Seminararbeit

Imputationsverfahren

Autor: Minh Ngoc Nguyen

Betreuer: Eva Endres

2. Januar 2015

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

Fehlende Daten oder unvollständige Daten stellen ein häufiges Problem im Rahmen von empirischer Untersuchung dar. Es gibt vielfältige andere Umstände, welche fehlende Daten verursachen. Werden die Daten beispielsweise durch eine Befragung erhoben, kann der Ausfall durch Antwortverweigerung oder mangelndes Wissen des Befragten verursacht werden. 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 ineffizienten Schätzern führen Little and Rubin (2002).
Zur Behandlung der fehlenden Daten stehen mehrere Vorgehensweisen zur Verfügung (für Details siehe (Little and Rubin, 2002)). Zum einen besteht die Möglichkeit, die Datensätze mit fehlenden Daten ganz oder teilweise zu entfernen, die so genannte complete oder available case analysis. Dieses Vorgehen reduziert jedoch die verfügbare Information meist erheblich und kann außerdem zu ineffizienten Schätzern führen. Ein weiteres Vorgehen ist die Verwendung von Methoden der Imputation. Man unterscheidet zwischen Single-Imputation und Multiple-Imputation. Für jeden fehlenden Wert ersetzt Single-Imputation nur einen plausiblen Wert, beispielsweise durch den Mittelwert der vorhandenen Beobachtung in diese Variable. Hauptproblem ist, dass dieses Verfahren die Unsicherheit bei der Imputation nicht berücksichtigt. Aus diesem Grund führt es in allgemein zu unterschätzten Varianzen und zu signifikanten p-Werten Little and Rubin (2002). Eine Weiterentwicklung der Single-Imputation stellen die Methoden der Mutiple-Imputation dar. Statt nur einen Wert einzusetzen, generiert man mehrere Imputationen und somit die Unsicherheit über die unbeobachteten tatsächlichen Werte Rechnung getragen wird.
In dieser Seminararbeit sollten theoretische Grundlagen der Multipe-Imputation vorstellen. Die geschieht in Kapitel 2 durch einen Einblick in deren Mechanismen, die zum Fehlen führen. 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 er-

setzt werden. Kapitel 5 beschäftigt mit einem Beispiel aus der Simulation, dabei zu deren Veranschaulichung zu dienen. Abschließend fasst Kapitel 6 mit einem Überblick über die Kernaussage der Arbeit deren wichtigste Punkte zusammen.

2 Fehlende Daten Mechanismen

Der Begriff des Fehlendmechanismus, auch Missing Data Mechanisms genannt, geht auf B.Rubin (1987) zurück. bzw. deren Neuauflage Little and Rubin (2002). Für folgende Notation lässt sich der beobachtete Teil als Y_{beob} und der nichbeobachtete als Y_{fehl} von vollständigen Datenmatrix Y dargestellt. $Y = (Y_{beob}, Y_{fehl})$. Man definiert zusätzliche dafür eine Indikatormatrix für die fehlenden Daten

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

R wird im Zusammenhang mit den Mechanismen als Zusammenschluss von Zufallsvariablen betrachtet, die durch eine gemeinsame Verteilung charakterisiert werden können. Diese Verteilung erfasst mögliche Beziehungen zwischen dem Auftreten von fehlenden Werten und Ausprägung der Zufallsvariablen. Man geht davon aus, dass der Mechanismus der fehlenden Daten durch die bedingte Verteilung von R gegeben Y beschrieben wird $g(R|Y, \xi)$. Unbekannte Parameter der Verteilung werden durch ξ charakterisiert.
Eine Betrachtung dieser bedingten Verteilung motiviert die Klassifizierung in MCAR, MAR, MNAR (Little and Rubin, 2002). Dabei sind jeweils die Werte derjenigen Fälle maßgeblich, die unbeobachtet vorliegen. Demzufolge unterscheidet man in
MCAR: Missing-Completely-at-Random
Missing-Completely-at-Random bedeutet, dass der Mechnismus der hinter den fehlenden Werten weder von den fehlenden Werten selbst noch von den kompletten Daten abhängig ist. Formal gilt

$$g(R|Y; \xi) = g(R|\xi) \tag{2.1}$$

Ein Beispiel hierfür wäre, liegt es MCAR vor, wenn die Wahrscheinlichkeit für deren Fehlen weder vom Einkommen selbst noch vom Alter abhängt. Aus der obigen Definition ist die Annahme eines MCAR-Mechanismus in den meisten Fällen restriktiv. Fehlende Werte, die MCAR sind, können in der Analyse ignoriert werden, weil sie kein Problem darstellen. In der Praxis tritt MCAR selten auf.
MAR: Missing-at-Random

Von Missing-at-Random spricht man, wenn die Verteilung der fehlenden Daten unabhängig von den fehlenden Daten selbst ist, kann aber von anderen beobachteten Daten abhängen. Die mathematische Formulierung lautet wie folgt

$$g(R|Y;\xi) = g(R|Y_{beob};\xi) \tag{2.2}$$

Im Beispiel mit beiden Variablen Einkommen und Alter, wären fehlende Einkommen dann MAR, wenn die Wahrscheinlichkeit des Fehlens bzw. des Beobachtens des Einkommens vom Alter abhängt, nicht aber zusätzlich von der Höhe des Einkommens.
MNAR:Missing-Not-at-Random
Wenn die Verteilung von Ausfällen der fehlenden Daten selbst auch von fehlenden Werten im Datensatz abhängt, dann bezeichnet man dies als Missing-Not-at-Random. Es gilt

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

Die fehlenden Einkommensangaben wären etwa dann MNAR, wenn die Wahrscheinlichkeit für das Fehlen von Werten der Variablen Einkommen, selbst nach Konditionieren auf das Alter, von der Höhe des Einkommens selbst abhängt.
In Ergänzung zu formalen Konzepten von MCAR, MAR und MNAR wurde von Little and Rubin (2002) die Ignorierbarkeit von Fehlendmechanismen definiert. Der Fehlendmechanismus wird als ignorierbar bezeichnet, wenn

- Die Daten MAR sind, d.h. $g(R|Y;\xi) = g(R|Y_{beob};\xi)$
- Die Parameter des Fehlendmechanismus ξ und die Parameter über die Parameter θ der interessierende Verteilung $f(Y_{beob}, Y_{fehl}|\theta)$ unabhängig sind.

3 Multiple-Imputation

3.1 Grundkonzepte der Multiple-Imputation

Die Idee der Multiple-Imputation wurde von B.Rubin (1987) entwickelt. Es handelt sich um Methoden, die für jeden fehlenden Wert m Werte ersetzen, wobei $m > 1$ gilt. Dies führt zu m verschiedenen vervollständigten Datensätzen. Diese m vervollständigten Datensätze enthalten somit die gleichen beobachteten Werte und unterschiedliche imputierte Werte an den ursprünglich fehlenden Stellen. An jedem der vervollständigten Datensätze kann nun standardmäßig eine statistische Analyse durchgeführt, wobei die dann notwendigerweise schwankenden Schätzwerte die Unsicherheit des Datenausfalls und damit die Imputation widerspiegeln. Die verschiedenen Ergebnisse lassen sich schließlich kombinieren. Das Vorgehen der Multiple-Imputation lässt sich am besten grafisch veranschaulichen (vgl. Abbildung 3.1).

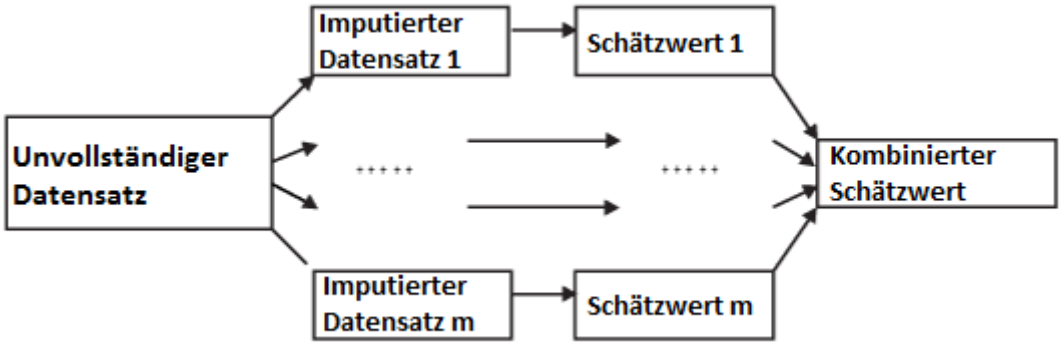

Abbildung 3.1: Grundkonzepte der Multiple-Imputation

3.2 Kombinationsregel der vervollständigten Datensätze

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

Regressionskoeffizienten, sollen anschließend zu einem einzelnen kombiniert werden. Die verschiedenen Schätzungen aus den m vervollständigten Datensätze zu kombinieren wurde von Little and Rubin (2002) folgende Formel veröffentlicht.
Sei Q der interessierende Parameter und V dessen zugehörigen Varianz. Somit erhält man aus der Analyse der m vervollständigten Datensätze 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 Varianz $\hat{V}_{MI}$ ist dann

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

mit der Within-Varianz

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

und der Between-Varianz

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

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 Stichprobe sind Test und Konfidenzintervalle approximativ t-verteilt. 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

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

3.3 Wahl der Anzahl m Imputationen

Vor der Durchführung einer Multiple-Imputation muss man erforderlich festlegen, wie viele Imputationen m ausgeführt werden sollen. Nach B.Rubin (1987) reichen üblicherweise schon kleine Anzahlen m zwischen 3 und 10 Wiederholungen aus, stabile Ergebnisse zu erhalten. In statistischen Programmpaketen wählt man meist als Standardwert $m = 5$. Nach B.Rubin (1987) beträgt die relative Effizienz (RE) einer auf m Imputationen basierenden Schätzung

$$RE = (1 + \frac{\gamma}{m})^{-1} \tag{3.7}$$

wobei γ Anteils der fehlenden Werte bezeichnet. Tabelle 3.3 zeigt die erzielten relativen Effizienzen für unterschiedliche Anzahl an Imputationen und verschiedenen Anteil an fehlenden Werten.

		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

Wie man sieht ist eine geringe Anzahl an Imputationen in den meisten Fällen vollkommen ausreichend. Nur bei einer sehr hohen Rate an fehlender Information lässt sich durch das Erhöhen dieser Anzahl ein entscheidender Effizienzgewinn erzielen.

4 Imputationsverfahren

In den folgenden Abschnitt werden nun verschiedene Verfahren vorgestellt, die bei der Durchführung einer MI dafür sorgen, dass die fehlenden Werte eines Datensatzes ersetzt werden.
Zur besseren Übersicht und leichterem Verständnis soll Notation, die im Folgenden gilt, eingeführt und erläutert werden. Im Kontext repräsentiert Y_j mit $j = 1, \ldots, p$ ein der Variablen, die fehlende Werte beinhalten. $Y = (Y_1, \cdots, Y_p)$. Es ist weiterhin notwendig, die beobachteten und fehlenden Elemente zu definieren. Der beobachtete Teil von Y_j lässt sich als $Y_{beob,j}$ darstellen und der fehlende als $Y_{fehl,j}$. Dementsprechend wird $Y_{beob} = (Y_{beob,1}, \cdots, Y_{beob,p})$ der beobachtete und $Y_{fehl} = (Y_{fehl,1}, \cdots, Y_{fehl,p})$ der unbeobachtete Teil von Y bezeichnet. θ stellen die unbekannten Parameter des interessierenden Modells dar.
Um die Notation einfach zu halten, ist nun der Fehlendmechanismus ignorierbar sofern die Ausprägung von Y_{fehl} zufällig fehlen (MAR).

4.1 Verfahren auf Basis des Maximum-Likelihood-Ansatzes

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 wie folgt

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

Wie bei normaler Maximum-Likelihood gilt es hier das Maximum dieser Funktion zu finden. Im Falle von fehlenden Daten sind aber kompliziert, da die benögtigte Fisher-Informationsmatrix bzw. erwartete Fisher-Informationsmatrix rechenaufwendig ist zu

berechnen. Das Maximierungsproblem kann in diesen Fällen durch die Anwendung von dem Expectation-Maximization Algorithmus Dempster et al. (1977), der von Little and Rubin (2002) ausführlich beschrieben wird, gelöst werden.
Die Grundidee des EM Algorithmus ist es, zunchst die fehlende Werte durch Schätzungen zu ersetzen, damit eine Parameterschätzung durchzuführen, auf derer wiederum die fehlenden Werten neu geschätzt werden. Die fehlende Werte und die Parameter werden so lange neu geschätzt bis es zur Konvergenz kommt Little and 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 Y_{beob} 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 and Rubin, 2002):

1. **E-Schritt**: Berechne den bedingten Erwartungswert von Log-Likelihood gegeben den beobachteten Werten und dem jeweilig aktuellen θ^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 it.

2. **M-Schritt**: Finde dass θ^{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 Iteration $|\theta^{(t+1)} - \theta^{(t)}| \leq \epsilon$, 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 and Rubin (2002) zeigen, dass unter bestimmten Bedingungen wie einer linearen loglikelihood der EM Algorithmus zuverlässig konvergiert. Das heißt jede Iteration erhöht die Log-Likelihood $l(Y_{obs}|\theta)$, deshalb EM-Algorithmus meist einfach zu konstruieren ist und sich jeder Schritt leicht interpretieren lässt.
Ein Nachteil stellt jedoch die unter Umsätnden langsame Rate der 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, vgl. (Little and Rubin, 2002).
Ein weiterer Nachteil dieser Vorgehensweise besteht darin, dass der Algorithmus lediglich die bedingten Erwartungswerte für die Imputation verwendet ist und somit die Varianz ist schwierig zu gewinnen.Somit wird die Unsicherheit der Schätzung nicht berücksichtigt. Aus diesem Grund sind bayesianische Ansätze in Verbindung mit Multiple-Imputation von fehlenden Werten vorzuziehen.

4.2 Verfahren auf Basis des Bayes-Ansatzes

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

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

Wird häufig ein zweistufiger Ziehungen verwendet. Durch diese Züge werden die Unsicherheit in der Vorhersage der einzelnen fehlenden Werte gegeben Parameter und die Unsicherheit über die Parameterschätzung widergespiegelt (ässler:2012**?**). 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.

Dan nun 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.

4.2.1 Data-Augmentation Algorithmus

Wie bereits erwähnt, werden innerhalb der bayesianischen Theorie alle Schlüsse für die unbekannten Parameter einer Verteilung aus der a-posteriori Verteilung gezogen. Da

die a-posterior häufig nicht direkt bestimmbar ist, wird auf die leichter zu ermittelnde Verteilung $f(\theta|Y_{beob}, Y_{fehl})$ zurückgegriffen. Hier wird der Data Augmentation Algorithmus vorgestellt, der nach Tanner and Wong (1987) 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. Bei DA Algorithmus werden die fehlenden Daten durch zufälliges Ziehen aus der bedingten Prädiktivverteilung unter Annahme eins Parameters, der aus der a-posteriori Verteilung in vorherige Iteration gezogen wird, ersetzt. Der Algorithmus gliedert sich in zwei Schritte, die bis zum Erreichen von Konvergenz t-mal iterativ wiederholt werden: Imputation-Schritt (I-Schritt) und Posterior-Schritt (P-Schritt).

1. **I-Schritt**: Ziehe die fehlende Werte

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

   gemäß der sog. prädiktiven bedingten Verteilung von Y_{fehl}, d.h. gegeben die beobachteten Werten und einen aktuellen Parameter von $\theta^{(t)}$.

2. **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

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

Imputations- und Posteriorischritt werden solange iterativ durchgeführt bis die approximierende a-posteriori Verteilung von $f(\theta^{(t)}|Y_{beob})$ gegeben eine stationäre Verteilung $f(\theta|Y_{beob})$ konvergiert. Das Ergebnis des Verfahrens liefert eine Markov Kette $(\theta^{(t)}, Y_{fehl}^{(t)}) : t = 0, 1, \cdots$ mit der stationären Verteilung von $f(\theta|Y_{beob})$.
Zu Beginn des Verfahrens muss für den Verteilungsparameter θ ein Startwert $\theta^{(0)}$ festgelegt werden, z.B. für multivariate Normalverteilung sind μ und Σ ausreichend. Die Ergebnisse des Expectation-Maximization Algorithmus werden hier häufig verwendet.

4.2.2 Chained-Equations Algorithmus

Aufgrund des unterschiedlichen Messniveaus der einzelnen Variablen, etwa eine Mischung aus stetigen, diskret-metrischen und qualitativen Variablen, stellt sich das Problem, ein multivariates Modell anzupassen bzw. multivariate Imputation der fehlenden Werte durchzuführen. Der Grund dafür liegt darin, dass es schwierig sein kann eine gemein-

same Spezifikation aller Variablen zu finden. Deshalb hat van Buuren and Groothuis-Oudshoorn (2011) den Ansatz von Chained-Equation vorgeschlagen, der die fehlenden Werte durch eine Verkettung von univariaten Imputationen ersetzt. Bei diesem Verfahren wird jede Variable separate behandelt, wobei für jede ein Imputationsmodell in Abhängigkeit vom Messniveau gewählt wird. Dieses Verfahren ist in der englischen Fachsprache auch unter anderen Namen bekannt wie Stochastic Relaxation, Regression Switching, Sequential Regression, Incompatible MCMC, usw. (van Buuren and Groothuis-Oudshoorn (2011)).
Die Grundidee dieses Ansatzes besteht darin, dass diese mit Hilfe von Zügen aus den bedingten Verteilungen konstruieren die gemeinsame Verteilung modelliert werden zu können. Jede Variable mit fehlenden Werten wird eine separate bedingte Verteilung spezifiziert ((van Buuren and Groothuis-Oudshoorn, 2011)). Ausgehende von einer anfänglichen Imputation, erzeugt MICE die Imputationen durch Iteration über die bedingte Dichte.
Bei CE-Algorithmus werden am Anfang die fehlenden Werte durch Werte ersetzt, die gemäß Randverteilung der beobachteten Variablen gezogen werden. Dann startet ein iteratives Verfahren (Gibbs Sampling), wo bei in jedem Schritt die Variablen sukzessive behandelt werden ((van Buuren and Groothuis-Oudshoorn, 2011)). Die Iteration t von Gibbs Sampler Algorithmus für Generierung $Y_{fehl}^{(t)}$ aus $Y_{fehl}^{(t-1)}$ ist dann gegeben durch

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

Wobei $Y_j^{(t)} = (Y_{beob,j}^{(t)}, Y_{fehl,j}^{(t)})$ die j-te vervollständigte Variable bei Iteration t ist. Konvergenz kann daher recht schnell, also die wird bis zu einer vorgegebenen Anzahl der Iteration $10 - 20$ weiterdurchgeführt ((van Buuren and Groothuis-Oudshoorn, 2011)). Diese Sequenz erzeugt wieder eine Mavkow-Kette, die unter geeigneten Bedingungen gegen die gemeinsame Verteilung von Y_{fehl} und θ gegeben Y_{beob} geht.
CE-Algorithmus bietet den Vorteil, dass komplexe Datenstrukturen berücksichtigt werden können. Es unterscheidet vier Typen von Variablen: Zählvariablen, stetige Variable, kategoriale 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.

4.2.3 Fazit

Die Multiple-Imputation mit CE-Algorithmus 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.
Bei den CE-Algorithmus gibt es keine Einschränkung für jede zu behandelnde Variable eine bedingte Verteilung gegeben die restlichen Variablen findet zu werden. Im DA-Algorithmus ist es hingegen schwer, eine gemeinsame Verteilung aller Variablen zu bestimmen. Im Unterschied zum DA-Algorithmus kommt die Konvergenz bei dem CE-Algorithmus meist schnell zustande, weil meist nur 5-10 Iterationen für jeden vervollständigten Datensatz bei dem CE-Algorithmus benötigt werden.

5 Simulationsstudie

Im Rahmen einer Simulationsstudie sollen in diesem Abschnitt die in dieser Arbeit angesprochenen Methoden der Datenergänzung auf unvollständige Datensätze angewendet werden.

5.1 Simulationsdesign

Für die Simulationsstudie wurden die drei Variablen X_1, X_2, Y mit jeweils 10000 Beobachtungen generiert. X_1 und X_2 werden aus einer bivariaten Standardnormalverteilung mit einer Korrelation von 0.5 gezogen. Zudem sei $Y = \beta_1 X_1 + \beta_2 X_2 + \epsilon$ mit $\epsilon \sim N(0, 1)$ und $\beta_1 = \beta_2 = 1$. Um aus dem Datensatz ohne fehlende Beobachtungen einen Datensatz mit fehlenden Werten zu erzeugen, wurden in Abhängigkeit von den jeweiligen Fehlendmechanismen Beobachtungen entfernt:

- MCAR: Beobachtungen in X_2 fehlen mit einer Wahrscheinlichkeit von $p\%$ unabhängig von Y und X_1
- MAR abhängig von X_1: Beobachtungen in X_2 fehlen wenn X_1 kleiner als das $p\%$ Quantil von X_1 ist
- MAR abhängig von Y: Beobachtungen in X_2 fehlen wenn Y kleiner als das $p\%$ Quantil von Y ist
- MNAR: Beobachtungen in X_2 fehlen wenn X_2 kleiner als das $p\%$ Quantil von X_2 ist.

Es sollen für jeden Mechanismus verschiedene Anteile von fehlenden Daten erzeugt werden, dies sind $30\%, 50\%, 70\%$ und 90%. Die Parameter, mit denen die Qualität der Verfahren geprüft wird, sind die Regressionskoeffizienten β_2 und seine Standardfehler sowie der Mittelwert μ_{X_2} und die Varianz $\sigma^2_{X_2}$und Korrelation $\rho_{X_1X_2}$zwischen X_1 und X_2.

5.2 Verwendung des Data-Augmentation Algorithmus

Zuerst wird die Multiple-Imputation auf Basis eines Data-Augmentation Algorithmus durchgeführt. Hier werden die vervollständigten Datensätze unter Verwendung einer variierenden Anzahl an zu imputierenden Werten betrachtet, mit $m = 3, 5$ und 10. Die Imputationen werden mit dem Package norm der Software R verwendet (Novo, 2013). Die jeweiligen unvollständigen Datensätze werden dem DA-Algorithmus mit 20 Iterationen unterworfen und pro Fehlendmechanismus sowie pro Fehlwertanteil ein vollständiger Datensatz erstellt.
Betrachtet man Tabelle 5.1, so ergibt Imputation mit DA-Algorithmus durchaus sinnvoll ersetzen. Für alle Fälle treffen die Parameterschätzer β_2 treffen den wahren Wert von 1 sehr gut. Lediglich die Standardfehler steigen mit zunehmendem Anteil der fehlenden Werte leicht an. Über alle m Imputationen eines unvollständigen Datensatzes hinweg schwanken die Mittelwerte μ_{X_2} in einem sehr engen Bereich um den wahren Wert herum, zwischen -0.2 und 0.2. Die mittlere geschätzte Varianz und Korrelation bleiben annähernd erhalten. Durch das Vorliegen eines MCAR Fehlendmechanismus konnten durch den DA-Algorithmus, sehr gute imputierte Datensätze erzeugt werden.

m	Fehlwertanteil	β_2	Standardfehler	μ_{X_2}	$\sigma^2_{X_2}$	$\rho_{X_1X_2}$
3	30	1.0066	0.0121	-0.0064	1.0078	0.50605
	50	1.0119	0.0123	-0.0075	1.0029	0.51158
	70	1.0121	0.0148	0.0170	0.9954	0.50539
	90	1.0176	0.0152	0.0330	1.0186	0.50417
5	30	1.0097	0.0119	-0.0046	1.0086	0.50224
	50	1.0085	0.0116	-0.0115	1.0308	0.51370
	70	1.0163	0.0151	0.0109	1.0178	0.51264
	90	1.0403	0.0163	0.0176	1.0429	0.53444
10	30	1.0066	0.0121	-0.0109	1.0061	0.50605
	50	1.0119	0.0123	-0.0109	1.0052	0.51158
	70	1.0121	0.0148	0.0167	0.9950	0.50540
	90	1.0179	0.0152	0.0291	1.0274	0.50418

Tabelle 5.1: Simulationsergebnisse der m imputierten Datensätze bei MCAR und Data-Augmentation Algorithmus

Ein ähnliches Bild ergibt sich bei Vorliegen eines MAR-Mechanismus der von X_1 in Tabelle 5.2 bzw. Y in Tabelle 5.3 abhängig ist. Die Paramterschätzer sind unverzerrt, die Standardfehler wachsen durchschnittlich nur leicht bei zunehmendem Anteil der fehlenden Werte. Auch die Struktur der Daten bleibt nach den Imputationen erhalten, was

sich in plausiblen für $\mu_{X_2}, \sigma^2_{X_2}, \rho_{X_1X_2}$ wiedergibt.

m	Fehlwertanteil	β_2	Standardfehler	μ_{X_2}	$\sigma^2_{X_2}$	$\rho_{X_1X_2}$
3	30	1.0068	0.0129	0.0177	0.9948	0.48672
	50	1.0259	0.0122	0.0325	0.9642	0.48616
	70	1.0091	0.0153	0.0561	0.9698	0.47689
	90	1.0453	0.0168	0.1362	0.9097	0.44518
5	30	1.0057	0.0126	0.0131	0.9952	0.48987
	50	1.0234	0.0119	0.0201	0.9781	0.49584
	70	1.0047	0.0172	0.1127	0.9067	0.43830
	90	1.0619	0.0159	0.1518	0.8998	0.43623
10	30	1.0068	0.0129	0.0132	0.9984	0.48672
	50	1.0233	0.0119	0.0179	0.9794	0.49584
	70	1.0046	0.0171	0.1142	0.9048	0.43831
	90	1.0619	0.0159	0.1497	0.8994	0.43623

Tabelle 5.2: Simulationsergebnisse der m imputierten Datensätze bei MAR abhängig von X_1 und Data-Augmentation Algorithmus

Betrachtet man Tabelle 5.4, so sind die Schätzer der einzelne Datensätze beim MNAR Fehlendmechanismus stark verzerrt. Die Standardfehler der Schätzungen sind im Vergleich zu den bereits erläuterten Ergebnissen bei anderen Fehlendmechanismen deutlich erhöht. Die Mittelwerte sowie die Varianzen weichen weichen stark vom wahren Wert ab. Interessant ist, dass die Korrelation zwischen X_1 und X_2 trotz allem relativ gut nachgebildet werden kann

Der DA Algorithmus ist nach den vorliegenden Erkenntnissen eine adäquate Möglichkeit unvollständige Datensätze zu vervollständigen. Zumindest gilt dies bei Vorliegen eine MCAR oder MAR Fehlendmechanismus. In diesen Fällen führt die Imputation zu unverzerrten Schätzern, basierend auf den ergänzten Datensätzen. Zudem bleiben Mittelwerte und Varianzen der vervollständigten Variablen erhalten und auch die Abhängigkeit zur Variablen X_1 wird durch die Ergänzungen beibehalten. Wie bei den bereits behandelten Ansätzen zur Datenergänzung, ist auch dieses Verfahren nicht in der Lage mit einem MNAR Fehlendmechanismus umzugehen. Alle Größen weichen stark von ihren wahren Werten ab. Lediglich die Korrelation zwischen X_1 und X_2 kann einigermaßen gut wiedergegeben werden.

m	Fehlwertanteil	β_2	Standardfehler	μ_{X_2}	$\sigma^2_{X_2}$	$\rho_{X_1X_2}$
3	30	1.0068	0.0127	0.0144	0.9831	0.49971
	50	1.0018	0.0126	0.0260	0.9747	0.49205
	70	1.0098	0.0139	0.0967	0.8941	0.47034
	90	1.0395	0.0165	0.1130	0.9064	0.44369
5	30	1.0067	0.0127	0.0123	0.9846	0.49971
	50	1.0017	0.0127	0.0257	0.9729	0.49205
	70	1.0097	0.0138	0.0952	0.8892	0.47034
	90	1.0394	0.0164	0.1110	0.9060	0.44369
10	30	1.0068	0.0127	0.0098	0.9881	0.49971
	50	1.0103	0.0128	0.0066	0.9974	0.50130
	70	1.0098	0.0139	0.0968	0.8885	0.47034
	90	1.0394	0.0165	0.1090	0.9066	0.44369

Tabelle 5.3: Simulationsergebnisse der m imputierten Datensätze bei MAR abhängig von Y und Data-Augmentation Algorithmus

5.3 Verwendung des Chained-Equations Algorithmus

Dieselbe Struktur der Simulationsstudie wie im vorangegangenen Abschnitt liegt auch bei der Bewertung der Imputation durch Chained-Equations Algorithmus, die mit Package mice in der Software R durchgeführt wird (van Buuren and Groothuis-Oudshoorn (2011)).

Die Tabelle 5.5 zeigt die Ergebnisse bei einem MCAR-Mechanismus. Die Parameterschätzer sind nahezu unverzerrt bei einem geringen Standardfehler. Betrachtet man die verbleibenden drei Größen, so ergibt sich die Tatsache, dass CE-Algorithmus diese Größen gut nachbilden kann. Sowohl Mittelwert als auch Varianz der Variable X_2 werden gut widerspiegelt.

Tabelle 5.6 und 5.7 zeigen auch, dass CE-Algorithmus in der Lage für ein MAR-Fall ist sinnvolle Werte zu ersetzen. Die imputierten Datensätze sind in der Lage unverzerrte Schätzer bei niedrigen Standardfehlern zu erzeugen. Mittelwert, Varianz und die Abhängigkeit der Variable X_1 und X_2 können zwar nicht so exakt wie bei dem MCAR-Mechanismus wiedergegeben werden. Aber diese Werte liegen immer noch nahe den wahren Werten.

Bei Vorhanden eines MAR Mechanismus (Tabelle 5.8) steht Problem da. Bei erhöhten Standardfehlern sind die Schätzer stark verzerrt. Mittelwerte und Varianzen werden nicht korrekt wiedergegeben.

Abschließend kann festgehalten werden, dass CE-Algorithmus bei MCAR und MAR-

m	Fehlwertanteil	β_2	Standardfehler	μ_{X_2}	$\sigma^2_{X_2}$	$\rho_{X_1X_2}$
3	30	1.1628	0.0196	0.3101	0.5630	0.46311
	50	1.2348	0.0217	0.5491	0.4198	0.42106
	70	1.3803	0.0270	0.8558	0.3050	0.39576
	90	1.4644	0.0355	1.4535	0.1859	0.33250
5	30	1.1628	0.0296	0.3089	0.5634	0.46311
	50	1.2358	0.0217	0.5489	0.4203	0.42106
	70	1.3803	0.0270	0.8547	0.3034	0.39576
	90	1.4644	0.0355	1.4524	0.1860	0.33250
10	30	1.1628	0.0196	0.3068	0.5660	0.46311
	50	1.1254	0.0237	0.5331	0.4313	0.44019
	70	1.3803	0.0138	0.8558	0.3034	0.39576
	90	1.4644	0.0143	1.4512	0.1859	0.33250

Tabelle 5.4: Simulationsergebnisse der m imputierten Datensätze bei MNAR und Data-Augmentation Algorithmus

Mechanismus in der Lage ist, die Verteilung sinnvoll zu reproduzieren.

5.4 Vergleich

Hier werden zwei Methoden der Multiple-Imputation vergleicht. Beide Methoden werden dabei mit $m = 5$ imputiert und mit 50% des Anteils der fehlenden Daten. Die Ergebnisse aus Tabelle 5.9 zeigen, dass sich sowohl für die vollständige beobachtete Variable X_1 als auch für die Variable X_2, die fehlenden Beobachtungen enthält, bei einem MCAR und MAR sehr gut Ergebnisse ergeben. Die gemittelten Parameterschätzer treffen den wahren Wert bzgl. Der Variable X_2 bei gleichbleibendem Standardfehler nahe exakt. A Bei einem MNAR Fehlendmechanismus ergibt sich, dass der Parameter bei zunehmenden Standardfehler stark überschätzt wird. Alle gemittelte Parameterschätzer sind bei Betrachtung der p-Wert hochsignifikant.

m	Fehlwertanteil	β_2	Standardfehler	μ_{X_2}	$\sigma^2_{X_2}$	$\rho_{X_1X_2}$
3	30	1.0083	0.0135	-0.0090	1.0100	0.5033
	50	1.0104	0.0152	-0.0135	1.0124	0.5162
	70	1.0162	0.0198	0.0137	1.0238	0.5046
	90	1.0183	0.0160	-0.0112	1.0242	0.5153
5	30	1.0088	0.0138	-0.0086	1.0088	0.5072
	50	1.0151	0.0141	-0.0140	1.0109	0.5156
	70	1.0154	0.0134	0.0087	1.0331	0.5095
	90	1.0147	0.0126	-0.0124	1.0246	0.5138
10	30	1.0109	0.0123	-0.0091	1.0108	0.5066
	50	1.0104	0.0133	-0.0129	1.0124	0.5137
	70	1.0172	0.0137	0.0128	1.0276	0.5098
	90	1.0222	0.0138	-0.0060	1.0230	0.5158

Tabelle 5.5: Simulationsergebnisse der m imputierten Datensätze bei MCAR und Chained Equations Algorithmus

m	Fehlwertanteil	β_2	Standardfehler	μ_{X_2}	$\sigma^2_{X_2}$	$\rho_{X_1X_2}$
3	30	1.0008	0.0117	0.0067	1.0084	0.4957
	50	1.0154	0.0152	0.0177	0.9814	0.4930
	70	1.0100	0.0153	0.0522	0.9495	0.4787
	90	1.0443	0.0136	0.0736	0.9251	0.4764
5	30	1.0023	0.0118	0.0086	1.0088	0.4950
	50	1.0176	0.0127	0.0164	0.9802	0.4952
	70	1.0103	0.0143	0.0470	0.9573	0.4793
	90	1.0412	0.0129	0.0764	0.9297	0.4690
10	30	1.0076	0.0126	0.0073	1.0012	0.4974
	50	1.0129	0.0136	0.0166	0.9866	0.4939
	70	1.0107	0.0136	0.0515	0.9528	0.4795
	90	1.0448	0.0143	0.0804	0.9257	0.4708

Tabelle 5.6: Simulationsergebnisse der m imputierten Datensätze bei MAR abhängig von X_1 und Chained Equations Algorithmus

m	Fehlwertanteil	β_2	Standardfehler	μ_{X_2}	$\sigma^2_{X_2}$	$\rho_{X_1X_2}$
3	30	1.0045	0.0124	0.0047	1.0033	0.5003
	50	1.0020	0.0132	0.0116	0.9915	0.5008
	70	0.9992	0.0186	0.0442	0.9564	0.4873
	90	1.0289	0.0137	0.0369	0.9649	0.4688
5	30	1.0038	0.0118	0.0064	0.9991	0.5025
	50	1.0089	0.0145	0.0111	0.9931	0.5002
	70	1.0022	0.0146	0.0399	0.9686	0.4836
	90	1.0250	0.0123	0.0393	0.9636	0.4682
10	30	1.0104	0.0128	0.0048	0.9974	0.5044
	50	1.0078	0.0135	0.0102	0.9911	0.4967
	70	1.0023	0.0145	0.0434	0.9630	0.4859
	90	1.0285	0.0132	0.0431	0.9613	0.4707

Tabelle 5.7: Simulationsergebnisse der m imputierten Datensätze bei MAR abhängig von Y und Chained Equations Algorithmus

m	Fehlwertanteil	β_2	Standardfehler	μ_{X_2}	$\sigma^2_{X_2}$	$\rho_{X_1X_2}$
3	30	1.1638	0.0180	0.3009	0.5765	0.4664
	50	1.2455	0.0246	0.5379	0.4315	0.4343
	70	1.3611	0.0343	0.8486	0.3151	0.4011
	90	1.4529	0.0336	1.4280	0.1923	0.3629
5	30	1.1623	0.0175	0.3032	0.5746	0.4671
	50	1.2534	0.0231	0.5374	0.4282	0.4340
	70	1.3613	0.0316	0.8451	0.3134	0.3957
	90	1.4515	0.0347	1.4295	0.1933	0.3602
10	30	1.1713	0.0170	0.3014	0.5726	0.4673
	50	1.2498	0.0216	0.5366	0.4290	0.4360
	70	1.3636	0.0291	0.8474	0.3126	0.3970
	90	1.4580	0.0342	1.4316	0.1913	0.3543

Tabelle 5.8: Simulationsergebnisse der m imputierten Datensätze bei MNAR und Chained Equations Algorithmus

			MCAR	MAR (abhängig von X1)	MAR
DA	X_1	Schätzwert	0.9889	1.0139	1.1794
		Standardfehler	0.0130	0.0139	0.0133
		p-Wert	0.0000	0.0000	0.0000
	X_2	Schätzwert	1.0085	1.0233	1.2358
		Standardfehler	0.0116	0.0119	0.0217
		p-Wert	0.0000	0.0000	0.0000
CE	X_1	Schätzwert	0.9895	1.0169	1.1605
		Standardfehler	0.0137	0.0126	0.0136
		p-Wert	0.0000	0.0000	0.0000
	X_2	Schätzwert	1.0151	1.0176	1.2534
		Standardfehler	0.0141	0.0127	0.0231
		p-Wert	0.0000	0.0000	0.0000

Tabelle 5.9: Ergebnisse der Multiple-Imputation

6 Zusammenfassung

Diese Arbeit beschäftigt sich mit der Möglichkeit zur Behandlung der unvollständigen Werte in komplexen Datensätzen. Dabei werden die Methode der Multiple-Imputation vorgestellt. 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. Der einzige Nachteil, den die multiple Imputation aufweist, ist der größere Aufwand um die Imputation und die Analyse durchzuführen. In Zeiten von leistungsstarken Rechnern fällt dies kaum noch ins Gewicht.
Dazu werden zwei Verfahren, die Spezialfälle der MCMC Methode, vorgestellt. Diese Verfahren, die Data Augmentation und Chain-Equations sind, beruhen dem Prinzip der Multiple-Imputation. Während Chain-Equations Zug um Zug bedingte Verteilung konstruieren, modelliert Data Augmentation eine gemeinsame Verteilung. Nach den Simulationen hat sich gezeigt, dass die Multiple-Imputation mit beiden Ansätzen für MCAR und MAR sehr gute Ergebnisse erzielen. Im Falle eines MNAR-Mechanismus der fehlenden Daten, erzielt kein Verfahren zufriedenstellende Ergebnisse.

Literaturverzeichnis

B.Rubin, D. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York, USA: John Wiley & Sons.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society,Series B 39*(1), 1–38.

Little, R. and D. Rubin (2002). *Statistical Analysis with Missing Data*. Hoboken, USA: Wiley & Sons.

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

Tanner, M. A. and W. Wong (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association 39*(1), 1–38.

van Buuren, S. and K. Groothuis-Oudshoorn (2011). mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software 45*(3), 1–67.