werte gering ist, kann man die Daten paarweise ausschließen, andererseits wäre der fallweiser Ausschluss in diesem Fall auch geeignet und deutlich einfacher. Die Forscher sind sich aber einig, dass beide Methoden generell nicht zufriedenstellend sind (Little and Rubin [2002]).

4.2 Imputationsverfahren

4.2.1 Mittelwertimputation

Die Mittelwertimputation ist ein Verfahren aus der Reihe deterministischer Ersetzungsmethoden. Die fehlende Werte werden dabei einfach durch einen festen vom Forscher bestimmten Wert ersetzt. Als solche als plausibel angesehene Werte können zum Beispiel Lagemaße wie Median, Modus oder Mittelwert verwendet werden. Genauer wird dieses Verfahren anhand der Mittelwertimputation erläutert.
Sei y_{ij} der Wert der metrischen Variable Y_j für i-te Beobachtung. Eine einfache Idee besteht darin, alle n_{mis} von n fehlenden Werten y_{ij} durch das arithmetische Mittel $\bar{y}_j^{(obs)}$ der beobachteten n_{obs} Werte zu ersetzen. Es ist offensichtlich, dass durch solche Imputation der Mittelwert der entsprechenden Variable sich nicht ändert:

$$\bar{y}_j = \frac{1}{n}\sum_{i=1}^{n} y_{ij} = \frac{1}{n_{obs}+n_{imp}}(\sum_{i=1}^{n_{obs}} y_{ij} + n_{imp}\bar{y}_j^{(obs)}) = \frac{1}{n_{obs}+n_{imp}}(\sum_{i=1}^{n_{obs}} y_{ij} + \frac{n_{imp}\sum_{i=1}^{n_{obs}} y_{ij}}{n_{obs}})$$
$$= \frac{1}{n_{obs}+n_{imp}}(\frac{\sum_{i=1}^{n_{obs}} y_{ij}(n_{obs}+n_{imp})}{n_{obs}}) = \frac{1}{n_{obs}}\sum_{i=1}^{n_{obs}} y_{ij}$$

Insbesondere bei nicht normalverteilten Daten wird aber solche Vervollständigung des Datensatzes die Verteilung der Variablen stark beeinflussen. Da alle Werte fehlende Werte duch einen Wert in der Mitte der Verteilung ersetzt werden, führt es dazu, dass die Parameter wie Varianz oder Schiefe unterschätzt werden. Laut Little and Rubin [2002]

wird die Varianz von beobachteten und imputierten Werten um den Faktor $\frac{n_{obs}-1}{n-1}$ unterschätzt. Die Kovarianz $\tilde{s}_{jk}^{jk}$ der vervollständigten Variablen Y_j und Y_k wird um den Faktor $\frac{n_{jk}-1}{n-1}$ unterschätzt, wobei n_{jk} Anzahl der Beobachtungen angibt, für die sowohl Y_j, als auch Y_k beobachtet wurden.
Durch die Korrektur mit $(n-1)/(n_{obs}-1)$ für die Varianz und $(n-1)/(n_{jk}-1)$ für die Kovarianz von Y_j und Y_k können die beiden Parameter prinzipiell konsistent geschätzt werden. Allerdings bedeutet dies Abweichungen von der Standardanalyse. Die korrigierten Schätzer sind auch genau die, die man auch bei available case Analyse verwenden würde. Die resultierenden Korrelationmatrizen können auch nicht positiv definit sein. Eine weitere Gefahr der Methode besteht darin, dass die durch Mittelwert imputierten Werte unrealistisch für die Variable sein können. Ein Beispiel dafür ist ein Mittelwert von 1,5 bei der Variable Änzahl der Kinder". Die fehlenden Werte durch diesen Wert zu ersetzen wäre zwar mathematisch korrekt, aber inhaltlich sinnlos.

Tabelle 4.1: Gehalt und Zitierungsniveau der Publikationen von Professoren

Analyse	**N**	**r**	β	**St.Error**
Complete cases	62	.55	310.747	60.95
Mittelwertimputation	69	.54	310.747	59.13

In der Tabelle 4.1 sind die Ergebnisse der Regressionsanalyse von Cohen et al. [2013] präsentiert. In Rahmen der Studie versuchten die Forscher ein Modell zu entwickeln, das den Gehalt der Professoren einer Universität allein anhand der Zitierungen ihrer Publikationen vorhersagt. Für 62 von 69 Personen lagen die vollständigen Informationen vor, in restlichen sieben Fällen fehlten die Angaben zum Zitierungsniveau. Die Ergebnisse der complete case Analyse ($N = 62$) sind in der ersten Zeile der Tabelle dargestellt, die Auswertung des mit Mittelvertimputation vervollständigten Datensatzes ($N = 69$) sieht man in der zweiten Zeile. Es fällt sofort auf, dass die β-Koeffizienten der beiden Modelle identisch sind. Die Standardabweichung des zweiten Koeffizienten ist aber deutlich kleiner. Durch Addition von sieben Fällen, dessen Abweichung von Mittelwert Null beträgt, hat man keine neue Information zugefügt. Die Stichprobegröße wurde dabei aber erhöht, was zur künstlichen Schrumpfung der Standardabweichung führt. Aus diesem Grund raten Little and Rubin [2002] von dieser Methode ab, schlagen aber die bedingte Mittelwertimputation als mögliche Verbesserung vor.

Wenn für die Analyse nicht metrische Daten vorliegen oder wenn man die Robustheit gegen Ausreißern erreichen will, kann man anstelle des arithmetischen Mittelwertes Me-

dian oder Modus einsetzen. Dies ist aber mit ähnlichen Problemen verbunden, wie bei der Verwendung unbedingter Mittelwerten.

4.2.2 Regressionsimputation

Zur Ersetzung fehlender Werte bei stetigen Variablen bietet sich die Regressionsimputation. Diese Methode setzt multivariate Normalverteilung und ein monotones Muster fehlender Werte voraus, bei denen mindestens MAR-Annahme erfüllt ist. Wie der Name dieser Methode bereits verrät, werden fehlende Werte durch anhand eines Regressionsmodells berechnete Werte ersetzt. Dabei werden im Datensatz diejenige Variable $y_{obs}^{(i)}$ ausgesucht, die mit der Responsevariable $y_j^{(i)}$ (der Variable mit teilweise fehlenden Werten) im Zusammenhang stehen und vollständig erhoben sind. Diese werden dann als unabhängige Variablen in Modell genommen:

$$y_j^{(i)} = y_{obs}^{(i)}\beta^{(i)} = \beta_0^{(i)} + \beta_1^{(i)}y_1^{(i)} + \cdots + \beta_{j-1}^{(i)}y_{j-1}^{(i)} \tag{4.4}$$

Wenn sich im Datensatz mit einem monotonem Muster fehlende Werte mehrere unvollständige Variablen befinden, kann die Regressionsmethode interativ f"r weiter Variablen durchgeführt werden, so Hohl [2007]. Diese Methode weist aber auch dieselbe Nachteile, wie die oben genannte Mittelwertimputation, nämlich mögliches Auftreten unrealistischer Werten, die mit dem Regressionsmodell berechnet werden, und eine Unterschätzung der Varianz, da das Ergänzen von Werten, die durch andere Größen im Datensatz perfekt abgeleitet werden können, addiert keine neue Information, vergrößert aber die Stichprobe.
Eine Erweiterung von Single Regressionsimputation stellt die Multiple Stochastische Regressionsmethode dar. Diese addiert zur vorhergesagten Werten einen stochastischen Term (Residuum der Regressionsgleichung). Damit wird die Streuung der ersetzten Werten erhöht und man kann mehrere Varianten des vollständigen Datensätzes generieren. Durch bilden von Mittelwerten von mehreren Datensätzen erhält man die endgültige Ersetzungswerte, die Unsicherheit der Schätzung berücksichtigen.

4.2.3 Hot-Deck und Cold-Deck-Methoden

Wie bereits erwähnt können bei Regressions- oder Mittelwertimputation durchaus unrealistische Werte auftreten, die in der Stichprobe nicht vorkommen oder sogar außerhalb eines beobachteten Intervalls liegen. Eine Lösung für dieses Problem bieten

"Hot-Deck"bzw. "Cold-DeckTechniken, die zur Gruppe nicht parametrischen Verfahren gehören.

Hot-Deck Methode

Um eine Hot-Deck-Methode anzuwenden, müssen fehlende Werte mindestens MAR sein. Muster fehlender Werte spielt hier keine entscheidende Rolle, wobei die Durchführung bei Vorliegen eines monotonen Musters unkomplizierter ist.

Der Name "Hot-Deck"geht auf das Lesen von sogenannten Hollerith-Lochkarten für IBM Computer zurück, die bei der Anfrage schnell heiß wurden. Die Idee hinter der Hot-Deck-Methode besteht darin, die fehlende Werte durch in derselben Stichprobe erhaltene Werte zu ersetzen. Dabei wird ein Spenderdatensatz erstellt, der aus der Beobachtungen besteht, die ähnliche Charakteristiken wie die Beobachtung mit fehlendem Wert aufweisen und für die interessierende Variable einen gültigen Wert haben. Betrachtet man eine sozio-ökonomische Befragung, bei der ein vierunddreißigjähriger Ingenieur aus Bayern keine Angabe zu seinem Einkommen gemacht. Um diesen fehlenden Wert zu ersetzen, erstellt man einen Spenderdatensatz aus ähnlichen Personen (männlich, Alter zwischen dreißig und fünfunddreißig, Ingenieur, Bundesland Bayern), die diese Frage beantwortet haben, und zieht aus diesem zufällig einen Einkommenswert. Die zufällige Ziehung kann mit und ohne Zurücklegen erfolgen, sowie unter Verwendung anderer Ziehungsdesigns, S. Little and Rubin [2002]. Zur Ersetzung kann man auch den Mittelwert der ähnlichsten k Beobachtungen aus den kompletten Daten verwenden.

Beim Vorliegen von Zeitreihen oder Längsschnittdaten bietet sich die sequentielle Hot-Deck-Methode. Dabei werden die Beobachtungen in Abhängigkeit vom Zeitpunkt der Erhebung sequentiell angeordnet. Zuletzt beobachtete Wert in jeder Sequenz und für jede Variable wird dann für fehlender Wert eingesetzt ("last observation carried forward"), vgl. Spieß [2008].

Der Vorteil der Hot-Deck-Methode ist die große Variabilität in den ersetzten Werten, was zu einem realistischen Datensatz mit besseren Verteilungseigenschaften (die Verteilung wird nicht so stark verzerrt wie bei Mittelwertimputation) führt. Ein Kritischer Aspekt dieser Methoden ist der Wahl der Variablen, welche Ähnlichkeit der Beobachtungen beurteilen, und wie stark die Ausprägungen dieser Variablen von der Beobachtung mit fehlendem Wert abweichen können, um dennoch als ahnlich"klassifiziert zu werden (Hohl [2007]). Im obigem Beispiel ist den Forschern alleine überlassen, welche Altersgrenze für die potenziellen Spender eingesetzt werden, ob Geschlecht, Beruf und Bundesland bei dem Wahl mitberücksichtigt werden. Der Wahl dieser Variablen ist subjektiv und kann nicht formalisiert werden. Ferner je mehr Variablen zur Erstellung des Spenderdaten-

satzes verwendet werden, desto weniger fehlender Werte sollen im Datensatz auftreten. Sonst kann es zu Situationen führen, wo der Spenderdatensatz für einige Einheiten mit fehlenden Werten leer ist, vgl. Hohl [2007]. Die Anwendung dieser Methode ist mit einem großen programmiertechnischen Aufwand verbunden ist, weil für jede Variable mit fehlenden Werten einen Spenderdatensatz definiert werden muss.

Cold-Deck Methode
Die Cold-Deck Methode hat keine Voraussetzungen bezüglich des Mechanismus oder Muster fehlenden Werte. Spenderdaten für Durchführung dieser Methode liefern externe Datensätze oder sonstige Informationsquellen derselben Population. Dabei können fehlende Werte durch einen glichen Wert oder durch verschiedene Werte, abhängig von den Werten anderer Hilfsvariablen, (Hohl [2007]).
Cold-Deck-Technik ist einfach anzuwenden, wenn die vorliegende Daten vergleichbar sind, was in der Praxis aber selten der Fall ist, was die Anwendbarkeit deutlich einschränkt. Außerdem muss es begründet werden, warum bestimmte externe Datenquelle ein geeigneter Ersatz für fehlende Werte bietet.

Sowohl Hot-, als auch Cold-Deck Methoden als Grundlage für eine konsistente Schätzung der Parameter verwendet werden. Man muss aber beachten, dass solche Imputationen, solange sie nicht ohne Unsicherheit aus der vorliegenden Informationen abgeleitet werden können, nur Schätzungen für fehlende Werte darstellen. Die mit jeder Schätzung verbundene Unsicherheit soll bei weiteren Analysen unbedingt berücksichtigt werden, sonst werden die Standardfehler entsprechender Parameter systematisch unterschätzt, was zu Fellinterpretationen führen kann.

4.2.4 Best/Worst Case Methode

Für fehlende Werte in klinischen Studien, in denen der Unterschied zwischen zwei Gruppen (Test- und Kontrollgruppe) statistisch untersucht wird, kommen zwei weitere Ersetzungstechniken in Frage, nämlich die Best bzw. Worts Case Methode. Ähnlich zu Hot- und Cold-Deck setzen diese keine bestimmte Muster und Mechanismen fehlender Werte voraus und können bei jedem Merkamlstyp angewandt werden.
Die worst case Methode steht für konservatives Testen. Dabei werden fehlende Werte in der Gruppe mit den vermuteten besseren Werten durch den schlechtesten beobachteten Wert dieser Gruppe ersetzt, in der anderen Gruppe wird umgekehrt den besten beobachteten Wert verwendet. Dadurch wird der mögliche Zielgrößenunterschiede zwischen

Gruppen verkleinert, was ein statistisch signifikanter Unterschied schwieriger nachweisbar macht, Hohl [2007]. Die best case methode funktioniert genau umgekehrt und stellt eine optimistische Schätzung des Effekts dar. Durch absichtliche Verkleinerung bzw. Vergrößerung der Zielgrößenunterschiede kann man Hinweise dazu gewinnen, wie sich die Effektgröße im schlimmsten bzw. besten Fall entwickelt wird, vgl. Mayer [2010].

5 Zusammenfassung

Im Rahmen dieser Seminararbeit wurde ein Überblick zum Problem fehlender Werte gegeben. Dabei wurden mögliche Ursachen und Konsequenzen des Problems erläutert und zwei wichtige, von Little and Rubin [2002] entwickelte, Konzepte der Muster und der Mechanismen fehlender Werte präsentiert. Anhand der Informationen über die globale Anordnung fehlender Werte im Datensatz und deren Natur können geeignete Behandlungsmethoden ausgewählt werden.
Die schließlich vorgestellten Techniken zählen zu sogenannten Ad-hoc Verfahren, die einerseits einfach anwendbar sind, andererseits viele Nachteile für weitere Analysen aufweisen. Einige von diesen Verfahren wurden in den Zeiten leistungsschwacher Technik entwickelt und sind heutzutage nicht mehr aktuell, andere, wie zum Beispiel complete case Analyse, werden in der modernen Forschung trotzdem sehr häufig eingesetzt. Was man bei kleinem Anteil an fehlende Daten mit geringerem Zeit- und Kostenaufwand rechtfertigen kann. In anderen Fällen ist es zu raten, auf kompliziertere Ersetzungsmethoden wie Multiple Imputation zuzugreifen.

Literaturverzeichnis

Jacob Cohen, Patricia Cohen, Stephen G West, and Leona S Aiken. *Applied multiple regression/correlation analysis for the behavioral sciences*. Routledge, 2013.

Kathrin Hohl. *Ersetzungsmethoden für fehlende Wertekategorialer Variablen in klinischen Datensätzen*. dissertation, Universitt Ulm, 2007. URL http://www.google.de/url?sa=t&rct=j&q=&esrc=s&source=web&cd=2&cad=rja&uact=8&ved=0CCcQFjAB&url=http%3A%2F%2Fvts.uni-ulm.de%2Fdocs%2F2007%2F6027%2Fvts_6027_8103.pdf&ei=Q4QFVZ6dOYHyUIOJgrgC&usg=AFQjCNFBzyv9et5TIkOBfrIw7pD3N3neuQl.

David C Howell. The treatment of missing data. *The Sage handbook of social science methodology*, pages 208–224, 2007.

Jae-On Kim and James Curry. The treatment of missing data in multivariate analysis. *Sociological Methods & Research*, 6(2):215–240, 1977.

JC Lin and Peter M Bentler. A probability based test for missing completely at random data patterns. In *meeting of the National Council on Measurement in Education, Vancouver, Canada*, 2012.

R.J.A. Little and D.B. Rubin. *Statistical Analysis with Missing Data*. Wiley Series in Probability and Statistics. Wiley, 2002. ISBN 9780471183860. URL https://books.google.ru/books?id=aYPwAAAAMAAJ.

Roderick JA Little. A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, 83(404):1198–1202, 1988.

Benjamin Mayer. *Fehlende Werte in klinischen Verlaufsstudien - Der Umgang mit Studienabbrechern*. dissertation, Universitt Ulm, 2010. URL http://www.google.de/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&uact=8&ved=0CCIQFjAA&url=http%3A%2F%2Fvts.uni-ulm.de%2Fdocs%2F2011%2F7633%2Fvts_7633_10939.pdf&ei=J7EFVaXiEMb5UrfHgOgL&usg=AFQjCNFkhRzm74nI1gwpDivzUpi3cu9R9w.

Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

Joseph L Schafer and John W Graham. Missing data: our view of the state of the art. *Psychological methods*, 7(2):147, 2002.

Joseph L Schafer and Maren K Olsen. Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate behavioral research*, 33(4): 545–571, 1998.

Rainer Schnell. Missing-data-probleme in der empirischen sozialforschung. 1986.

Martin Spieß. *Missing-Data-Techniken: Analyse von Daten mit fehlenden Werten*. LIT Verlag Münster, 2008.

Ting Yan and Richard Curtin. The relation between unit nonresponse and item nonresponse: A response continuum perspective. *International Journal of Public Opinion Research*, page edq037, 2010.