Ludwig-Maximilians-Universität München

Institut für Statistik

Seminararbeit

Imputationsverfahren

Autorin:
Minh Ngoc Nguyen

Betreuerin:
Eva Endres

15. März 2015

Seminararbeit

Imputationsverfahren

Autorin: Minh Ngoc Nguyen

Betreuerin: Eva Endres

15. März 2015

Abstract

Fehlende Daten stellen ein häufiges Problem im Rahmen empirischer Untersuchungen dar. Multiple-Imputation oder Mehrfachergänzung in den fehlenden Werten werden sich in den letzten Jahren als intuitive und flexible Methode, die mit unvollständigen Datensätzen umgeht, erworben. Deshalb soll in dieser Seminararbeit ein Überblick über Multiple-Imputation geschaffen werden. Die multiple imputierte Datensätze können wie vollständige Datensätze mit statistischen Standardmethoden ausgewertet werden. Unter anderem werden zwei Verfahren vorgestellt, die bei der Durchführung einer Multiple-Imputation dafür sorgen, dass die fehlenden Werte ersetzt werden: likelihood-basierte Verfahren mit Expectation Maximization (EM) und bayes-basierte Verfahren mit Data Augmentation (DA) und Multiple Imputation by Chained Equation (MICE).

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einführung

Fehlende Beobachtungen bzw. Daten stellen ein häufiges Problem in vielen Erhebungen oder Experimenten in Wissenschaft und Praxis dar. Die Gründe für die fehlenden Beobachtungen von Daten können dabei sehr vielfältig sein, z.B. ein Interview wird abgebrochen, Teilnehmer verweigern die weitere Teilnahme in einer Langzeitbeobachtung, Fragen werden übersehen, Informationen sind nicht mehr verfügbar etc. Problematisch ist dies, da die Theorie der meisten statistischen Analysemethoden nur auf den Idealfall vollständiger Datensätze ausgerichtet ist. Wendet man diese auf fehlende Daten an, so kann dies zu folgenden Problemen wie Informationsverlust, geringere Fallzahlen für die statistische Analyse und Verzerrung der Ergebnisse führen (Little u. Rubin, 2002), da Einheiten mit fehlenden Daten aus der Schätzung teil oder ganz entfernt werden. Vor diesem Hintergrund ist es naheliegend, anstelle alle, d.h. die fehlenden Beobachtungen aufzufüllen, d.h. zu ersetzen. Im englisch sprachigen Raum hat sich dafür der Begriff der „Imputation“ eingebürgert, der sich ebenso im Deutschen verwenden lässt. Es gibt nun eine Vielzahl von Imputationsmethode, die alle ihre Vor- und auch Nachteile haben (Rässler, 2000). Dabei unterscheidet man grob zwischen der Single- und der Multiple-Imputation. Single-Imputation ersetzt für jeden fehlenden Wert nur einen plausiblen Wert, beispielsweise durch den Mittelwert der vorhandenen Beobachtung in diese Variable. Hauptproblem ist, dass diese Methode die Unsicherheit bei der Imputation nicht berücksichtigt. Aus diesem Grund führt es in allgemein zu unterschätzten Varianz bzw. zu fälschlicherweise signifikanten Ergebnissen (Little u. Rubin, 2002). Eine Weiterentwicklung stellen die Methoden der Mutiple-Imputation, die von Rubin (1987) eingeführt wird, dar. Statt nur einen Wert einzusetzen, generiert man mehrere Imputationen und somit die Unsicherheit über die unbeobachteten tatsächlichen Werte Rechnung getragen wird. Unter anderem spielen zwei Verfahren wichtige Rolle, die bei der Durchführung einer Multiple-Imputation dafür sorgen, dass die fehlenden Werte ersetzt werden. Ersten ist das likelihood-basierte Verfahren mit Expectation Maximization (EM). Zweiten ist bayes-basierte Verfahren mit Data Augmentation (DA) und Multiple Imputation by Chained Equation (MICE).

In dieser Seminararbeit sollten theoretische Grundlagen der Multipe-Imputation vorstellen werden. Im Kapitel 2 wird daher kurz auf die Ursachen fehlender Daten, die sogenannte Fehlendmechanismus, und die Ignorierbarkeit eingegangen. Daraus folgend präsentiert Kapitel 3 die theoretischen Konzepte der Multiple-Imputation. Anschließend werden verschiedene Verfahren in Kapitel 4 vorgestellt, die bei der Durchführung einer Multiple-Imputation dafür sorgen, dass die fehlenden Werte ersetzt werden. Kapitel 5 beschäftigt sich mit einem Beispiel aus der Simulation, und soll dabei zu deren Veranschaulichung dienen. Abschließend fasst Kapitel 6 mit einem Überblick über die Kernaussage der Arbeit deren wichtigste Punkte zusammen.

2 Fehlendmechanismen und Ignorierbarkeit

Der Begriff des Fehlendmechanismus, auch Missing Data Mechanisms genannt, geht auf Rubin (1987) zurück, bzw. deren Neuauflage Little u. Rubin (2002). Fehlendmechanismus formuliert Annahme über das Verursachen für das Auftreten fehlender Werte. Dies ist wichtig, weil man nur auf Basis der Kenntnis des Fehlendmechanismus eine angemessene Behandlung des Problems durchführen kann.
Im Folgenden soll Y die Datenmatrix bezeichnen, in der die n Zeilen die Beobachtungen und die p Spalten die Variablen repräsentieren. Der Teil der beobachteten Werte lässt sich als Y_{beob} und der fehlende als Y_{fehl} dargestellen. Es besteht $Y = (Y_{beob}, Y_{fehl})$. Man definiert zusätzliche dafür eine Indikatormatrix, welche das Fehlen bzw. Nichtfehlen eines Wertes ausdrückt.

$$R = \begin{cases} 1 & falls\ Y\ fehlend\ ist \\ 0 & falls\ Y\ beobachtet\ ist \end{cases}$$

Man behandelt diese Indikatorvariablen als Zufallsvariablen und ordnet ihnen eine Verteilung zu. Der Fehlendmechanismus lässt sich als die bedingte Verteilung von R gegeben Y beschrieben

$$g(R|Y, \xi) = g(R|Y_{beob}, Y_{fehl}), \xi) \tag{2.1}$$

Dabei steht ξ für die unbekannten Parameter, der den Fehlendmechanismus steuert.
Little u. Rubin (2002) hat die folgenden Mechanismen benannt: Missing Completely At Random (MCAR), Missing At Random (MAR) und Missing Not At Random (MNAR).

MCAR: Missing Completely At Random
Fehlende Werte sind MCAR, wenn die Wahrscheinlichkeit für das Fehlen der Werte im gesamten Datensatz weder von den beobachteten noch von den unbeobachteten Werten abhängt. Formal gilt:

$$g(R|Y,\xi) = g(R|\xi) \tag{2.2}$$

Das heißt, dass in diesem Fall die Wahrscheinlichkeit für das Fehlen eines Wertes sogar komplett unabhängig vom Datensatz ist. Als Beispiel betrachtet man die Variablen Alter und Einkommen mit fehlenden Einkommensangaben (Spieß, 2008). MCAR liegt vor, wenn die Wahrscheinlichkeit für das Fehlen des Einkommens unabhängig von Einkommen und Alter ist. Dies ist beispielsweise der Fall, wenn Daten zufällig verloren gegangen sind, z.B. Datenverlust des Rechners. In der Praxis tritt MCAR aber selten auf.

MAR: Missing At Random
Fehlende Werte sind MAR, wenn das Fehlen der Werte von den beobachteten, nicht aber von den unbeobachteten Werten abhängt. Formal gilt:

$$g(R|Y,\xi) = g(R|Y_{beob},\xi) \tag{2.3}$$

Dies bedeutet, dass die Wahrscheinlichkeit für das Fehlen eines Wertes von einem anderen beobachteten Variablenwerten abhängig ist. Im Beispiel mit den Variablen Alter und Einkommen, liegt MAR vor, wenn die Wahrscheinlichkeit des Fehlens des Einkommens vom Alter abhängt. Man könnte sich vorstellen, dass ältere Personen nicht mehr bereit sind, ihr Einkommen zu geben, als jüngere Personen. Diese Wahrscheinlichkeit hängt allerding nicht von der Höhe des Einkommens selbst ab.

MNAR: Missing not at random
Das Vorhandensein eines MNAR Mechanismus stellt den schwierigsten Fall dar. Fehlende Werte sind MNAR, wenn das Fehlen der Werte von den unbeobachteten Werten abhängig ist und zwar selbst dann, wenn für beobachtete Variablen kontrolliert wird. Formal gilt:

$$g(R|Y,\xi) = g(R|Y_{beob},Y_{fehl},\xi) \tag{2.4}$$

Das Fehlen der Werte hängt also von den fehlenden Werten ab. Als plausibles Beispiel kann man das Verweigern von Angabe zum Einkommen von Personen, die ein hohes oder niedriges Einkommen habe, betrachten. Die Wahrscheinlichkeit, dass das Einkommen fehlt, hängt selbst nach Konditionieren auf das Alter, vom Einkommen selbst ab.

In Ergänzung zu formalen Konzepten von MCAR, MAR und MNAR wurde von Little u. Rubin (2002) die Ignorierbarkeit von Fehlendmechanismen definiert. Die Ignorierbarkeit kann anhand von zwei Bedingungen an den Fehlendmechanismus formuliert werden (Ru-

bin, 1976). Ersten, dass die Daten MAR oder MCAR sind. Zweitens, dass die Parameter des Ausfallmechanismus ξ keine Informationen über die Parameter θ der interessierenden Verteilung $f(Y_{beob}, Y_{fehl}|\theta)$ enthalten und vice versa dass ξ und θ distinkt sind.

Das Vorliegen eines ignorierbaren Fehlendmechanismus ist Voraussetzung für die meisten Imputationsverfahren.

3 Multiple-Imputation

3.1 Grundkonzepte

Multiple-Imputation wurden von Rubin (1987) vorgeschlagen.Es handelt sich um Methoden, die für jeden fehlenden Wert m plausible Werte ersetzen, wobei $m \geq 1$ gilt. Dies führt zu m verschiedenen vervollständigten Datensätzen.
Das Verfahren der Multiple-Imputation lässt sich in drei Schnitte aufteilen: Imputation, Analyse und Kombination. Im ersten Schritt werden für jeden fehlenden Wert m Werte geschätzt und eingesetzt. Je nach Anzahl der unterschiedlichen Werte, die für einen fehlenden Wert imputieren sollen, ergibt sich im ersten Schritt eine gewisse Anzahl an Datensätzen. Will man beispielsweise für einen fehlenden Wert 3 verschiedene Werte imputieren, erhält man dementsprechend 3 unterschiedliche Datensätze. Nach dem Imputationsschritt liegen m vervollständigte Datensätze, die somit die gleichen beobachteten Werte und unterschiedliche imputierte Werte an den ursprünglich fehlenden Stellen enthalten, vor. In dem Analyseschritt werden dieselben Methoden und Standardsoftware durchgeführt, die man auch für einen vollständigen Datensatz anwenden würde. Dies muss allerdings m-mal getrennt voneinander geschehen, so dass schließlich m unterschiedliche Ergebnisse vorliegen. Die Ergebnisse werden anschließend im letzten Schritt zu einem Ergebnis zusammengefasst. Das Vorgehen der Multiple-Imputation lässt sich am besten grafisch veranschaulichen (vgl. Abbildung 3.1).
Die m imputierten Datensätze spiegeln die Unsicherheit wider, welche durch das Schätzen entstanden ist. Somit werden im letzten Schritt der Integration durch Kombinieren der Ergebnisse (Rubin, 1987) allgemeine Schätzer und Standardfehler erzeugt, welche die Unsicherheit der Daten reflektieren.

3.2 Kombination der Ergebnisse

Diese m Datensätze können nun jeweils einzeln mit den Standardmethoden ausgewertet werden. Deren Ergebnisse wie Punktschätzer und Standardfehler für Mittelwert oder

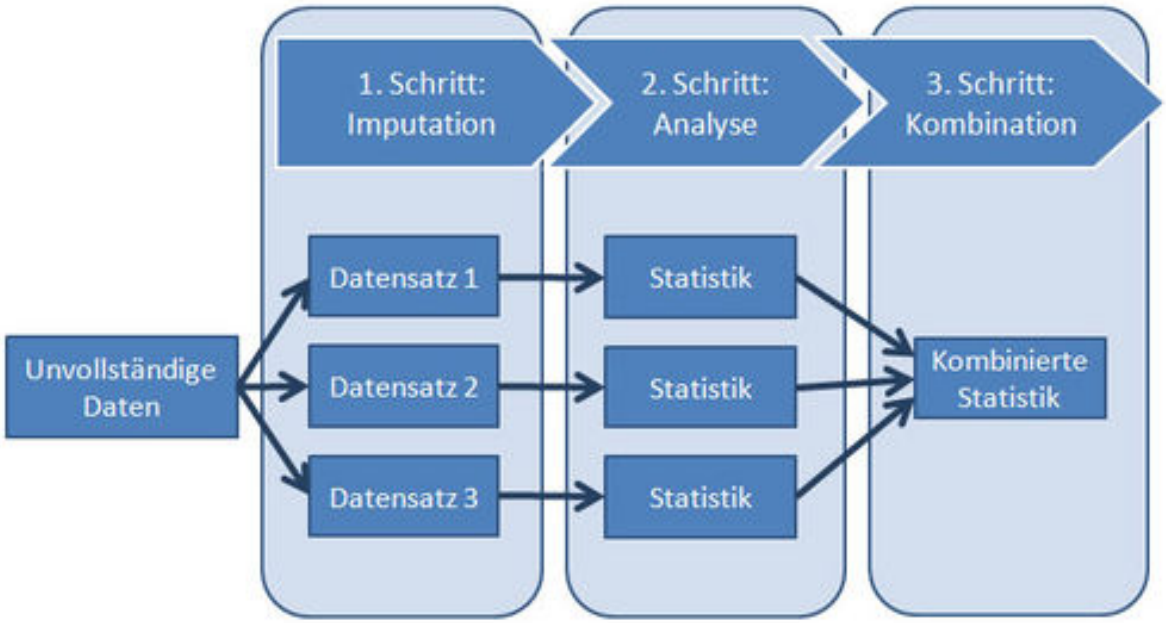

Abbildung 3.1: Grundkonzepte der Multiple-Imputation (z.B für $m = 3$)

Regressionskoeffizienten, sollen anschließend zu einem Einzelnen kombiniert werden. Die verschiedenen Schätzungen aus den m vervollständigten Datensätzen zu kombinieren wurde von Little u. Rubin (2002) folgende Formel veröffentlicht.
Sei Q der interessierende Parameter und V dessen zugehörigen Varianz. Somit erhält man aus der Analyse der m vervollständigten Datensätzen die Schätzer $\hat{Q}_1, \cdots, \hat{Q}_m$ und die entsprechenden geschätzten Varianzen $\hat{V}_1, \cdots, \hat{V}_m$ die alle gleich plausibel sind. Der kombinierte Schätzwert $\hat{Q}_{MI}$ kann berechnet werden durch

$$\hat{Q}_{MI} = \frac{1}{m} \sum_{i=1}^{m} \hat{Q}_i \tag{3.1}$$

Die zugehörige Gesamtvarianz $\hat{V}_{MI}$ ist dann

$$\hat{V}_{MI} = (1 + \frac{1}{m})\hat{B} + \hat{W} \tag{3.2}$$

mit der Within-Varianz (Varianz innerhalb der Imputation):

$$\hat{W} = \frac{1}{m} \sum_{i=1}^{m} \hat{V}_i \tag{3.3}$$

und der Between-Varianz (Varianz zwischen der m Imputation):

$$\hat{B} = \frac{1}{m-1}\sum_{i=1}^{m}(\hat{Q}_i - \hat{Q}_{MI})^2 \tag{3.4}$$

Dabei fließt in die geschätzte Gesamtvarianz $\hat{V}_{MI}$ durch den Faktor $(1+\frac{1}{m})$ die Between-Varianz in erhöhtem Maße ein und somit die Unsicherheit, die durch die m Imputationen entsteht, berücksichtigt wird.
Für große Stichproben sind Tests und Konfidenzintervalle approximativ t-verteilt (Little u. Rubin, 2002). Der Gesamtstandardfehler ist folglich die Quadratwurzel aus $\hat{V}_{MI}$. Konfidenzintervalle können damit berechnet werden als

$$KI = \hat{Q}_{MI} \pm t_{1-\frac{1}{\alpha}}\sqrt{\hat{V}_{MI}} \tag{3.5}$$

Mit den Freiheitsgraden der t-Verteilung

$$df = (m-1)(1 + \frac{1}{(m+1)}\frac{\hat{W}}{\hat{B}})^2 \tag{3.6}$$

wobei

$$\frac{1}{(m+1)}\frac{\hat{W}}{\hat{B}} \tag{3.7}$$

den relativen Anstieg der Varianz bedingt durch fehlende Werte beschreibt, der mit steigendem m sinkt.

3.3 Wahl der Anzahl m der Imputationen

Vor der Durchführung einer Multiple-Imputation ist es erforderlich festzulegen, wie viele Imputationen m ausgeführt werden sollen. Nach Rubin (1987) reichen üblicherweise schon kleine Anzahlen m zwischen 3 und 10 Wiederholungen aus, stabile Ergebnisse zu erhalten. Geht man von m Imputationen für einen fehlenden Wert aus, so zeigte Rubin (1987), dass sich die Relative Effizienz (RE) approximativ mit

$$RE = (1 + \frac{\gamma}{m})^{-\frac{1}{2}} \tag{3.8}$$

darstellen lässt. γ stellt den Anteil der fehlenden Werte im Datensatz.
Tabelle 3.3 zeigt die erzielten Relativen Effizienzen für unterschiedliche Anzahl an Imputationen m und verschiedenen Anteil an fehlenden Werten γ. In dieser Tabelle kann

		γ				
		0.1	0.3	0.5	0.7	0.9
	3	97	91	86	81	77
m	5	98	94	91	88	85
	10	99	97	95	93	92
	20	100	99	98	97	96

Tabelle 3.1: Effizienz (in %) von Multiple-Imputation (Enders, 2010)

man erkennen, dass eine geringe Anzahl an Imputationen in den meisten Fällen vollkommen ausreichend ist. Nur bei einem sehr hohen Fehlanteil lässt sich durch das Erhöhen dieser Anzahl ein entscheidender Effizienzgewinn erzielen.

In statistischen Programmpaketen wählt man meist als Standardwert $m = 5$.

3.4 Vor- und Nachteile

Der große Vorteil der Multiple-Imputation liegt zunächst in der einfachen Analyse und der Verwendung von sämtlicher zur Verfügung stehender Information. Die Daten können mit jeder, für die vollständigen Daten geeigneten Methode analysiert werden. Zusätzlich wird die Unsicherheit der imputierten Datensätze bei Multiple-Imputation widergespiegelt (Rubin, 1987). Eine hohe Effizienz kann mit wenigen Imputationen m erreicht werden (Tabelle 3.3).
Als Nachteil ist aufzuführen, dass Multiple-Imputation eine deutlich aufwendige Methode ist, die künstlich vervollständigten Datensätzen zu erzeugen. Zum Anderen benötigt man größere Rechen- und Speicherkapazität, weil sie ohnehin meist ohne Unterstützung von Computern nicht möglich sind. Außerdem ist jeder einzelne imputierte Datensatz separat zu berücksichtigen, was die Analysezeit erheblich steigert. Wie das oben Tabelle 3.3 zeigt, ist bei einem niedrigen Anteil fehlender Werte eine geringe Imputationsanzahl ausreichend und damit der Aufwand vertretbar. Wenn der Anteil fehlender Werte groß ist, muss auch Anzahl der Imputation gesteigert werden und der Aufwand steigt ebenfalls.

Trotz aller Vorteilen stellt die Anwendung von Multiple-Imptation Probleme dar. Ob die Analyse allerdings zu, im Sinne validen Aussagen führt, hängt von der gewählten Imputationsverfahren ab. In den folgenden Kapitel werden nun zwei Verfahren vorgestellt, die bei der Durchführung einer Multiple-Imputation dafür sorgen, dass die fehlenden Werte

eines Datensatzes ersetzt werden.

4 Imputationsverfahren

4.1 Notation

Zur besseren Übersicht und leichterem Verständnis soll die Notation, die im Folgenden verwendet wird, eingeführt und erläutert werden. Eine Datenmatrix Y mit n unabhängigen Beobachtungen und p Variablen liegt vor. Y_j repräsentiert eine der Variablen mit $Y_j = (Y_{1j}, \cdots, Y_{nj})$ mit $j = 1, \cdots, p$. Der Teil der beobachteten Werte sei mit Y_{beob}, der der fehlenden Werte mit Y_{fehl} bezeichnet $Y = (Y_{beob}, Y_{fehl})$ mit $Y_{beob} = (Y_{beob,1}, \cdots, Y_{beob,p})$ und $Y_{fehl} = (Y_{fehl,1}, \cdots, Y_{fehl,p})$. Weiter werden die fehlenden bzw. beobachteten Werte einer Variablen mit $Y_j = (Y_{beob,j}, Y_{fehl,j})$ bezeichnet. θ stellen die unbekannten Parameter des interessierenden Modells dar.
Um die Notation einfach zu halten, ist nun der Fehlendmechanismus hier ignorierbar sofern die Ausprägung von Y_{fehl} zufällig fehlen (MAR).

4.2 Likelihood-basierte Verfahren

Die gemeinsame Verteilung der beobachteten und fehlenden Daten wird dargestellt als

$$f(Y|\theta) = f(Y_{beob}, Y_{fehl}|\theta) \tag{4.1}$$

Falls die fehlenden Daten MAR sind, kann die Likelihood nach Little u. Rubin (2002) wie folgt bestimmt werden

$$L(\theta|Y_{beob}) = \int f(Y_{beob}, Y_{fehl}|\theta) dY_{fehl} \tag{4.2}$$

Wie bei normaler Maximum-Likelihood gilt es hier das Maximum dieser Funktion zu finden. Im Falle von fehlenden Daten ist dies aber kompliziert, da die benötigte Fisher-Informationsmatrix bzw. erwartete Fisher-Informationsmatrix aufwendig ist zu berechnen. Das Maximierungsproblem kann in diesen Fällen durch die Anwendung von dem

Expectation-Maximization Algorithmus gelöst werden (Dempster u. a., 1977).

4.2.1 Expectation-Maximization Algorithmus (EM Algorithmus)

Der Expectation Maximization Algorithmus ist ein relativ bekanntes Verfahren zur Bestimmung von Maximum-Likelihood-Schätzern unter unvollständigen Daten. Der geht auf Dempster u. a. (1977) zurück. Expectation Maximization wird im Folgenden als EM abgekürzt.
Die Grundidee des EM Algorithmus ist es, zunächst die fehlenden Werte durch Schätzungen zu ersetzen, damit eine Parameterschätzung durchzuführen, auf derer wiederum die fehlenden Werte neu geschätzt werden. Die fehlende Werte und die Parameter werden so lange neu geschätzt bis es zur Konvergenz kommt Little u. Rubin (2002).
Es besteht aus 2 Schritten, die iterativ wiederholt werden. Der E-Schritt (Expectation) bildet den bedingten Erwartungswert der fehlenden Werten gegeben auf gegebenen Ybeob und die aktuellen geschätzten Parameter. Im M-Schritt (Maximization) wird der erhaltene Erwartungswert unter den Parametern maximiert. Der EM-Algorithmus kann nun wie folgt beschrieben werden (Little u. Rubin, 2002):

E-Schritt: Berechne den bedingten Erwartungswert von Log-Likelihood gegeben den beobachteten Werten und dem jeweilig aktuellen $\theta^{(t)}$

$$Q(\theta|\theta^{(t)}) = E(l(\theta, Y|Y_{beob}, \theta^{(t)}) = \int l(\theta|Y)P(Y_{fehl}|Y_{beob}, \theta = \theta^{(t)})dY_{fehl} \tag{4.3}$$

wobei θ eine interessierende Größe ist.

M-Schritt: Finde dass $\theta^{(t+1)}$ um $Q(\theta|\theta^{(t)})$, welches im ersten Schritt bestimmt wurde, zu maximieren

$$Q(\theta^{(t+1)}|\theta^{(t)}) \geq Q(\theta|\theta^{(t)}) \tag{4.4}$$

für alle θ

Dieser iterative Prozess wird solange fortgeführt, bis die Parameterschätzer konvergieren, d.h. sie verändern sich nur minimal von Iteration zu $|\theta^{(t+1)} - \theta^{(t)}| \leq \epsilon$, mit ϵ einem beliebig klein wählbaren Wert. Für den ersten Iteration Schritt muss der Startwert $\theta^{(0)}$ bestimmt werden. Die Startwerte wie Mittelwerte und Kovarianzen erhält man mit Hilfe von Fallweisen oder Paarweisen Ausschluss.
Little u. Rubin (2002) zeigen, dass unter bestimmten Bedingungen wie einer linearen loglikelihood der EM Algorithmus zuverlässig konvergiert. Das heißt jede Iteration erhöht

die Log-Likelihood $lnL(Y_{obs}|\theta)$ deshalb EM-Algorithmus meist einfach zu konstruieren ist und sich jeder Schritt leicht interpretieren lässt.
Ein Nachteil stellt jedoch die unter Umständen langsame Konvergenz dar. In der Literatur finden sich verschiedene Ansätze zur Erhöhung der Geschwindigkeit des Algorithmus durch Verbindung mit anderen Algorithmen wie Newton-Raphson oder dem Scoring-Algorithmus (Little u. Rubin, 2002). Ein weiterer Nachteil dieser Vorgehensweise besteht darin, dass der Algorithmus lediglich die bedingten Erwartungswerte für die Imputation verwendet hat und somit die Varianz schwierig ist zu gewinnen. Somit wird die Unsicherheit der Schätzung nicht berücksichtigt. Als Verfahren, das direkt zur Multiple-Imputation fehlender Werte eingesetzt werden kann, ist EM Algorithmus nicht geeignet. Es ist dennoch sinnvoll aufgrund der wichtigen Bedeutung der Ergebnisse des EM Algorithmus im Hinblick auf die Verwendung der DA Algorithmus.

4.3 Bayes-basierte Verfahren

Im Rahmen des Bayes-Ansatzes sollte es sich um m unabhängige Zufallsziehungen für die fehlenden Daten Y_{fehl} aus ihren a-posteriori Prädiktivverteilung $f(Y_{fehl}|Y_{beob})$ der fehlenden Werte Y_{fehl} gegeben die beobachteten Werte Y_{beob} handeln (Schafer, 1997). Diese Verteilung lässt sich darstellen als

$$f(Y_{fehl}|Y_{beob}) = \int f(\theta, Y_{fehl}|Y_{beob})d\theta = \int f(Y_{fehl}|Y_{beob}, \theta)f(\theta|Y_{beob})d\theta \qquad (4.5)$$

Häufig wird zweistufige Ziehung verwendet. Durch diese Züge werden die Unsicherheiten in der Vorhersage der einzelnen fehlenden Werte gegebenen Parameter und die Unsicherheit über die Parameterschätzung widergespiegelt (Rässler u. a., 2013). Diese können beispielsweise direkt realisiert werden,

1. Ein Wert des Parameter θ wird aus seiner a-posteriori Verteilung von θ gegeben die beobachteten Daten $f(\theta|Y_{beob})$ zufällig gezogen.

2. Der fehlende Wert Y_{fehl} wird gemäß den bedingt Prädiktivverteilung $f(Y_{fehl}|Y_{beob}, \theta)$ für aktuellen Wert von θ erzeugt.

Das Problem bei diesem Vorgehen liegt üblicherweise in der Komplexität von $f(\theta|Y_{beob})$. Die a-posteriori Verteilungen $f(\theta|Y_{beob})$ sind häufig unhandlich und schwierig zu bestimmen. Daher werden zur Durchführung von Multiple-Imputation verstärkt Markov-

Chain-Monte-Carlo (MCMC) Methoden eingesetzt. Anschließend werden zwei MCMC-Techniken zur Beschaffung solcher Zufallsziehungen vorgestellt.

4.3.1 Data Augmentation Algorithmus (DA Algorithmus)

Der Data Augmentation Algorithmus, der auf den grundlegenden Artikel von Tanner u. Wong (1987) zurückgeht, übersetzt in etwa "Datenmehrung" bedeutet. Dies ist ein iteratives Verfahren zur Simulation der a-posteriori Verteilungen von θ. Dieser Algorithmus ist eine stochastische Bayesversion des EM Algorithmus. Data Augmentation wird im Folgenden als DA abgekürzt.
Beim DA Algorithmus werden die fehlenden Daten durch zufälliges Ziehen aus der bedingten Prädiktivverteilung unter Annahme eines Parameters, der aus der a-posteriori Verteilung in vorherige Iteration gezogen wird, ersetzt. Eine gemeinsame Verteilung aller Variable wird vorgenommen, z.B. multivariate Normalverteilung.
Analog zum EM Algorithmus gliedert der DA Algorithmus sich in zwei Schritte, die iterativ wiederholt werden: I-Schritt (Imputation) und P-Schritt (Posterior). Der DA Algorithmus kann nun wie folgt beschrieben werden (Little u. Rubin, 2002):

I-Schritt: Ziehe die fehlende Werte

$$Y_{fehl}^{(t+1)} \sim f(Y_{fehl}|Y_{beob}, \theta^{(t)}) \tag{4.6}$$

gemäß der sogenannt bedingt Prädiktivverteilung von Y_{fehl}, d.h. gegeben die beobachteten Werten und einen aktuellen Parameter von $\theta^{(t)}$.

P-Schritt: Gegeben die berechneten Werte $Y_{fehl}^{(t+1)}$, ziehe einen neuen Wert für θ aus seiner vollständigen a-posteriori Verteilung, d.h. gemäß

$$\theta^{(t+1)} \sim f(\theta|Y_{beob}, Y_{fehl}^{(t+1)}) \tag{4.7}$$

Ausgehen von einem Startwert $\theta^{(0)}$ erhält man eine Markov Kette $\left\{\left(\theta^{(t)}, Y_{fehl}^{(t)}\right) : t = 0, 1, \cdots\right\}$ deren stationäre Verteilung die gemeinsame Verteilung $f(\theta, Y_{fehl}|Y_{beob})$ ist. Für $t \to \infty$ können die Werte $\theta^{(t)}$ und $Y_{fehl}^{(t)}$ daher als approximative Ziehungen aus den stationären Verteilungen $f(\theta|Y_{beob})$ und $f(Y_{fehl}|Y_{beob})$ aufgefasst werden (Schafer, 1997). Beim DA Algorithmus werden von Enders (2010) ca. 200 Iterationen vor dem ersten Datensatz und zwischen zwei unabhängigen Datensätzen empfohlen. Zu Beginn des Verfahrens muss für den Verteilungsparameter θ ein Startwert $\theta^{(0)}$ festgelegt werden, z.B. für multi-

variate Normalverteilung sind μ und Σ ausreichend. Die Ergebnisse des EM Algorithmus werden hier häufig verwendet.
Der DA Algorithmus ist einfach zu implementieren. Wenn die zugrunde gemeinsame Verteilung des Datensatzes richtig spezifiziert wird, könnte DA Algorithmus gültige Ergebnisse mit dem imputierten Datensatz gewährleisten. Aber die einzelnen Variablen sind in den empirischen Datensätzen oft unterschiedlichen Typs, z.B. Mischung aus numerischen und kategorialen Variablen, stellt sich das Problem, ein multivariates Verteilung anzupassen (Spieß, 2008). Der Grund dafür liegt darin, dass es schwierig sein kann eine gemeinsame Verteilung aller Variablen zu finden. Für solchen Datensätzen kann man den DA Algorithmus nicht anwenden.

4.3.2 Multivariate Imputation by Chained Equation Algorithmus (MICE Algorithmus)

Der Multivariate Imputation by Chained Equation Algorithmus, der von van Buuren (2007) vorgestellt wird, ersetzt die Multiple-Imputation durch iterative univariate Imputation. Es wird im Folgenden als MICE abgekürzt. Dieses Verfahren ist in der englischen Fachsprache auch unter anderen Namen bekannt wie Stochastic Relaxation, Regression Switching, Sequential Regression, Incompatible MCMC, usw. (van Buuren u. Groothuis-Oudshoorn, 2011).
Die Grundidee dieses Algorithmus besteht darin, die Verteilung mit Hilfe von Ziehungen aus den bedingten Verteilungen zu schätzen. Für jede Variable mit fehlenden Werten wird eine separate bedingte Verteilung spezifiziert (van Buuren u. Groothuis-Oudshoorn, 2011). Ausgehende von einer anfänglichen Imputation, erzeugt MICE die Imputationen durch Iteration über die bedingte Verteilung.
Beim MICE Algorithmus werden am Anfang die fehlenden Werte durch Hot-deck Imputation ersetzt. Dann startet ein iteratives Verfahren (Gibbs Sampling), in dem bei in jedem Schritt die Variablen sukzessive behandelt werden (van Buuren u. Groothuis-Oudshoorn, 2011). Die Iteration t von Gibbs Sampler Algorithmus für Generierung $Y_{fehl}^{(t)}$ aus $Y_{fehl}^{(t-1)}$ ist dann gegeben durch

$$\theta_1^{(t)} \sim f(\theta_1 | Y_{beob,1}, Y_2^{(t-1)}, \cdots, Y_p^{(t-1)})$$
$$Y_{fehl,1}^{(t)} \sim f(Y_{fehl,1} | Y_{beob,1}, Y_2^{(t-1)}, \cdots, Y_p^{(t-1)}, \theta_1^{(t)})$$

$$\vdots$$

$$\theta_p^{(t)} \sim f(\theta_p | Y_{beob,p}, Y_1^{(t)}, Y_2^{(t)}, \cdots, Y_{p-1}^{(t)})$$
$$Y_{fehl,p}^{(t)} \sim f(Y_{fehl,p} | Y_{beob,p}, Y_1^{(t)}, \cdots, Y_p^{(t)}, \theta_1^{(t)}, \cdots \theta_p^{(t)})$$

wobei $Y_j^{(t)} = (Y_{beob,j}^{(t)}, Y_{fehl,j}^{(t)})$ die j-te vervollständigte Variable bei Iteration t ist. Diese Sequenz erzeugt wieder eine Mavkow-Kette, die unter geeigneten der Ziehung aus der gemeinsamen Verteilung von Y_{fehl} und θ gegeben Y_{beob} entspricht. Die Konvergenz ist daher recht schnell, also wird der Algorithmus bis zu einer vorgegebenen Anzahl von Iteration im Bereich von 10–20 weiterdurchgeführt (van Buuren u. Groothuis-Oudshoorn, 2011).

Der MICE Algorithmus bietet den Vorteil, dass man hier die Annahme der gemeinsamen Verteilung aller Variablen vermeiden kann. Zum Anderen können mit komplexe Datenstrukturen berücksichtigt werden (van Buuren u. Groothuis-Oudshoorn, 2011). Man unterscheidet dabei vier Typen von Variablen: Zählvariablen, stetige, kategoriale und gemischte (0 als kategorialer Wert, sonst stetig) Variable. Zählvariablen werden mit einer Poisson Regression imputiert. Für stetige Variable wird das normale lineare Regressionsmodell zur Schätzung verwendet und für kategoriale ein logistisches oder verallgemeinertes logistisches Modell. Gemischte Variablen werden zweistufig imputiert.

5 Simulation

Im Rahmen einer Simulationsstudie solle die einige der in vorangegangen Abschnitt Methoden der Multiple-Imputation auf unvollständige Datensätze angewendet werden. Die drei Variablen X_1, X_2 und Y mit jeweils 10000 Beobachtungen werden generiert. X_1 und X_2 werden aus einer bivariaten Standardnormalverteilung mit einer Korrelation von ρ gezogen. Zudem sei $Y = \beta_1 X_1 + \beta_2 X_2 + \epsilon$ mit $\epsilon \sim N(0,1)$ und $\beta_1 = \beta_2 = 1$.
Die Verteilung von $R|X_1, X_2, Y$ sei von X_1, X_2 und Y abhängig und durch ein logistisches Modell

$$P(R = 1|X_1, X_2, Y, \alpha) = \frac{exp(\alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 Y)}{1 + exp(\alpha_0 + \alpha_0 X_1 + \alpha_2 X_2 + \alpha_3 Y} \tag{5.1}$$

Die Parameter α_1, α_2 und α_3 legen fest, ob die fehlenden Daten MCAR, MAR oder MNAR sind:

- MCAR: $\alpha_0 = 2.5$ und $\alpha_1 = \alpha_2 = \alpha_3 = 0$
- MAR abhängig von X_1: $\alpha_0 = 2.5$, $\alpha_1 = -1$ und $\alpha_2 = \alpha_3 = 0$
- MAR abhängig von Y: $\alpha_0 = 2.5$, $\alpha_3 = -1$ und $\alpha_1 = \alpha_2 = 0$
- MNAR: $\alpha_2 = -4$ und $\alpha_0 = \alpha_1 = \alpha_3 = 0$

Um erkennen zu können, wie gut die Multiple-Imputation durch die einzelnen Algorithmen ist, werden verschiedene Größen in Abhängigkeit von der Korrelation ρ zwischen X_1 und X_2. Insgesamt werden bei den Datensätzen 50% der Beobachtungen entfernt. Für die unterschiedlichen Datensituationen wird nun jeweils Regressionskoeffizienten β_2 und seine Standardfehler, der Mittelwert und die Varianz von X_2 sowie der Korrelationskoeffizient zwischen X_1 und X_2 bestimmt, die Qualität der Verfahren zu prüfen

5.1 Verwendung des Data-Augmentation Algorithmus

Zuerst wird die Multiple-Imputation auf Basis eines DA Algorithmus durchgeführt. Die Multiple-Imputation wird mithilfe des *"norm"* packages aus dem Programm R durch-

geführt (Novo, 2013). Der Paket muss für die Multiple-Imputation lediglich der unvollständige Datensatz, die Anzahl an Imputationen m, die durchgeführt werden sollen. Die Anzahl der Datensätze, welche die Multiple-Imputation erzeugen soll, wird auf $3, 5$ und 10 festgelegt. Die jeweiligen unvollständigen Datensätze werden dem DA Algorithmus mit 20 Iterationen unterworfen und pro Fehlendmechanismus ein vollständiger Datensatz erstellt.
Für die Berechnung des jeweiligen Multiple-Imputation Schätzer, ist es somit notwendig, die einzelnen Schätzer der m vervollständigten Datensätze gemäß Abschnitt 3.2 zu kombinieren. Die Tabelle 5.1 bis 5.4 zeigen die Ergebnisse der Multiple-Imputation unter Verwendung des DA Algorithmus.
In Tabelle 5.1 sind die Ergebnisse bei einem MCAR Mechanismus. Für alle Fälle treffen die Parameterschätzer β_2 den wahren Wert von 1 sehr gut. Dies gilt vor allem unabhängig von der Korrelation zwischen X_1 und X_2. Lediglich steigen die Standardfehler mit zunehmender Korrelation leicht an. Über alle m Imputationen eines unvollständigen Datensatzes beschreiben die Mittelwerte μ_{X_2} bzw. Varianzen trotz leichter Überschätzung sehr gute Ergebnisse. Auch die Korrelationskoeffizienten bleiben erhalten. Durch das Vorliegen eines MCAR Mechanismus konnten durch den DA Algorithmus, sehr gute imputierte Datensätze erzeugt werden.

m	ρ	β_2	Std.fehler	μ_{X_2}	$\sigma^2_{X_2}$	$\rho_{X_1X_2}$
3	0.0	0.99973	0.01032	0.02265	1.01802	-0.01992
	0.3	1.00936	0.01082	0.01352	1.01301	0.30561
	0.5	1.00438	0.01231	0.02131	1.02384	0.50439
	0.8	1.01636	0.01924	0.00963	1.01099	0.80119
5	0.0	1.00082	0.01096	0.02203	1.01845	-0.01706
	0.3	1.01044	0.01149	0.01291	1.01144	0.30602
	0.5	1.00532	0.01249	0.02074	1.02444	0.50694
	0.8	1.01789	0.01853	0.00917	1.00996	0.80132
10	0.0	0.99883	0.01081	0.01907	1.01945	-0.01744
	0.3	1.00271	0.01131	0.01906	1.02551	0.30783
	0.5	1.00299	0.01249	0.01801	1.02514	0.50604
	0.8	1.00446	0.01866	0.01338	1.02176	0.80245

Tabelle 5.1: Ergebnisse der m imputierten Datensätze bei MCAR und DA Algorithmus, $\gamma = 50\%$

Ein ganz ähnliches Bild ergibt sich bei Vorliegen eines MAR Mechanismus. Die Ergebnisse für den MAR Mechanismus, der von X_1 bzwischen von Y abhängig ist, sind in Tabelle 5.2 bzw. Tabelle 5.3 zusammengefasst. Alle Werte zeigen keine Auffälligkeiten.

Die Parameterschätzer sind unverzerrt mit niedrigen Standardfehlern. Auch die Struktur der Daten bleibt nach der Imputation erhalten, was sich in plausiblen Werten für $\mu_{X_2}, \sigma^2_{X_2}, \rho_{X_1X_2}$ widerspiegelt.

m	ρ	β_2	Std.fehler	μ_{X_2}	$\sigma^2_{X_2}$	$\rho_{X_1X_2}$
3	0.0	0.99242	0.0103	0.01331	1.02807	-0.02616
	0.3	0.99624	0.01139	0.00457	1.00477	0.30227
	0.5	1.01285	0.01225	0.00559	0.99917	0.49847
	0.8	1.01272	0.01681	0.00767	1.01684	0.80017
5	0.0	0.99346	0.01096	0.01269	1.02504	-0.02546
	0.3	0.99749	0.01178	0.00395	1.00484	0.30222
	0.5	1.01395	0.01299	0.00503	0.99951	0.50025
	0.8	1.01624	0.02029	0.00721	1.02098	0.80091
10	0.0	0.99151	0.01077	0.0097	1.03075	-0.02422
	0.3	0.98987	0.01147	0.01055	1.01709	0.30174
	0.5	1.0121	0.01252	0.00232	1.00452	0.50025
	0.8	0.99738	0.01893	0.01162	1.02856	0.80247

Tabelle 5.2: Ergebnisse der m imputierten Datensätze bei MCAR abhängig von X_2 und DA Algorithmus, $\gamma = 50\%$

m	ρ	β_2	Std.fehler	μ_{X_2}	$\sigma^2_{X_2}$	$\rho_{X_1X_2}$
3	0.0	1.00543	0.01026	0.00483	1.01809	-0.01384
	0.3	1.01845	0.0109	0.0019	0.99325	0.29959
	0.5	1.00766	0.01177	-0.00355	1.02283	0.50312
	0.8	1.02619	0.01677	-0.00942	1.0297	0.80062
5	0.0	1.00653	0.01093	0.00422	1.01499	-0.01159
	0.3	1.01965	0.01154	0.00129	0.99414	0.29818
	0.5	1.00921	0.01219	-0.00412	1.02117	0.50109
	0.8	1.02786	0.01726	-0.00988	1.02839	0.80066
10	0.0	1.00457	0.01079	0.00128	1.01433	-0.01262
	0.3	1.01204	0.01126	0.00808	1.00593	0.29223
	0.5	1.00647	0.01249	-0.00684	1.01932	0.50034
	0.8	1.00627	0.01945	-0.00412	1.0262	0.80091

Tabelle 5.3: Ergebnisse der m imputierten Datensätze bei MCAR abhängig von Y und DA Algorithmus, $\gamma = 50\%$

Die Probleme werden bei MNAR Mechanismus sichtbar (Tabelle 5.4). Hier erzielen DA keine sehr guten Ergebnisse. Die Parameterschätzer sind verzerrt. Die Standardfehler der Schätzungen sind im Vergleich zu den bereits erläuterten Ergebnissen bei anderen Fehlendmechanismen deutlich erhöht. Die Mittelwerte der imputierten Variablen X_2

werden nicht korrekt wiedergegeben und auch die Varianzen weichen stark vom wahren Wert von 1 ab. Die Korrelation zwischen X_1 und X_2 kann nicht gut nachgebildet werden.

m	ρ	β_2	Std.fehler	μ_{X_2}	$\sigma^2_{X_2}$	$\rho_{X_1X_2}$
3	0.0	1.22161	0.01582	0.501	0.5334	-0.0029
	0.3	1.23203	0.01669	0.477	0.53884	0.26469
	0.5	1.17577	0.01765	0.46201	0.53767	0.45136
	0.8	1.07774	0.02176	0.31659	0.63928	0.76702
5	0.0	1.22311	0.01704	0.5005	0.53258	-0.00077
	0.3	1.23278	0.01773	0.4765	0.53657	0.26402
	0.5	1.17907	0.01833	0.46155	0.54006	0.45395
	0.8	1.07857	0.02229	0.31619	0.63857	0.76737
10	0.0	1.22126	0.0163	0.49814	0.5357	-0.00028
	0.3	1.21029	0.01689	0.48885	0.52588	0.26295
	0.5	1.17649	0.01823	0.45933	0.54183	0.45371
	0.8	1.05278	0.0228	0.32534	0.62895	0.76312

Tabelle 5.4: Ergebnisse der m imputierten Datensätze bei MNAR und DA Algorithmus, $\gamma = 50\%$

Multiple-Imputation unter Anwendung des DA Algorithmus ist nach den vorliegenden Ergebnissen eine adäquate Methode die unvollständige Datensätze zu vervollständigen. Zumindest gilt dies, wenn ein MCAR oder MAR Mechanismus vorliegt. In diesen Fällen führt die Multiple-Imputation zu unverzerrten Schätzern, basierend auf den imputierten Datensätzen. Zudem bleiben Mittelwerte und Varianzen der vervollständigten Variablen erhalten und auch die Abhängigkeit zur Variablen X_1 wird durch die Multiple-Imputation beibehalten. Aber dieses Verfahren ist nicht in der Lage mit einem MNAR Mechanismus umzugehen. Alle Größen weichen stark von ihren wahren Werten ab.

5.2 Verwendung des MICE Algorithmus

Dieselbe Struktur der Simulationsstudie wie im vorangegangenen Abschnitt liegt auch bei der Bewertung der Imputation durch MICE Algorithmus, die mit Package *"mice"* in der Software R durchgeführt wird (van Buuren u. Groothuis-Oudshoorn, 2011). Für alle unvollständigen Datensätze werden im Rahmen einer jeden Imputation $m = 3, 5, 10$ vervollständige Datensätze erzeugt. Für die Variable X_2 werden Imputationen auf der Basis eines linearen Modells erzeugt. Die Tabelle 5.5 bis 5.8 zeigen die Ergebnisse der Multiple-Imputation unter Verwendung des MICE Algorithmus.
In Tabelle 5.5 sind die Ergebnisse bei einem MCAR Mechanismus. Die Parameterschätzer

sind nahezu unverzerrt bei einem geringen Standardfehler. Betrachtet man die verbleibender drei Größen, so ergibt sich, dass der MICE Algorithmus diese Größen gut nachgebildet. Sowohl Mittelwert als auch Varianzen der Variable X_2 liegen beim MICE Algorithmus näher an ihrem wahren Wert von 0 bzw. 1. Auch die Korrelation zwischen X_1 und X_2 wird durch dieses Vorgehen gut widergespiegelt.

m	ρ	β_2	Std.fehler	μ_{X_2}	$\sigma^2_{X_2}$	$\rho_{X_1X_2}$
3	0.0	1.00005	0.01607	0.00952	1.03105	-0.00656
	0.3	1.00433	0.01645	0.00951	1.03685	0.31614
	0.5	1.00546	0.01723	0.00865	1.03856	0.51199
	0.8	1.01193	0.02041	0.00531	1.03405	0.80399
5	0.0	1.00656	0.0106	0.01473	1.00305	-0.01575
	0.3	1.01046	0.01109	0.01468	1.00972	0.30737
	0.5	1.01084	0.01225	0.01356	1.01039	0.50617
	0.8	1.01334	0.01822	0.00916	1.01006	0.80292
10	0.0	1.0065	0.01134	0.01324	1.0078	-0.01500
	0.3	1.01052	0.01175	0.01319	1.01521	0.30839
	0.5	1.01137	0.01273	0.01215	1.01581	0.50691
	0.8	1.01588	0.01873	0.00805	1.01451	0.80318

Tabelle 5.5: Ergebnisse der m imputierten Datensätze bei MCAR und MICE Algorithmus, $\gamma = 50\%$

Die Ergebnisse aus Tabelle 5.6 und 5.7 zeigen auch, dass MICE-Algorithmus in der Lage für ein MAR Mechanismus ist sinnvolle Werte zu ersetzen. Die imputierten Datensätze sind unverzerrte Schätzer bei niedrigen Standardfehlern zu erzeugen. Die Mittelwerte, Varianzen und die Korrelationen zwischen X_1 und X_2 liegen immer noch nahe den wahren Werten.

Bei Vorhandensein eines MNAR Mechanismus (Tabelle 5.8) treten Problem auf. Bei erhöhten Standardfehlern sind die Parameterschätzer stark verzerrt. Die Mittelwerte und Varianzen werden nicht korrekt wiedergegeben. Die Korrelationskoeffizienten zwischen X_1 und X_2 können zwar nicht so exakt wie bei dem MCAR und MAR Mechanismus nachgebildet.

Abschließend kann festgehalten werden, dass MICE Algorithmus in der Lage mit einem MCAR oder MAR Fehlendmechanismus umzugehen. Dies gilt aber nicht bei Vorliegen eines MNAR Mechanismus.

m	ρ	β_2	Std.fehler	μ_{X_2}	$\sigma^2_{X_2}$	$\rho_{X_1X_2}$
3	0.0	0.99282	0.01602	-0.00049	1.03445	-0.02376
	0.3	0.99194	0.01622	0.0002	1.0317	0.30912
	0.5	1.01283	0.01581	-0.00763	1.00641	0.50436
	0.8	1.00348	0.01732	0.00288	1.03785	0.80356
3	0.0	0.99916	0.01054	0.00523	1.01674	-0.0275
	0.3	0.99744	0.01136	0.00594	1.00106	0.30234
	0.5	1.01956	0.01227	-0.00225	0.99013	0.50213
	0.8	1.00737	0.01836	0.00735	1.01974	0.80337
10	0.0	0.99913	0.01126	0.00359	1.02295	-0.02323
	0.3	0.99821	0.01193	0.00429	1.01316	0.30219
	0.5	1.01993	0.01279	-0.00379	0.9944	0.50332
	0.8	1.01323	0.01967	0.00606	1.02375	0.80265

Tabelle 5.6: Ergebnisse der m imputierten Datensätze bei MAR abhängig von X_1 und MICE Algorithmus, $\gamma = 50\%$

m	ρ	β_2	Std.fehler	μ_{X_2}	$\sigma^2_{X_2}$	$\rho_{X_1X_2}$
3	0.0	1.00559	0.01613	-0.01003	1.02439	0.00099
	0.3	1.0133	0.01631	-0.00323	1.01327	0.30147
	0.5	1.00786	0.01756	-0.01864	1.02213	0.50453
	0.8	1.01385	0.01862	-0.01409	1.03632	0.80197
5	0.0	1.0122	0.01052	-0.00305	1.00584	-0.01337
	0.3	1.01984	0.01111	0.00394	0.9948	0.29677
	0.5	1.01426	0.01227	-0.0115	1.01193	0.50594
	0.8	1.02128	0.01718	-0.00834	1.01921	0.80093
10	0.0	1.01212	0.01134	-0.00506	1.00292	-0.00903
	0.3	1.01986	0.01186	0.00189	0.99604	0.29573
	0.5	1.01547	0.01258	-0.01354	1.0186	0.50485
	0.8	1.02447	0.01818	-0.00997	1.02459	0.8011

Tabelle 5.7: Ergebnisse der m imputierten Datensätze bei MAR abhängig von Y und MICE Algorithmus, $\gamma = 50\%$

m	ρ	β_2	Std.fehler	μ_{X_2}	$\sigma^2_{X_2}$	$\rho_{X_1X_2}$
3	0.0	1.23038	0.02176	0.48444	0.54726	0.00339
	0.3	1.22263	0.02128	0.47296	0.54069	0.26858
	0.5	1.19039	0.02066	0.44376	0.55478	0.45707
	0.8	1.07414	0.02104	0.30738	0.65392	0.77031
5	0.0	1.23389	0.01661	0.49329	0.5333	-0.00047
	0.3	1.2265	0.01772	0.48232	0.52826	0.26779
	0.5	1.19021	0.01732	0.45311	0.541	0.45909
	0.8	1.0741	0.0242	0.31603	0.63772	0.76845
10	0.0	1.23553	0.01604	0.4912	0.53626	0.00288
	0.3	1.22801	0.0164	0.48011	0.52955	0.26484
	0.5	1.19294	0.01716	0.45096	0.54373	0.45809
	0.8	1.07422	0.02322	0.31425	0.63929	0.76889

Tabelle 5.8: Ergebnisse der m imputierten Datensätze bei MNAR und MICE Algorithmus, $\gamma = 50\%$

5.3 Simulation bei unterschiedlichen Anteil fehlender Werten

Die Analyse in den vorangegangenen Abschnitt bezogen sich jeweils auf einen Anteil an fehlenden Werten je Datensatz von 50%. Im Folgenden wird auf die Ergebnisse eingegangen, die sich bei unterschiedlichen Anteil fehlender Werten und verschiedenen Korrelationen ρ zwischen X_1 und X_2 eingestellt haben.
Da wie im vorherigen Abschnitt beschrieben wurde, dass es nicht möglich ist mit einem MNAR Mechanismus umzugehen, sollten hier die fehlenden Daten nicht mit dem NMAR Mechanismus erzeugt werden.
Die Ergebnisse für die Multiple-Imputation durch DA Algorithmus sind in den Tabellen 5.9 bis 5.11 enthalten. Die Ergebnisse des MICE Algorithmus können aus den Tabellen 5.12 bis 5.14 abgelesen werden. Bezüglich der beiden ergeben sich beim unterschiedlichen Fehlanteil kaum Unterschiede. Die Parameterschätzer sind unverzerrt mit niedrigen Standardfehlern. Die Mittelwerte, Varianzen und Korrelationskoeffizienten zwischen X_1 und X_2 werden gut nachgebildet.

5.4 Fazit

Bezogen auf die Multiple-Imputation kann abschließend festgehalten werden, dass bei Multiple-Imputation durch die Anwendung von DA bzw. MICE Algorithmus sinnvoll

γ	ρ	β_2	Std.fehler	μ_{X_2}	$\sigma^2_{X_2}$	$\rho_{X_1X_2}$
10	0.0	1.00840	0.01028	-0.00004	1.01079	-0.01705
	0.3	1.00893	0.01072	0.00203	1.02101	0.30688
	0.5	1.01007	0.01179	0.00102	1.02016	0.50496
	0.8	1.01329	0.01712	0.00154	1.01744	0.80235
30	0.0	1.00583	0.01078	0.00870	1.00035	-0.02030
	0.3	1.00508	0.01121	0.01597	1.01566	0.30094
	0.5	1.00459	0.01219	0.00775	1.01690	0.50557
	0.8	1.00660	0.01918	0.01011	1.01175	0.80139
50	0.0	1.00082	0.01096	0.02203	1.01845	-0.01706
	0.3	1.01044	0.01149	0.01291	1.01144	0.30602
	0.5	1.00532	0.01249	0.02074	1.02444	0.50694
	0.8	1.01789	0.01853	0.00917	1.00996	0.80132
80	0.0	0.99931	0.01333	0.01700	1.04100	0.00306
	0.3	0.98841	0.01435	0.02617	1.03009	0.27926
	0.5	0.99902	0.01495	0.02866	1.06605	0.49895
	0.8	0.97934	0.02568	0.02122	1.00226	0.78766

Tabelle 5.9: Ergebnisse der $m = 5$ imputierten Datensätze bei MCAR und DA Algorithmus

γ	ρ	β_2	Std.fehler	μ_{X_2}	$\sigma^2_{X_2}$	$\rho_{X_1X_2}$
10	0.0	1.00879	0.01024	-0.00227	1.01227	-0.01858
	0.3	1.01367	0.01076	-0.00044	1.01179	0.30575
	0.5	1.01124	0.01182	-0.00369	1.01372	0.50641
	0.8	1.00894	0.01712	-0.00125	1.01275	0.80165
30	0.0	0.99813	0.01057	0.00260	1.02778	-0.01863
	0.3	0.99990	0.01110	0.00076	1.02068	0.30075
	0.5	1.01036	0.01244	-0.00452	1.00315	0.50669
	0.8	1.00381	0.01793	0.00470	1.01059	0.80178
50	0.0	0.99346	0.01096	0.01269	1.02504	-0.02546
	0.3	0.99749	0.01178	0.00395	1.00484	0.30222
	0.5	1.01395	0.01299	0.00503	0.99951	0.50025
	0.8	1.01624	0.02029	0.00721	1.02098	0.80091
80	0.0	0.99604	0.01355	0.01708	1.02519	-0.03123
	0.3	1.01337	0.01481	0.00046	0.98514	0.30725
	0.5	1.01014	0.01543	0.02057	1.01151	0.49860
	0.8	0.97717	0.02090	0.02968	1.00453	0.79133

Tabelle 5.10: Ergebnisse der $m = 5$ imputierten Datensätze bei MAR abhängig von X_1 und DA Algorithmus

γ	ρ	β_2	Std.fehler	μ_{X_2}	$\sigma^2_{X_2}$	$\rho_{X_1X_2}$
10	0.0	1.01575	0.01024	-0.00306	1.00638	-0.01252
	0.3	1.01327	0.01081	-0.00214	1.00972	0.30742
	0.5	1.01407	0.01188	-0.00383	1.01320	0.50594
	0.8	1.01135	0.01714	0.00072	1.01488	0.80224
30	0.0	1.01589	0.01056	0.00123	1.00912	-0.01556
	0.3	1.01538	0.01124	0.00746	1.00130	0.30720
	0.5	1.00990	0.01213	-0.01062	1.01537	0.50232
	0.8	1.01410	0.01779	0.00391	1.01537	0.80280
50	0.0	1.00653	0.01093	0.00422	1.01499	-0.01159
	0.3	1.01965	0.01154	0.00129	0.99414	0.29818
	0.5	1.00921	0.01219	-0.00412	1.02117	0.50109
	0.8	1.02786	0.01726	-0.00988	1.02839	0.80066
80	0.0	0.99126	0.01432	0.02926	0.98782	0.02229
	0.3	1.01174	0.01458	0.03804	0.97941	0.26668
	0.5	1.02825	0.01609	0.01127	0.98632	0.49323
	0.8	1.01343	0.02323	0.01552	0.98255	0.78619

Tabelle 5.11: Ergebnisse der $m = 5$ imputierten Datensätze bei MAR abhängig von Y und DA Algorithmus

γ	ρ	β_2	Std.fehler	μ_{X_2}	$\sigma^2_{X_2}$	$\rho_{X_1X_2}$
10	0.0	1.00817	0.01014	0.00147	1.00819	-0.01747
	0.3	1.00902	0.01060	0.00237	1.01870	0.30780
	0.5	1.00947	0.01171	0.00237	1.01903	0.50656
	0.8	1.01142	0.01699	0.00161	1.01726	0.80325
30	0.0	1.00660	0.01031	0.00840	0.99811	-0.01657
	0.3	1.00560	0.01086	0.00821	1.01465	0.30765
	0.5	1.00527	0.01239	0.00752	1.01582	0.50687
	0.8	1.00436	0.02101	0.00472	1.01625	0.80365
50	0.0	1.00656	0.01060	0.01473	1.00305	-0.01575
	0.3	1.01046	0.01109	0.01468	1.00972	0.30737
	0.5	1.01084	0.01225	0.01356	1.01039	0.50617
	0.8	1.01334	0.01822	0.00916	1.01006	0.80292
80	0.0	0.99637	0.02115	-0.00799	1.03322	-0.01839
	0.3	0.99441	0.02152	0.00499	1.05222	0.28881
	0.5	0.99469	0.02315	0.00618	1.04622	0.48698
	0.8	1.00297	0.03253	0.00649	1.02661	0.78994

Tabelle 5.12: Ergebnisse der $m = 5$ imputierten Datensätze bei MCAR und MICE Algorithmus

γ	ρ	β_2	Std.fehler	μ_{X_2}	$\sigma^2_{X_2}$	$\rho_{X_1X_2}$
10	0.0	1.00860	0.01014	-0.00074	1.00825	-0.02005
	0.3	1.01391	0.01060	-0.00020	1.01273	0.30661
	0.5	1.01111	0.01172	-0.00247	1.01414	0.50640
	0.8	1.00845	0.01737	-0.00124	1.01217	0.80210
30	0.0	0.99884	0.01006	0.00199	1.02922	-0.01902
	0.3	1.00025	0.01073	-0.00744	1.01687	0.30621
	0.5	1.01326	0.01179	-0.00497	1.01298	0.50683
	0.8	1.00248	0.01727	-0.00106	1.01543	0.80385
50	0.0	0.99916	0.01054	0.00523	1.01674	-0.02750
	0.3	0.99744	0.01136	0.00594	1.00106	0.30234
	0.5	1.01956	0.01227	-0.00225	0.99013	0.50213
	0.8	1.00737	0.01836	0.00735	1.01974	0.80337
80	0.0	0.99243	0.02141	-0.00533	1.00997	-0.03346
	0.3	1.01859	0.02222	-0.02499	0.97959	0.30859
	0.5	1.00995	0.02204	0.00286	1.03085	0.48970
	0.8	0.98973	0.03464	0.00997	1.00431	0.79141

Tabelle 5.13: Ergebnisse der $m = 5$ imputierten Datensätze bei MAR abhängig von X_1 und MICE Algorithmus

γ	ρ	β_2	Std.fehler	μ_{X_2}	$\sigma^2_{X_2}$	$\rho_{X_1X_2}$
10	0.0	1.01555	0.01010	-0.00149	1.00402	-0.01406
	0.3	1.01338	0.01065	-0.00173	1.00794	0.30850
	0.5	1.01332	0.01175	-0.00234	1.00915	0.50615
	0.8	1.01033	0.01689	0.00080	1.01273	0.80247
30	0.0	1.01677	0.01011	0.00046	1.00371	-0.01332
	0.3	1.01606	0.01076	-0.00088	1.00517	0.31104
	0.5	1.01018	0.01189	-0.01160	1.01383	0.50559
	0.8	1.01431	0.02043	-0.00206	1.02180	0.80471
50	0.0	1.01220	0.01052	-0.00305	1.00584	-0.01337
	0.3	1.01984	0.01111	0.00394	0.99480	0.29677
	0.5	.01426	0.01227	-0.01150	1.01193	0.50594
	0.8	1.02128	0.01718	-0.00834	1.01921	0.80093
50	0.0	0.98863	0.02144	0.00754	1.01981	-0.01180
	0.3	1.01703	0.02191	0.01403	0.98891	0.27335
	0.5	1.02533	0.02298	-0.00373	0.99795	0.48987
	0.8	1.02952	0.03793	-0.00244	0.98944	0.78616

Tabelle 5.14: Ergebnisse der $m = 5$ imputierten Datensätze bei MAR abhängig von Y und MICE Algorithmus

imputierte Datensätze erzeugt werden. Unproblematisch sind dabei MCAR und MAR Mechanismen, weil sich nach der Multiple-Imputation unverzerrte Parameterschätzer ergeben. Auch die Mittelwerte, Varianzen sowie die Abhängigkeitsstruktur werden durch die Imputation erhalten. Bei diesen Verfahren ist es aber nicht möglich mit einem MNAR Mechanismus umzugehen.

6 Diskussion

In dieser Arbeit wurde ein Überblick zu Multiple-Imputation und der Imputationsverfahren dargestellt. Multiple Imputation erweisen sich eindeutig als eine gute Methode für den Umgang mit fehlenden Werten. Der große Vorteil der Multiple-Imputation liegt zunächst in der einfachen Analyse und der Verwendung von sämtlicher zur Verfügung stehender Information. Die Daten können mit jeder, für die vollständigen Daten geeigneten Methode analysiert werden. Bei der Multiple-Imputation wird die Unsicherheit der unbekannten fehlenden Werte berücksichtigt. Der einzige Nachteil, den die Multiple-Imputation aufweist, ist der größere Aufwand um die Imputation und die Analyse durchzuführen.
Dazu werden Data Augmentation und MICE Algorithmus als Imputationsverfahren vorgestellt, die bei der Durchführung einer Multiple-Imputation dafür sorgen, dass die fehlenden Werte eines Datensatzes ersetzt werden. Die Multiple-Imputation mit MICE bzw. DA Algorithmus haben als wichtige Gemeinsamkeiten. Beide Verfahren verwenden einen iterativen Algorithmus, der als MCMC-Methode beschrieben werden kann. Auch für jeden vervollständigten Datensatz wird eine unabhängige Ziehung von Verteilungsparametern aus der a-posteriori Verteilung im Sinne der Bayes-Theorie vorgenommen. Während DA Algorithmus eine gemeinsame Verteilung modelliert, konstruiert MICE Algorithmus Zug um Zug bedingte Verteilung. Außerdem kommt die Konvergenz bei dem MICE Algorithmus meist schnell zustande, weil meist nur $10 - 20$ Iterationen für jeden vervollständigten Datensatz benötigt werden. Beim DA Algorithmus werden ca. 200 Iterationen vor dem ersten Datensatz und zwischen zwei unabhängigen Datensätzen empfohlen.
Nach den Simulationen hat sich gezeigt, dass DA und MICE Algorithmus für MCAR und MAR Mechanismus sehr gute Ergebnisse erzielen. Im Falle eines MNAR Mechanismus der fehlenden Daten, erzielt kein Verfahren zufriedenstellende Ergebnisse. Die Bestimmung des Fehlendmechanismus, vor allem die Unterscheidung zwischen MAR und MNAR, ist wichtig, um sicherzustellen, dass die Voraussetzungen für eine Imputation gegeben sind.

Die Methode der Multiple-Imputation zu den fehlenden Daten bezogen sich in dieser Seminararbeit nur auf Daten mit ignorierbarem Fehlendmechanismus. Zum Vorgehen bei nicht ignorierbarem Fehlendmechanismus werden zum Beispiel bei Little u. Rubin (2002) verschiedene Modelle, wie etwa spezielle Likelihoodmodelle, vorgeschlagen.

Literaturverzeichnis

[van Buuren 2007] Buuren, Stef van: Multiple Imputation of Discrete and Continuous Data by Fully Conditional. In: *Statistical Methods in Medical Research* 16 (2007), S. 219–242

[van Buuren u. Groothuis-Oudshoorn 2011] Buuren, Stef van ; Groothuis-Oudshoorn, Karin: mice: Multivariate Imputation by Chained Equations in R. In: *Journal of Statistical Software* 45 (2011), Nr. 3, S. 1–67

[Dempster u. a. 1977] Dempster, A. P. ; Laird, N. M. ; Rubin, D. B.: Maximum likelihood from incomplete data via the EM algorithm. In: *Journal of the Royal Statistical Society,Series B* 39 (1977), Nr. 1, S. 1–38

[Enders 2010] Enders, Craig K.: *Applied Missing Data Analysis*. New York : Guilford Press, 2010

[Little u. Rubin 2002] Little, R. ; Rubin, D.: *Statistical Analysis with Missing Data*. Hoboken, USA : Wiley & Sons, 2002

[Novo 2013] Novo, Alvaro A.: Package 'norm': Analysis of multivariate normal datasets with missing values. (2013)

[Rässler 2000] Rässler, Susanne: Ergänzung fehlender Daten in Umfragen. In: *Journal of Economics and Statistics (Jahrbuecher fuer Nationaloekonomie und Statistik)* 220 (2000), Nr. 1, S. 64–94

[Rässler u. a. 2013] Rässler, Susanne ; Rubin, Donald B. ; Zell, Elizabeth R.: Imputation. In: *Wiley Interdisciplinary Reviews: Computational Statistics* 5 (2013), Nr. 1, S. 20–29

[Rubin 1976] Rubin, Donald B.: Inference and Missing Data. In: *Biometrica* 63 (1976), Nr. 3, S. 581–592

[Rubin 1987] Rubin, Donald B.: *Multiple Imputation for Nonresponse in Surveys*. New York, USA : John Wiley & Sons, 1987

[Schafer 1997] Schafer, Joseph L.: *Analysis of Incomplete Multivariate Data*. New York : Chapman and Hall, 1997

[Schafer u. Olsen 1998] Schafer, Joseph L. ; Olsen, Maren K.: Multiple Imputation for Multivariate Missing-Data Problems: A Data Analyst's Perspective. In: *Multivariate Behavioral Research* 33 (1998), Nr. 4, S. 545–571

[Spieß 2008] Spiess, Martin: *Missing-Data Techniken: Analyse von Daten mit fehlenden Werte*. Hamburg : Lit Verlag, 2008

[Tanner u. Wong 1987] Tanner, M. A. ; Wong, W: The calculation of posterior distributions by data augmentation. In: *Journal of the American Statistical Association* 39 (1987), Nr. 1, S. 1–38

Anhang

Statistische Software und benutzte Funktionen

Die Simulationen in dieser Arbeit wurden mit dem statistischen Programmpaket R programmiert. In diesem Kapitel des Anhangs werden die wichtigsten implementierten Funktionen für die Behandlung fehlender Daten beschrieben.

prelim.norm(): Führt für den EM-Algorithmus oder für Imputation vorbereitende Transformationen der unvollständigen Datenmatrix durch. Unter anderem werden die beobachteten Daten standartisiert. Des Weiteren werden die Zeilen nach ihrem Fehlendmuster sortiert und der Code der fehlenden Daten wird zu NA geändert, falls noch nötig.

em.norm(): Berechnet die Maximum-Likelihood-Schätzungen der Parameterwerte von unvollständigen Datensätzen mit dem EM Algorithmus.

da.norm(): Berechnet die Maximum-Likelihood-Schätzungen der Parameterwerte von unvollständigen Datensätzen mit dem DA Algorithmus.

imp.norm(): Führt Multiple-Imputation unter der Annahme einer multivariaten Normalverteilung aus.

mice(): Führt Multiple-Imputation unter Anwendung von MICE Algorithmus aus.

pooling(): Kombiniert die Analyseergebnisse aus den m imputierten vollständig Datensätzen zu einem Gesamtergebnis