PROFESSUR FÜR
ANGEWANDTE STATISTIK
DER FREIEN UNIVERSITÄT BERLIN

Seminararbeit

Inferenz mit unvollständigen Daten - ein Vergleich von singulären und multiplen Imputationsverfahren

Tobias Flörchinger

Gutachter(in):	Prof. Dr. Natalia Rojas-Perilla
Semester:	Wintersemester 2020/2021
Verfasser:	Tobias Flörchinger
Matrikel-Nr.:	608227
Studienfach:	Master Statistik

Abgabetermin: 21.02.2021

Hiermit erkläre ich an Eides Statt, dass ich die vorliegende Arbeit selbständig und ohne unerlaubte fremde Hilfe angefertigt, andere als die angegebenen Quellen und Hilfsmittel nicht benutzt und die den benutzten Quellen und Hilfsmitteln wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.

Frankfurt am Main, den 21. Februar 2021

Tobias Flörchinger

Abstract

In dieser Arbeit werden singuläre und multiple Imputationsverfahren hinsichtlich valider Inferenz miteinander verglichen. Das heißt, auf einem per Imputation vervollständingten Datensatz soll trotz der Ergänzungen korrekt auf die zugrundeliegende Population geschlossen werden können. Fehlende Werte in Datensätzen sind dabei durch Schweigeverzerrungen in Umfragen, bei Datenfusionen oder in synthetischen Daten möglich. Ist valide Inferenz nicht gegeben, sind Methoden des statistischen Schließens wie Konfidenzintervalle und Hypothesentest fehlführend. Diesbezüglich werden Daten mit Ausfällen simuliert und unter Verwendung des MICE Algorithmus sechs singuläre sowie drei multiple Imputationsverfahren verglichen, wobei die vollständigen Stichproben ohne Ausfälle als Benchmark dienen. Die zu schätzenden Vergleichsparameter sind der Erwartungswert einer Zufallsvariable sowie zwei Regressionskoeffizienten. Die Vergleichskriterien sind der relative Fehler, die Deckungsrate und die Breite der Konfidenzintervalle. Es zeigt sich, dass univariate bzw. deterministische singuläre Imputationsverfahren für alle Parameter hohe Verzerrungen in den Schätzungen aufweisen, welche die Deckungsraten in Richtung 0 drücken. Auch durch die übrigen singulären Imputationsverfahren ergibt sich stets eine zu niedrige Deckungsrate. Nur zu multiplen Imputationsverfahren kann valide Inferernz festgestellt werden. Es wird ersichtlich, dass multivariate Verfahren mit hinreichend stochastischen Komponenten und die multiple Imputation benötigt werden, um die mit dem Datenausfall verbundene Unsicherheit in der Realisierung der Ergänzung abzubilden. Das führt zu einer adäquaten Erhöhung der Varianzschätzer, was bei Unverzerrtheit der Parameterschätzung valide Inferenz ermöglicht.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

Umfragen sind ein zentrales Mittel, um Informationen über eine zugrundeliegende Population zu gewinnen. So wird beispielsweise über das Sozio-oekonomische Panel jährlich eine repräsentative Stichprobe deutscher Haushalte hinsichtlich ihrer sozio-oekonomischen Lage befragt und damit auf die Bundesrepublik Deutschland als Population geschlossen. Für die Durchfühung sind neben praktischen Schwierigkeiten insbesondere theoretische Herausforderungen gegeben. Neben der Bestimmung des Umfragedesigns sowie des Stichprobenverfahrens sind Überlegungen zu Problemen bei der Erhebung der benötigten Daten erforderlich. Mitunter empfinden Personen bestimmte Fragen als sensibel und sind nicht dazu bereit Antworten zu geben. Die Angabe des Alters kann beispielsweise als weniger kritisch angesehen werden als Einblick in das Gehalt zu gewähren. Unter anderem deswegen können zu Fragen Antwortausfälle zu verzeichnen sein. Jedoch erfordern insbesondere multivariate Analysen einen vollständigen Datensatz. Eine Person aufgrund der fehlenden Antwort aus der Umfrage zu entfernen, kann zum einen wegen des Informationsverlusts hinsichlich der ansonsten beantworteten Fragen und zum anderen aufgrund der mit der Anzahl der fehlenden Antworten schwindenden Datenbasis, problematisch werden. Die Imputation der fehlenden Werte ist hier ein adäquater Lösungsansatz, wobei fraglich ist, wie stetige Daten ergänzt werden sollen, damit trotzdem richtig auf die Population geschlossen wird.

Dementsprechend ist das Ziel dieser Arbeit, singuläre und multiple Imputationsverfahren im Kontext der validen Inferenenz zu vergleichen. Aus einer Auswahl von Imputationsverfahren sollen jene identifiziert werden, die zu valider Inferenz führen. Die Gründe weswegen ein Verfahren valide Inferenz liefert oder dazu nicht in der Lage ist, sollen hier ersichtlich werden. Somit wird auf allgemeine Anforderungen an valide Imputationverfahren geschlossen.

Dazu definiert Kapitel 2 das theoretische Fundament sowie die damit ermöglichten singulären und multiplen Imputationsverfahren. Das daran anschließende Kapitel widmet sich der Methodik des Vergleichs. Der Simulationsstudie folgend, werden nacheinander die Datensimulation mit Ausfällen, die Parameterschätzung und die Auswahl der Vergleichskriterien thematisiert. In Kapitel 4 werden die Ergebnisse zu valider Inferenz anhand der Vergleichskriterien zusammengetragen und im Rahmen der Zielsetzung diskutiert. Abschließend fasst Kapitel 5 zusammen und legt thematische Anknüpfungspunkte dar.

2 Grundlagen

Ist $\mathcal{Y}$ eine Population mit der Realisation $(y_{i1}, \ldots, y_{ip})$ des stetigen Zufallsvektors $(X_1, \ldots, X_p)$, so ist durch n-faches ziehen aus $\mathcal{Y}$ auf Stichprobenebene ein vollständiger Datenmatrix $\boldsymbol{Y}$ der Dimension $n \times p$ mit den Spaltenvektoren $(\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_p)$ und den Elementen y_{ij} mit $i = 1, \ldots, n$ und $j = 1, \ldots, p$ gegeben. Darauf aufbauend wird sich die Notation der Arbeit an Van Buuren (2018) orientieren. Die multivariate Verteilung des Datensatzes $\boldsymbol{Y}$ ist mit den Parametern $\boldsymbol{\theta} \in \Omega_\theta$ durch $f(\boldsymbol{Y}|\boldsymbol{\theta})$ bestimmt. Sind fehlende Werte in $\boldsymbol{Y}$ enthalten, so werden diese mit einer Indikatormatrix $\boldsymbol{R}$ lokalisiert. Für jedes Element von $\boldsymbol{R}$ gilt

$$r_{ij} = \begin{cases} 1 & \text{, falls } y_{ij} \text{ nicht beobachtet wurde,} \\ 0 & \text{, ansonsten} \end{cases} \tag{2.1}$$

und $f(\boldsymbol{Y}|\boldsymbol{\theta}) = f(\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}|\boldsymbol{\theta})$ wird nach beobachteten und nicht-beobachteten Elementen aufgeteilt vgl. (Van Buuren (2018), S.35). Die Verteilung der Datenausfälle $\boldsymbol{R}$ mit den Parametern $\boldsymbol{\psi} \in \Omega_\psi$ kann von $\boldsymbol{Y}_{obs}$ und $\boldsymbol{Y}_{mis}$ abhängen. Diese Beziehung wird als Datenausfallmechanismus bezeichnet und es werden nach Abhängigkeit drei Fälle unterschieden. Lediglich bedingt auf die Parameter $f(\boldsymbol{R}|\boldsymbol{\psi})$ sind die Ausfälle $\boldsymbol{R}$ vollkommen zufällig (MCAR), für eine zusätzliche Abhängigkeit von den beobachteten Daten $f(\boldsymbol{R}|\boldsymbol{Y}_{obs}, \boldsymbol{\psi})$ zufällig (MAR) und im Fall der Abhängigkeit des Ausfalls von dem Ausfallgrund $f(\boldsymbol{R}|\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}, \boldsymbol{\psi})$ nicht zufällig (MNAR) vgl. (Little & Rubin (2019), S.13f.).

Der tatsächlich beobachtete Datensatz kann mit $\boldsymbol{Y}_{obs}$ sowie $\boldsymbol{R}$ beschrieben werden und $f(\boldsymbol{Y}_{obs}, \boldsymbol{R}|\boldsymbol{\theta}, \boldsymbol{\psi})$ entspricht der dazugehörigen gemeinsamen Dichtefunktion. Entgegen der unbekannten Parameter $\boldsymbol{\psi}$, sind die Parameter $\boldsymbol{\theta}$ schätzbar und für die Inferenz von Interesse. Um $\boldsymbol{\theta}$ ohne $\boldsymbol{\psi}$ schätzen zu können, wird die Proportionalität der Likelihood-Funktion von $\boldsymbol{\theta}$ und $\boldsymbol{\psi}$ zur gemeinsamen Dichte ausgenutzt vgl. (Van Buuren (2018), S.38).

$$L_{full}(\boldsymbol{\theta}, \boldsymbol{\psi}|\boldsymbol{Y}_{obs}, \boldsymbol{R}) \propto f(\boldsymbol{Y}_{obs}, \boldsymbol{R}|\boldsymbol{\theta}, \boldsymbol{\psi}) \qquad \text{und} \qquad L_{ign}(\boldsymbol{\theta}|\boldsymbol{Y}_{obs}) \propto f(\boldsymbol{Y}_{obs}|\boldsymbol{\theta}) \tag{2.2}$$

Wie Little & Rubin (2019) zeigen, kann $\boldsymbol{\psi}$ vernachlässigt und die in 2.2 zweitgenannte reduzierte Likelihood-Funktion genutzt werden, falls das Datenausfallmodell MAR ist und die Parameterräume von $\boldsymbol{\theta}$ und $\boldsymbol{\psi}$ unabhängig sind, so dass $\Omega_{\theta,\psi} = \Omega_\theta \times \Omega_\psi$ gilt. Hierbei

fällt die Annahme zufälliger Datenausfälle (MAR) höher ins Gewicht, da keine erklärende Beziehung zwischen $\boldsymbol{\theta}$ und $\boldsymbol{\psi}$ ersichtlich ist vgl. (Schafer (1997), S. 11). Unter diesen Bedingungen kann die in 2.2 erstgenannte Funktion $f(\boldsymbol{Y}_{obs}, \boldsymbol{R}|\boldsymbol{\theta}, \boldsymbol{\psi})$ nun zu

$$f(\boldsymbol{Y}_{obs}|\boldsymbol{\theta})f(\boldsymbol{R}|\boldsymbol{Y}_{obs}, \boldsymbol{\psi}) \propto L_{ign}(\boldsymbol{\theta}|\boldsymbol{Y}_{obs})L(\boldsymbol{\psi}|\boldsymbol{Y}_{obs}, \boldsymbol{R}) \tag{2.3}$$

faktorisiert werden, wobei auf 6.51 und 6.50 aus Little & Rubin (2019) verwiesen sei. Mit 2.3 sind die Likelihood-Quotienten hinsichtlich $\boldsymbol{\theta}$, durch das Wegfallen von $L(\boldsymbol{\psi}|\boldsymbol{Y}_{obs}, \boldsymbol{R})$, für beide Likelihood-Funktionen in 2.2 equivalent und $\boldsymbol{\psi}$ daher vernachlässigbar. Dies gilt auch für die Bayes-Inferenz mit der Unabhängigkeit der Prior Verteilungen $f(\boldsymbol{\theta\psi}) = f(\boldsymbol{\theta})f(\boldsymbol{\psi})$. Die Vernachlässigbarkeit von $\boldsymbol{\psi}$ ist ein zentrales Resultat, da so der Zusammenhang

$$f(\boldsymbol{Y}|\boldsymbol{Y}_{obs}, \boldsymbol{R} = 1) = f(\boldsymbol{Y}|\boldsymbol{Y}_{obs}, \boldsymbol{R} = 0) \tag{2.4}$$

gilt. Dies impliziert, dass die Verteilung der vollständigen Stichprobendaten identisch mit der Verteilung der beobachteten Daten ist und so unabhängig von $\boldsymbol{R}$ sowie $\boldsymbol{\psi}$ vgl. (Van Buuren (2018), S.38f). Damit werden Imputationsmodelle auf den beobachteten Daten spezifiziert, um folglich für die Imputation der fehlenden Datenpunkte genutzt zu werden.

2.1 MICE Algorithmus

Die mit 2.1 identifizierten Datenausfälle $\boldsymbol{R}$ nehmen – nach Anzahl der betroffenen Datenvektoren und der Anzahl der Ausfälle pro Vektor – verschiedene Ausfallmuster an. Abbildung 2.1

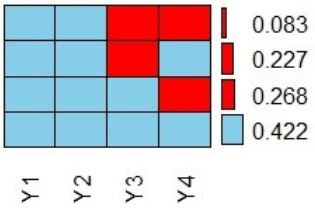

Abbildung 2.1: Generelles Ausfallmuster

beschreibt exemplarisch das in dieser Arbeit simulierte generelle Ausfallsmuster. Demnach lassen sich Anforderungen an den zugrundeliegenden Imputationsalgorithmus stellen. Wie anhand der rot markierten Felder zu erkennen ist, überschneiden sich die Datenausfälle innerhalb der Zeilen. Ist j jeweils einer der zu ergänzende Datenvektoren, so sind bei der Imputation von $\boldsymbol{Y}_j$ durch multivariate Modelle in den Kovariaten $\boldsymbol{Y}_{-j}$ ebenfalls fehlende Werte enthalten. Als Strategie gegen diese Problematik wird auf Multiple Imputation by Chained

Equations (MICE) zurückgegriffen vgl. (Buuren & Groothuis-Oudshoorn (2010), S.6f.). Der Algorithmus basiert auf der Annahme, dass die multivariate Verteilung der Daten $f(\boldsymbol{Y}|\boldsymbol{\theta})$ durch iteratives Ziehen aus den bedingten Verteilungen $f(\boldsymbol{Y}_1|\boldsymbol{Y}_{-1}, \boldsymbol{\theta}_1), \ldots, f(\boldsymbol{Y}_p|\boldsymbol{Y}_{-p}, \boldsymbol{\theta}_p)$ bestimmt werden kann. Folgt $f(\boldsymbol{Y}_j^{mis}|\boldsymbol{Y}_j^{obs}, \boldsymbol{Y}_{-j}, \boldsymbol{R})$ einem multivariaten Imputationsmodell, so is der Ablauf in MICE mit $t = 1, \ldots, T$ Iterationen vgl. (Van Buuren (2018), S.120):

1. wird $\forall j$ eine initiale Imputation $\boldsymbol{Y}_j^0$ durch Ziehen mit Zurücklegen aus $\boldsymbol{Y}_j^{obs}$ bestimmt,
2. ist $\forall j, t$ $\boldsymbol{Y}_j^t$ der zu ergänzende Datenvektor und $\boldsymbol{Y}_{-j}^t = (\boldsymbol{Y}_1^t, \ldots, \boldsymbol{Y}_{j-1}^t, \boldsymbol{Y}_{j+1}^{t-1}, \ldots, \boldsymbol{Y}_p^{t-1})$ der übrige Datensatz aus vollständigen und ergänzten Vektoren,
3. entspricht eine Rotation in j einer Iteration t. $\forall j$ wird aus der Parameterverteilung $\dot{\boldsymbol{\theta}}_j^t \sim f(\boldsymbol{\theta}_j^t|\boldsymbol{Y}_j^{obs}, \boldsymbol{Y}_{-j}^t, \boldsymbol{R})$ gezogen und damit $\dot{\boldsymbol{Y}}_j^t \sim f(\boldsymbol{Y}_j^{mis}|\boldsymbol{Y}_j^{obs}, \boldsymbol{Y}_{-j}^t, \boldsymbol{R}, \dot{\boldsymbol{\theta}}_j^t)$ ergänzt.

Der letzte Schritt wird durch Monte-Carlo-Integration ermöglicht und die nicht direkt bestimmbare posteriore Vorhersageverteilung des multivariaten Imputationsmodells

$$f(\boldsymbol{Y}_j^{mis}|\boldsymbol{Y}_j^{obs}, \boldsymbol{Y}_{-j}, \boldsymbol{R}) = \int_{\Omega_\theta} f(\boldsymbol{\theta}_j|\boldsymbol{Y}_j^{obs}, \boldsymbol{Y}_{-j}, \boldsymbol{R}) f(\boldsymbol{Y}_j^{mis}|\boldsymbol{Y}_j^{obs}, \boldsymbol{Y}_{-j}, \boldsymbol{R}, \boldsymbol{\theta}_j) d\boldsymbol{\theta}_j \tag{2.5}$$

wird in die posteriore Verteilung der beobachteten Daten und die bedingte Vorhersageverteilung der fehlenden Daten aufgeteilt. Die Imputationen der Datenvektoren mit fehlenden Werten bedingen einander über den zweiten Schritt. $\boldsymbol{Y}_j^0$ wird mit jeder Iteration t um neue Werte ersetzt, bis der Prozess nach T Iterationen zu einer stationären Verteilung der Parameter θ_j konvergiert. Wie Abbildung 2.2 verdeutlicht können bereits wenige Iterationen

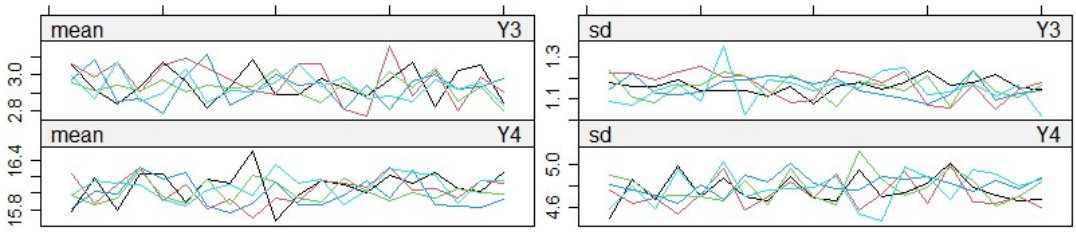

Abbildung 2.2: Konvergenz der Parameter der Simulationsdaten bis $T = 20$

genügen. In der Literatur werden 5 bis 10 Iterationen vorgeschlagen, hier findet sich auch eine ausführliche Diskussion der Bedinungen für Konvergenz vgl. (Van Buuren (2018), S.120ff). Da die Ergänzungen in $t + 1$ nur auf den Werten aus t basieren, entspricht der Prozess einer Markow-Kette. Der MICE-Algorithmus ist somit eine Variante des Gibbs-Samplings nach Geman & Geman (1984) und bildet das Gerüst für alle multivariaten Imputationsverfahren dieser Arbeit. Es wird auf die Implementierung im R-Packet `mice` zurückgegriffen.

2.2 Imputationsverfahren

Insgesamt werden die Mittelwertergänzung, eine einfache Variante des Hot-Deck Verfahrens, die Lineare Regression sowie das Bayessche Lineare Regressionsmodell, das Predictiv Mean Matching und der Regressionsbäum als Imputationsverfahren verwendet. Im Folgenden soll der Fokus auf der formalen Einführung der Verfahren und deren Funktionsweise im Kontext der Datenimputation liegen. Falls genutzt, werden die Verfahren gemäß ihrer Implementierung in `mice` eingeführt.

2.2.1 Singuläre Verfahren

Zunächst sollen die univariaten oder ad-hoc Verfahren vorgestellt werden. Diese sind die Hot-Deck Ergänzung (HD) und die Mittelwertergänzung (ME). HD ist ein Resampling Verfahren. In der hier betrachteten einfachen Ausführung werden, im Imputationsprozess von $\boldsymbol{Y}_j$, keine Verbindungen zwischen Variablen in $\boldsymbol{Y}_{-j}$ berücksichtigt. Sind fehlende Werte in $\boldsymbol{Y}_j$ enthalten, so wird lediglich zufällig mit Zurücklegen aus $\boldsymbol{Y}_j^{obs}$ gezogen und damit wiederholend ergänzt, bis jeder fehlende Wert in $\boldsymbol{Y}_j^{mis}$ vervollständigt ist. Diese Lösung entspricht ebenfalls der initialen Imputation $\boldsymbol{Y}_j^0$ des MICE-Algorithmus.

Für fehlende Werte in $\boldsymbol{Y}_j$ kann auch der Mittelwert der beobachteten Werte $\boldsymbol{Y}_j^{obs}$ zur Imputation verwendet werden. Dies entspricht der ME. Die Ergänzung für alle Elemente in $\boldsymbol{Y}_j^{mis}$ ergibt sich demnach mit der Anzahl der beobachteten Elemente n_{obs} durch

$$\widehat{Y}_j^{mis} = \frac{1}{n_{obs}} \sum_{i=1}^{n_{obs}} y_{ij}^{obs} \tag{2.6}$$

Um die Informationen anderer Variablen für die Imputation zu nutzen, werden multivariate Methoden verwendet. Konventionell bedingt sei der zu ergänzende Vektor $\boldsymbol{Y}_j$ als $\boldsymbol{y}$ bezeichnet und die Realisationen der erklärenden Variablen $\boldsymbol{Y}_{-j}$ in der Designmatrix $\boldsymbol{X}$ gegeben. Jede Beobachtung i aus $\boldsymbol{Y}$ besteht so aus einer der Realisationen der abhängigen Variable $y_i \in \mathcal{R}$ und $\boldsymbol{x}_i = [x_0, x_{i1}, \ldots, x_{ij-1}, x_{ij+1}, \ldots, x_{ip}] \in \mathcal{R}^p$ aus $p-1$ unabhängigen Variablen. Nach der initialen Befüllung durch MICE, teilen sich $\boldsymbol{X}$ und $\boldsymbol{y}$ $\forall i$ – durch zu ergänzende und beobachtete Elemente in $\boldsymbol{y}$ – zeilenweise in $\boldsymbol{X}_{mis}$ und $\boldsymbol{y}_{mis}$ sowie in $\boldsymbol{X}_{obs}$ und $\boldsymbol{y}_{obs}$ auf. Kann $\boldsymbol{y}_{obs}$ durch $\boldsymbol{X}_{obs}$ hinreichend erklärt werden, so stellt die lineare Regression (LR) eine gängige Variante zur Modellierung des Zusammenhangs dar. Das Modell ist linear in den Parametern des Koeffizientenvektors $\boldsymbol{\beta} = [\beta_0, \beta_1, \ldots, \beta_{j-1}, \beta_{j+1}, \ldots, \beta_p]^T$ und enthält den Fehlervektor $\boldsymbol{\varepsilon} = [\varepsilon_1, \ldots, \varepsilon_{n_{obs}}]^T$, der unsystematische Abweichungen beschreibt. Damit ist

ein LR Modell in der allgemein Darstellung mit

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad \text{mit} \quad \varepsilon \sim N(0, \sigma^2) \quad \text{und} \quad \sigma^2 > 0 \tag{2.7}$$

gegeben. Als Imputationsverfahren können zur Modellierung nur zeilenweise vollständige Daten aus $\boldsymbol{Y}$ berücksichitgt werden, weshalb auf $\boldsymbol{y}_{obs}$ und $\boldsymbol{X}_{obs}$ zurückgegriffen wird. Zur Bestimmung des linearen Zusammenhangs wird der Parametervektor mit der Methode der kleinsten Quadrate geschätzt, die den quadrierten Fehlerterm minimiert und

$$\widehat{\boldsymbol{\beta}}_{obs} = \underset{\beta \in \mathcal{R}^p}{\arg\min} \, \|\boldsymbol{y}_{obs} - \boldsymbol{X}_{obs}\boldsymbol{\beta}\|_2^2 \tag{2.8}$$

vgl. (Hastie et al. (2009), S.45f). $\widehat{\boldsymbol{\beta}}_{obs} = (\boldsymbol{X}_{obs}^T\boldsymbol{X}_{obs})^{-1}\boldsymbol{X}_{obs}^T\boldsymbol{y}_{obs}$ entspricht der erhaltenen Lösung aus 2.8 und nach dem Gauß-Markov-Theorem besitzt dieser Schätzer – mit vollem Rang der Matrix $\boldsymbol{X}_{obs}$ und unkorrelierten sowie homoskedastischen Fehlern $\boldsymbol{\varepsilon}$ – unter allen linearen sowie erwartungstreuen Schätzern die kleinste Varianz vgl. (Hastie et al. (2009), S.51). Mit dem Resultat 2.4 kann $\boldsymbol{y}_{mis}$ nun ergänzt werden. Der bestimmte Parametervektor liefert mit den – dem Modell unbekannten – Daten $\boldsymbol{X}_{mis}$ die Imputationen durch

$$\widehat{\boldsymbol{y}}_{mis} = \boldsymbol{X}_{mis}\widehat{\boldsymbol{\beta}}_{obs} \tag{2.9}$$

Für ein gleichbleibendes Datenausfallmuster in Y sind die LR sowie die ME deterministisch. Die Unsicherheit bezüglich der Parameter, die durch den Datenausfall entsteht, wird hier nicht berücksichtigt und Ergänzungen werden letztlich wie tatsächliche Werte behandelt.

2.2.2 Multiple Verfahren

Um Ergänzungen adäquat zu behandeln wird mehrfach ergänzt und – die Unsicherheit zu den Parametern beachtend – Bayes-Methoden genutzt. In diesem Kontext sind Parameter keine unveränderelichen Ergebnisse einer Schätzung der wahren Werte, sondern Zufallsvariablen mit Verteilungen. Diese sind für das LR Modell durch die gemeinsame Verteilung

$$f(\sigma^2, \boldsymbol{\beta}|\boldsymbol{X}_{obs}, \boldsymbol{y}_{obs}) \propto L(\boldsymbol{y}_{obs}|\boldsymbol{X}_{obs}, \sigma^2, \boldsymbol{\beta}) f(\sigma^2, \boldsymbol{\beta}) \quad \text{mit} \quad f(\sigma^2, \boldsymbol{\beta}) = f(\sigma^2) f(\boldsymbol{\beta}|\sigma^2) \tag{2.10}$$

bestimmt. Die Posterior-Verteilung der Parameter ist durch die Likelihood-Funktion sowie einer zu wählende Prior-Verteilung gegeben. Zufallszüge aus der gemeinsamen Verteilung der Parameter sind direkt nicht möglich, allerdings kann 2.10 umgeformt werden. Sei $f(\sigma^2, \boldsymbol{\beta}) \propto \sigma^{-2}$ die Kombination einer flachen Prior-Verteilung $f(\boldsymbol{\beta}|\sigma^2)$ und Jeffreys Prior

$f(\sigma^2)$, so entspricht die gemeinsame Parameterverteilung einer Inversen-χ^2 Verteilung für σ^2 und einer multivariaten Normalverteilung für $\boldsymbol{\beta}$ vgl. (Enders (2010), S.176ff). Es gilt

$$f(\sigma^2, \boldsymbol{\beta}|\boldsymbol{X}_{obs}, \boldsymbol{y}_{obs}) \propto f(\sigma^2|\boldsymbol{X}_{obs}, \boldsymbol{y}_{obs}) f(\boldsymbol{\beta}|\sigma^2, \boldsymbol{X}_{obs}, \boldsymbol{y}_{obs}). \tag{2.11}$$

Wird aus den posterioren Verteilungen der Modellparameter nacheinander $\dot{\sigma}^2$ sowie $\dot{\boldsymbol{\beta}}_{obs}$ gezogen und gegenüber 2.9 durch zufällige Züge aus $\dot{\varepsilon} \sim N(0, \dot{\sigma}^2)$ eine stochastische Komponente hinzugefügt, so ist die Methode mit den Ergebnissen aus 2.8 wie folgt gegeben:

1. $\dot{\sigma}^2$ wird aus $(\boldsymbol{y}_{obs} - \boldsymbol{X}_{obs}\widehat{\boldsymbol{\beta}}_{obs})^T(\boldsymbol{y}_{obs} - \boldsymbol{X}_{obs}\widehat{\boldsymbol{\beta}}_{obs})\chi^{-2}_{(n_{obs}-p)}$ gezogen,
2. mit $\dot{\sigma}^2$ wird $\dot{\boldsymbol{\beta}}_{obs}$ aus $mvN(\widehat{\boldsymbol{\beta}}_{obs}, (\boldsymbol{X}^T_{obs}\boldsymbol{X}_{obs})^{-1}\dot{\sigma}^2)$ gezogen,
3. mit $\dot{\boldsymbol{\varepsilon}}$ sind durch $\dot{\boldsymbol{y}}_{mis}$ aus $N(\boldsymbol{X}_{mis}\dot{\boldsymbol{\beta}}_{obs}, \dot{\sigma}^2)$ die Imputationen $\widehat{\boldsymbol{y}}_{mis}$ gegeben.

Das resultierende Imputationsmodell folgt der in `mice` implementierten Bayesschen Linearen Regression (BLR), dabei wird die Erweiterung auf eine Bayesschen Ridge Regression vernachlässigt vgl. (Van Buuren (2018), S.68). Wird der letzte Schritt umformuliert, folgt

$$\widehat{\boldsymbol{y}}_{mis} = \boldsymbol{X}_{mis}\dot{\boldsymbol{\beta}}_{obs} + \dot{\boldsymbol{\varepsilon}} \quad \text{mit} \quad \dot{\varepsilon} \sim N(0, \dot{\sigma}^2) \tag{2.12}$$

für die Ergänzungen. Aufgrund der zufälligen Komponenten führt 2.12 in jeder Schätzung zu untschiedlichen Ergänzungen und die Unsicherheit hinsichlich der Parameter, die wegen der reduzierten Datenbasis aus $\boldsymbol{y}_{obs}$ und $\boldsymbol{X}_{obs}$ in 2.8 entsteht, wird abgebildet.

Die vorgestellten Regressionsmodelle basieren auf den restriktiven Annahmen nach Gauß-Markov aus 2.7 und 2.8. Sind diese Annahmen durch die Datenlage nicht erfüllt, kann dies zu fehlerhaften Imputationen durch Misspezifikationen führen. Um trotzdem valide Ergänzungen zu generieren, können robustere Methoden verwendet werden, die eine geringere Parametrisierung aufweisen. Die semiparametrische Methode Predictiv Mean Matching (PMM) ist hierzu das Standardverfahren in `mice`. Allgemein nutzt das Verfahren eine initiale Ergänzung $\dot{\boldsymbol{y}}_{mis}$ um ein Abstandsmaß η zu minimieren und so Werte aus $\widehat{\boldsymbol{y}}_{obs} = \boldsymbol{X}_{obs}\widehat{\boldsymbol{\beta}}_{obs}$ zu ziehen. So ist das PMM eine Hot-Deck-Methode, die – gegenüber der einfachen Variante – multivariate Zusammenhänge berücksichtigt und mit dem Abstandsmaß nur realistische Werte aus $\widehat{\boldsymbol{y}}_{obs}$ idenifiziert. Das Vorgehen in `mice` entspricht vgl. (Van Buuren (2018), S.81):

1. das Abstandsmaß $\dot{\eta}(i,l) = |\boldsymbol{X}_{obs}\widehat{\boldsymbol{\beta}}_{obs} - \boldsymbol{X}_{mis}\dot{\boldsymbol{\beta}}_{obs}| = |\widehat{\boldsymbol{y}}^{obs}_i - \widehat{\boldsymbol{y}}^{mis}_l|$ wird mit $\dot{\boldsymbol{\beta}}_{obs}$ aus 2.12 sowie $\widehat{\boldsymbol{\beta}}_{obs}$ nach 2.8 und $i = 1, \ldots, n_{obs}$ sowie $l = 1, \ldots, n_{mis}$ bestimmt,
2. $\forall l$ werden die d Werte aus n_{obs} bestimmt die $\sum_d \dot{\eta}(i,l)$ minimieren und $\forall l$ wird zufällig ein Wert $\dot{\boldsymbol{y}}^{obs}_{i_l}$ aus diesen d Werten gezogen. Dieser entspricht der Ergänzung $\widehat{\boldsymbol{y}}^{mis}_l$.

Die Berechnungschritte sind damit per se deterministisch, allerdings verhält sich PMM durch $\dot{\boldsymbol{\beta}}_{obs}$ aus 2.12 und dem Zufallszug $\dot{\boldsymbol{y}}_{i_l}^{obs}$ probabilistisch. Die Anzahl der berücksichitgten Werte d liegt bei der `mice` bei 5 und beschreibt einen Kompromiss aus Ergebnissen empirischer Untersuchungen vgl. (Van Buuren (2018), S.82).

Auch nichtparametrische Regressionsbäume (CART) nach Breiman et al. (1984) können robuste Imputationsergebnisse bei Verletzung der Annahmen aus 2.8 und 2.7 liefern. Das Modell wird mittels rekursiver Partitionierung der Daten $\boldsymbol{y}_{obs}$ durch binäre Aufteilungen der $\boldsymbol{X}_{obs}$ geschätzt. Unter allen erklärenden Datenvektoren in $\boldsymbol{X}_{obs}$ wird jener mit dem

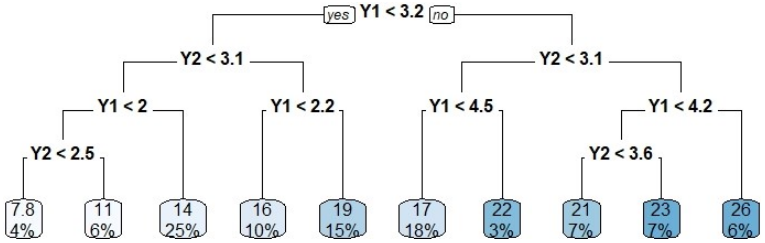

Abbildung 2.3: Exemplarische Darstellung von CART

optimalen Schwellenwert gewählt, sodass die Varianz hinsichtlich der zu erklärenden Daten $\boldsymbol{y}_{obs}$ in den zwei entstehenden Untergruppen minimiert wird vgl. (Hastie et al. (2009), S.307). Dementsprechend wird $\boldsymbol{y}_{obs}$ aufgeteilt. Wie in Abbildung 2.3 gezeigt, wird dies auf jeder Seite des Regressionsbaumes fortgeführt solange in allen Endknoten eine Mindestanzahl an Werten vorhanden ist und damit entstehen immer homogenere Untergruppen aus $\boldsymbol{y}_{obs}$. Die Imputation in `mice` ist durch den folgenden Ablauf gegeben vgl. (Doove et al. (2014), S.95):

1. mit den Daten $\boldsymbol{X}_{obs}$ und $\boldsymbol{y}_{obs}$ wird ein Regressionsbaum $f(\boldsymbol{X}_{obs})$ mit k Endknoten g_k modelliert, die d_k Werte aus $\boldsymbol{y}_{obs}$ enthalten und $\boldsymbol{y}_{obs}$ so disjunkt partitionieren,

2. $\forall l = 1, \ldots, n_{mis}$ wird mit $f(\boldsymbol{X}_{mis})$ der entsprechende Endknoten g_{lk} des Regressionsbaums für jedes Element der zu ergänzenden Daten $\boldsymbol{y}_{mis}$ bestimmt,

3. für $l = 1, \ldots, n_{mis}$ wird aus den über $\boldsymbol{X}_{mis}$ identifizierten Endknoten g_{lk} zufällig ein Wert $\dot{y}_{i_l}^{obs}$ aus d_k gezogen. Dieser entspricht der Imputation $\hat{y}_l^{mis}$.

Damit emuliert auch CART ein probabilistisches Modell mittels zufälligen Ziehen aus einer deterministischen Lösung.

Neben den multipen Imputationen durch BLR, PMM und CART, werden in Kapitel 3 auch singuläre Varianten der Verfahren untersucht. Dies soll eine unmittelbare Gegenüberstellung der singulären und multiplen Imputation ermöglichen.

3 Methodik

Um die eingangs aufgeworfene Fragestellung zu beantworten, wurde eine Simulationsstudie durchgeführt. Das Vorgehen soll in diesem Kapitel genauer erläutert werden. Dabei liegt der Fokus zunächst auf der Simulation von stetigen Daten, auf denen fehlende Werte generiert werden. Daran anschließend folgt für alle Imputationsverfahren I die Ergänzung wie in Kapitel 2 beschrieben. Darauf aufbauend wird die Schätzung der Populationsparameter von Interesse Q thematisiert. Letztlich werden die Vergleichskriterien eingeführt, welche die Schätzungen auf den ergänzten Daten jedes Imputationsverfahrens I hinsichtlich valider Inferenz überprüfen.

3.1 Datensimulation

Jeder Datensatz $\boldsymbol{Y}^{(z)} = (\boldsymbol{Y}_1^{(z)}, \ldots, \boldsymbol{Y}_p^{(z)})$ mit $p = 4$ und den Iterationen $z = 1, \ldots, Z$ der Simulation, wurde durch $n = 1000$ Zufallszüge aus stetigen Variablen der Population $\mathcal{Y}$ generiert. Dabei wurden Abhängigkeiten zwischen den Variablen unterstellt. Die Verteilungen der Variablen sind durch $X_1 \sim N(3, 1)$ und $X_2 \sim U(2, 4)$ sowie mit

$$X_3 = 0.75X_1 + 0.1X_2 + N(0, 1) \quad \text{und} \quad X_4 = 3X_1 + 0.75X_2^2 + 0.5X_3 + (\chi^2_{(3)} - 3) \quad (3.1)$$

gegeben. Über die Interaktionen in 3.1 stehen die aus X_4 gezogenen Daten derartig in Beziehung zu den Realisationen der übrigen Variablen, dass die Schätzung einer LR zweckmäßig ist. Der Diskussion robuster Imputationsverfahren in Unterpunkt 2.2.2 folgend, werden Misspezifikationen zu 2.7 und 2.8 eingeführt. So ist X_2 und der Fehlerterm von X_4 nicht normalverteilt, X_3 induziert Multikollinearität durch die Abhänigkeit von X_1 sowie X_2, und X_4 ist nicht linear abhängig von X_2. Dies soll die perfekte Anpassung linearer Imputationsverfahren verhindern und eine nicht ideale Struktur empirischer Daten imitieren.

Auf jedem Datensatz $\boldsymbol{Y}^{(z)}$ werden fehlende Werte erzeugt. Da MCAR in der Anwendung als nicht realistisch angesehen wird und alle Verfahren auf Ignorierbarkeit nach 2.4 beruhen, wird der Datenausfallmechanismus MAR simuliert vgl. (Van Buuren (2018), S.8). Dies betrifft die Realisationen der Variablen X_3 und X_4. Für die Modellierung der Datenausfälle in

$\boldsymbol{Y}_3$ und $\boldsymbol{Y}_4$ wird zu jedem Wert y_{i3} und y_{i4} mit $i = 1, \ldots, n$ über die logistischen Funktionen

$$P(\boldsymbol{r}_3 = 1) = \frac{exp(\boldsymbol{Y}_1 + \boldsymbol{Y}_2 - 7)}{1 + exp(\boldsymbol{Y}_1 + \boldsymbol{Y}_2 - 7)} \quad \text{und} \quad P(\boldsymbol{r}_4 = 1) = \frac{exp(0.725\boldsymbol{Y}_2 - \boldsymbol{Y}_1)}{1 + exp(0.725\boldsymbol{Y}_2 - \boldsymbol{Y}_1)} \tag{3.2}$$

die Wahrscheinlichkeit des Datenausfalls p_{i3} bzw. p_{i4} mit $0 \leq p_{ij} \leq 1$ bestimmt vgl. (Van Buuren (2018), S.70f). Diese sind durch 3.2 mit den Vektoren $\boldsymbol{p}_3$ und $\boldsymbol{p}_4$ gegeben. Nun wird für alle Elemente $i = 1, \ldots, n$ der Vektoren die Realisationen der Bernoulliversuche mit $Ber(p_{i3})$ bzw. $Ber(p_{i4})$ bestimmt und nacheinander in die $n \times 1$ Indikatorvektoren $\boldsymbol{r}_3$ sowie $\boldsymbol{r}_4$ übertragen. Ist ein Element der resultierenden Indikatormatrix R mit $r_{ij} = 1$ als ausgefallen markiert, so wird der entsprechende Wert y_{ij} aus dem Datensatz $\boldsymbol{Y}^{(z)}$ entfernt. Wie in Abbildung 3.1 angedeutet, betrifft dies im Mittel 31% der Werte in $\boldsymbol{Y}_3$ und 34%

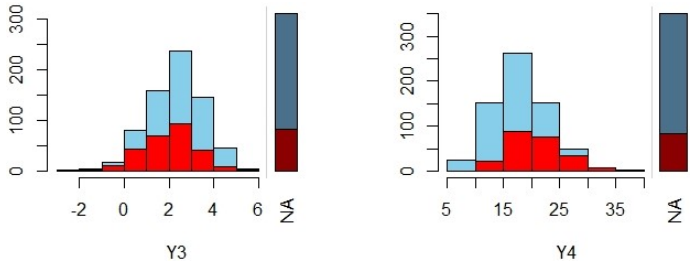

Abbildung 3.1: Datenausfälle auf einem Datensatz $\boldsymbol{Y}^{(z)}$

der Werte in $\boldsymbol{Y}_4$. In 3.2 wurde beachtet, nicht nur normalverteilte Daten zu nutzen sowie ein erkennbares Ausfallproblem zu generieren. Per Design ist der Datenausfallmechanismus von $\boldsymbol{Y}_{obs}$ abhängig, aber unabhängig von $\boldsymbol{Y}_{mis}$ und somit MAR. Damit ist Ignorierbarkeit im Sinne von 2.4 gegeben und die Datensätze können mit den Imputationensverfahren aus Unterpunkt 2.2 ergänzt werden. Ein simulierter Datensatz mit Datenausfällen ist in der Iteration z für alle Imputationsverfahren identisch. Zwischen den Iterationen z und $z + 1$ bestehen jedoch Unterschiede durch die zufällige Realisation der Daten sowie der Datenausfälle. Diese zufällige Variation kann, neben der Handhabe der Imputationverfahren I mit dem Datenausfallmechanismus, auch ursächlich für die Ergebnisse sein vgl. (Van Buuren (2018), S.51f). Dieser Simulationsmechanismus kann nicht vollends ausgeschlossen werden, allerdings konvergieren die Ergebnisse durch Erhöhung der Iterationen Z zu den tatsächlichen bezüglich des Datenausfallmechanismus vgl. 3.8. Dazu wird auch mehr Rechenkapazität benötigt. Abwägend werden $Z = 1000$ Iterationen für die vorliergende Fragestellung als hinreichend angesehen. Demnach werden 1000 unvollständige Datensätze generiert. Durch die singuläre und muliple Berücksichtigung von BLR, PMM und CART finden daraufhin neun Imuptationsverfahren I Anwendung. Multivariate Verfahren nutzen alle verfügbaren Kovariaten sowie `mice` gemäß Kapitel 2.1 mit $T = 5$ Iterationen vgl. Abbildung 2.2.

3.2 Parameterschätzung

Die Parameter der Population $\mathcal{Y}$ können in den ergänzten Datensätzen der Iteration z geschätzt werden. Als zu schätzende Populationsparameter – welche die Grundlage des Vergleichs bilden – werden μ_4 aus dem Erwartungswert der Variable X_4 und die linearen Zusammenhangsmaße β_2 sowie β_3 einer LR für den bedingten Erwartungswert $E[X_4|X_{-4}]$ herangezogen. Diese werden stellvertretend als Populationsparameter von Interesse mit Q beschrieben. Die Regressionskoeffizienten wurden hier wegen der zentralen Bedeutung der LR zur Modellierung von Abhängigkeiten zwischen Variablen eines wissenschaftlichen Systems durchgeführt. Es ist für die diesbezügliche statistische Inferenz das wohl am meisten verwendete Modell vgl. (Myers & Myers (1990), S.1). Eine exakte Schätzung der Parameter und die damit verbundene Inferenz soll daher auch mit einer unvollständigen Datenlage valide umsetzbar sein.

Die Schätzungen der Populationsparameter und deren Varianzen erfolgt für jeden der Z Simulationsdurchläufe. Die simulierten unvollständigen Daten werden in z durch s singuläre und m multiple Imputationsverfahren $s, m \in I$ vervollständigt. Damit werden die Populationsparameter geschätzt. Die genutzten Schätzer sind $\forall z, I$ das arithmetische Mittel

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_{i4} \quad \text{und} \quad \widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathcal{R}^4}{\arg\min} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 \tag{3.3}$$

die Methode der kleinsten Quadrate. Dabei entspricht $\boldsymbol{y} = \boldsymbol{Y}_4$ und mit dem Einsvektor $\boldsymbol{Y}_0$ folgt $\boldsymbol{X} = (\boldsymbol{Y}_0, \boldsymbol{Y}_1, \boldsymbol{Y}_2, \boldsymbol{Y}_3)$. Die Varianz der Parameterschätzungen wird $\forall z, I$ durch

$$V[\widehat{\mu}] = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_{i4} - \widehat{\mu})^2}{n} \quad \text{und} \quad V[\widehat{\beta}_j] = \frac{\widehat{\boldsymbol{\varepsilon}}^T\widehat{\boldsymbol{\varepsilon}}}{n-4}(\boldsymbol{X}^T\boldsymbol{X})_{jj} \tag{3.4}$$

bestimmt vgl. (Hastie et al. (2009), S.47). $\forall z, s$ müssen die Schätzungen 3.3 auf den singulär ergänzten Datensätzen einmalig durchgeführt werden. Diese seien in Anhängigkeit von z sowie s stellvertetend als Parameterschätzung von Interesse $\widehat{Q}_{z,s}$ mit der Varianz $V[\widehat{Q}_{z,s}]$ aus 3.4 beschrieben.

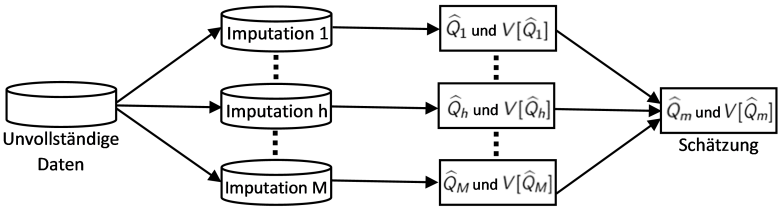

Abbildung 3.2: Parameterschätzung für multiple Imputationen

Durch die multiplen Imputationen der Modelle m aus Kapitel 2.2.2 entstehen M ergänzte Datensätze aus einem Simulationslauf z, die in der Schätzung von Q zweckmäßig zusammengeführt werden müssen vgl. Abbildung 3.2. Dies geschieht mit Hilfe von Kombinationsregeln vgl. (Rubin (2009), S.81ff). Sind m die multiplen Imputationsverfahren aus I und M die Anzahl der Imputationen, so gilt nach Abbildung 3.2 $\forall z, m$ in der Parameterschätzung

$$\widehat{Q}_m = \frac{1}{M} \sum_{h=1}^{M} \widehat{Q}_h \tag{3.5}$$

wobei $\widehat{Q}_h$ den Schätzungen 3.3 für die ergänzten Datensätze $h = 1, \ldots, M$ entspricht. Die Varianz von $\widehat{Q}_m$ kann mit dem Satz von der totalen Varianz in zwei Komponenten

$$\bar{U} = \frac{1}{M} \sum_{h=1}^{M} V[\widehat{Q}_h] \qquad \text{und} \qquad B = \frac{1}{M-1} \sum_{h=1}^{M} (\widehat{Q}_h - \widehat{Q}_m)^2 \tag{3.6}$$

zerlegt werden. Die erste Komponente ist der Durchschnitt der auf den M ergänzten Datensätzen bestimmten Varianzen $V[\widehat{Q}_h]$ und wird mit $\bar{U}$ beschrieben. Die zweite Komponente in 3.6 sei mit B bezeichnet und entspricht der Varianz zwischen den Schätzungen $\widehat{Q}_h$ und $\widehat{Q}_m$ aus 2.6. Mit dem letztgenannten basiert B selbst auf einer Schätzung, was für die Darstellung der Varianz berücksichtigt werden muss vgl. (Van Buuren (2018), S.43). Somit entspricht die totale Varianz der Schätzer für multiple Imputationsverfahren

$$V[\widehat{Q}_m] = \bar{U} + (1 + \frac{1}{M})B \tag{3.7}$$

wobei der Anteil $\frac{B}{M}$ auf die Schätzung von $\widehat{Q}_m$ entfällt. Dieser ist ein Berichtigungsfaktor, der die Validität der Ergebnisse für eine geringe Anzahl von multiplen Imputationen M sicherstellen soll vgl. (Van Buuren (2018), S.58). $\forall z, m$ werden die Schätzungen um Rubin's Rules vgl. 3.5 - 3.7 erweitert und sind so mit $\widehat{Q}_{z,m}$ sowie $V[\widehat{Q}_{z,m}]$ gegeben.

Die wahren Werte Q sind, insbesondere aufgrund des nicht linearen Zusammenhangs zwischen X_2 und X_4 in 3.1, nicht direkt aus dem datengenerierenden Prozess ersichtlich. Daher werden diese mit Hilfe des schwachen Gesetzes der großen Zahlen nach Khinchine (1929) näherungsweise durch Simulation bestimmt. Auf K simulierten vollständigen Stichproben gemäß 3.1 ist der Parameter $\widehat{Q}_k$ selbst eine identische und unabhängig verteilte Zufallsgröße mit existierendem Erwartungswert $E[\widehat{Q}_k] = Q$ vgl. 3.9. Damit gilt

$$\frac{\widehat{Q}_1 + \widehat{Q}_2 + \cdots + \widehat{Q}_K}{K} \xrightarrow{p} Q \tag{3.8}$$

und das arithmetische Mittel $\widehat{Q}_k$ über K iterationen konvergiert in Wahrscheinlichkeit zum wahren Wert Q. Die Anzahl der Simulationen wurde hier auf $K = 100000$ festgelegt.

Als Maßstab der Schätzungen auf den ergänzten Daten der I Imputationsverfahren dienen die Schätzungen 3.3 und 3.4 auf den vollständigen Stichproben (BD) vor Entfernen von Werten durch Datenausfallmechanismen. Diese sind unabhängig von I durch $\widehat{Q}_z$ und $V[\widehat{Q}_z]$ gegeben. Um festzustellen, inwieweit ein Imputationsverfahren zu besseren Inferenzergebnissen führt, werden Vergleichskriterien benötigt. Diese gelten für $\widehat{Q}_{z,s}$ sowie $\widehat{Q}_{z,m}$ equivalent. Daher wird eine gemeinsame Darstellung durch $\widehat{Q}_{z,I}$ genutzt.

3.3 Vergleichskriterien

Die Schätzungen $\widehat{Q}_{z,I}$ auf den ergänzten Daten sollen durch die Imputationsverfahren, gegenüber der Schätzungen auf den vollständigen Stichproben $\widehat{Q}_z$, für die Inferenz wichtige Eigenschaften beibehalten. Die damit einhergehende Validität von Inferenz mit vervollständigten Daten lässt sich nach Rubin (2009) formalisieren. Ist $V[\widehat{Q}_z]$ die Varianz bezüglich der vollständigen Stichprobe, $U = V[\widehat{Q}_{z,I}]$ die Varianz auf einem ergänzten Datensatz und $E[U_{z,I}]$ der Erwartungswert der Varianz über alle möglichen ergänzten Datensätze, so ist valide Inferenz hinsichtlich der Parameter Q mit den Eigenschaften

$$E[\widehat{Q}_z] = Q \quad \text{und} \quad E[U_{z,I}] \geq V[\widehat{Q}_z] \tag{3.9}$$

gegeben. Diese sind die Erwartungstreue, das heißt der Erwartungswert des Schätzers entspricht dem wahren Parameter und die Validität der Konfidenzintervalle vgl. (Van Buuren (2018), S.41). Die zweitgenannte Eigenschaft in 3.9 fordert eine hinreichend große Varianz der geschätzten Parameter $\widehat{Q}_{z,I}$ in den vervollständigten Datensätzen, so dass Konfidenzintervalle nicht zu schmal werden bzw. durch statistische Tests nicht fehlführend und verfrüht auf ein Verwerfen der Hypothese H_0 geschlossen wird.

Aus der Erwartungstreue nach 3.9 ergibt sich direkt ein Kriterum für den Vergleich der Imputationsverfahren. Ein Verfahren I ist einem anderen vorzuziehen, wenn der Erwartungswert der Schätzer $\widehat{Q}_{z,I}$ – über die Z ergänzten Datensätze aggregiert – eine geringere Verzerrung bezüglich der wahren Werte Q aufweist.

$$RBias_I = \frac{\frac{1}{Z}\sum_{z=1}^{Z}(\widehat{Q}_{z,I} - Q)}{Q} \tag{3.10}$$

Dieser Vergleich wird – für eine einheitliche Skala – mit dem relativen Fehler durchgeführt.

Um die Validität der Konfidenzintervalle der Schätzer $\widehat{Q}_{z,I}$ nach 3.9 im Modellvergleich zu berücksichtigen, wird als Kriterium auf die Deckungsrate

$$CIrate_I = \frac{1}{Z}\sum_{z=1}^{Z} \mathbb{1}_{\{Q \in CI(\alpha, \widehat{Q}_{z,I})\}} \tag{3.11}$$

zurückgegriffen. Hierbei wird α auf das gängige Signifikanzniveau von 5% festgelegt. Damit ist – auf den durch I Verfahren ergänzten Daten – für valide Inferenz zu erwarten, dass die wahren Werte Q in $100(1-\alpha)\%$ der Fälle innerhalb der Konfidenzintervalle der Schätzungen $\widehat{Q}_{z,I}$ liegen. Somit ist ein Imputationsverfahren einem anderen vorzuziehen, falls eine Deckungsrate näher an dem angestrebten Niveau von 95% liegt. Für die Konfidenzintervalle wird approximativ eine t-Verteilung t_v mit v Freiheitsgraden unterstellt, da die Varianzen nach 3.4 bzw. 3.7 geschätzt werden müssen. Somit gilt

$$CI(\alpha, \widehat{Q}_{z,I}) = \widehat{Q}_{z,I} \pm t_{v,1-\frac{\alpha}{2}} \sqrt{V[\widehat{Q}_{z,I}]} \quad \text{mit} \quad \frac{Q - \widehat{Q}_{z,I}}{\sqrt{V[\widehat{Q}_{z,I}]}} \sim t_v \tag{3.12}$$

wobei die Freiheitsgrade für singuläre Imputationsverfahren mit $v = n - 1$ gegeben sind. Bei multiplen Imputationsverfahren werden die Freiheitsgrade adjustiert. Da die Anzahl der Beobachtungen mit $n = 1000$ hinreichend groß ist, kann mit der Imputationensanzahl M

$$v_m = (M-1)\left(1 + \frac{1}{r^2}\right) \quad \text{mit} \quad r = \frac{B + \frac{B}{M}}{\bar{U}} \tag{3.13}$$

genutzt werden vgl. (Raghunathan (2015), S.78f). Die Varianzkomponenten B und $\bar{U}$ enstammen dabei den Gleichungen aus 3.6. Dementsprechend gilt für multiple Imputationsverfahren $v = v_m$ in 3.12 und folgend in 3.14.

Haben zwei Imputationsverfahren hinsichtlich 3.10 und 3.11 ähnlich günstige Ausprägungen, so ist jenes Verfahren effizienter, das eine geringere Weite der Konfidenzintervalle aufweist. Über alle Z Simulationen aggregiert, ist dieses Kriterium mit

$$CIwidth_I = \frac{1}{Z}\sum_{z=1}^{Z} 2\left(t_{v,1-\frac{\alpha}{2}} \sqrt{V[\hat{Q}_{z,I}]}\right) \tag{3.14}$$

gegeben und im wesentlichen – wie bereits von der zweitgenannten Eigenschaft für valide Inferenz in 3.9 gefordert – von der Varianz der Schätzer abhängig.

4 Ergebnisse

In diesem Kapitel sollen die Ergebnisse der Simulationsstudie beschrieben und im Kontext der Fragestellung diskutiert werden. Anhand der Vergleichskriterien aus Unterpunkt 3.3 erfolgt zunächst eine Gegenüberstellung der Imputationsverfahren bezüglich der Schätzung des Erwartungswertes von X_4. An BD in 4.1 sowie 4.2 orientierend, liefert ein Imputationsverfahren heuristisch valide Inferenz, wenn der relative Fehler $\pm 0,01$ um 0 und $\pm 1\%$ um die anzustrebende Deckungsrate von 95% liegt. Daraufhin werden die Imputationsverfahren hinsichtlich der Schätzung der Koeffizienten β_2 und β_3 verglichen. Abschließend werden die Ergebnisse zusammengefasst und anhand der Theorie verallgemeinert.

4.1 Erwartungswert

Tabelle 4.1 zeigt die Ergebnisse hinsichltich der Schätzungen von μ_4. Die Verfahren werden nach singulärer und mehrfacher Imputation unterteilt dargestellt. Abgesehen von der ME

		singulär						mehrfach		
Kriterium	BD	ME	HD	LR	BLR	PMM	CART	BLR	PMM	CART
$RBias$	0,00	0,04	0,04	0,00	0,00	0,00	0,00	0,00	0,00	0,00
$V[\widehat{\mu}_4]$	0,025	0,016	0,024	0,023	0,025	0,024	0,024	0,030	0,029	0,026
$CIrate$	95,6	1,9	5,1	91,3	91,1	90,2	89,3	95,2	94,6	90,7
$CIwidth$	0,616	0,492	0,606	0,590	0,615	0,610	0,609	0,680	0,673	0,637

Tabelle 4.1: Ergebnisse zu μ_4

und dem HD sind alle Verfahren bis auf die zweite Nachkommastelle unverzerrt. Abbildung 3.1 verdeutlicht, anhand der Verschiebungen in der Verteilung der ausgefallenen Daten durch den Ausfallmechanismus MAR, wie die Ergänzungen univariater Verfahren hier zu relativen Fehlern führen. Beziehungen zu anderen Variablen werden vernachlässigt, aber wie 3.2 zeigt, liegt diese mit dem Ausfallmechanismus MAR vor. Daher könnten hier lediglich Ausfälle MCAR für die ME und HD zu unverzerrten Ergebnissen führen. Zudem kann die ME die Varianz der Daten nicht abbilden, da jede Ergänzung per Definition – vgl. Zähler in 3.4 – einem Null-Varianz-Wert entspricht. Die Varianzen der übrigen singulären Verfahren liegen nahe an der BD Varianz. Wie für valide Inferenz in 3.9 gefordert, liefern die multiplen

Imputaionsverfahren BLR und PMM eine größere Varianz als BD. Multiple Imputation berücksichtigt die mit den unvollständigen Daten verbundene Unsicherheit explizit, was durch B in 3.7 und den Varianzen der Tabelle 4.1 widergespiegelt wird. Mit der Varianz als treibende Größe – vgl. 3.14 – sind proportional die Weiten der Konfidenzintervalle gegeben. Demensprechend sind die Konfidenzintervalle der multipelen Imputationsverfahren breiter als für den Maßstab BD. Zusammen mit der Unverzerrtheit werden somit für BLR und PMM Deckungsraten erzielt, die hinreichend nah an dem tatsächlichen Wert von 95% liegen. Für die multiple Variante von CART ist dies nicht der Fall. Das kann gegenüber PMM und BLR mit der verringerten Berücksichtigung der Unsicherheit innerhalb der Methode zusammenhängen. Eine Erweiterung durch initiales Bootstrapping bei jeder Imputation könnte hier zu einer Verbesserung führen vgl. (Van Buuren (2018), S.86). Unter allen singulären Imputationsverfahren sind die Deckungsraten zu niedrig. Durch die Verzerrung liegen die ME und das HD hier im einstelligen Bereich und für die übrigen singulären Verfahren reicht die berücksichtigte Unsicherheit innerhalb der Methoden alleine nicht aus, um nah genug an die Deckungsrate von 95% zu gelangen. Damit liefern die multiplen Imputationsverfahren BLR und PMM ergänzte Datensätze, die für Schätzungen von $E[X_4]$ zu validier Inferenz führen, wobei PMM durch schmalere Konfidenzintervalle effizienter ist.

4.2 Regressionskoeffizienten

Für die Regressionskoeffizienten sind die Ergebnisse in Tabelle 4.2 gegeben. Der obere Teil der Tabelle bezieht sich auf die Vergleichskriterien der Schätzungen von β_2 und der untere hingegen auf die Ergebnisse zu β_3. Anstelle der Varianz wird konventionell der Standardfehler

		singulär						mehrfach		
Kriterium	BD	ME	HD	LR	BLR	PMM	CART	BLR	PMM	CART
$RBias$	0,00	-0,30	-0,30	-0,02	-0,01	-0,02	0,00	-0,01	-0,02	0,00
$SE(\widehat{\beta}_2)$	0,136	0,158	0,222	0,106	0,135	0,136	0,139	0,176	0,174	0,164
$CIrate$	94,3	0,0	0,0	71,2	81,1	79,3	79,8	93,4	92,3	87,3
$CIwidth$	0.532	0.622	0.873	0.417	0.531	0.532	0.545	0.732	0.722	0.663
$RBias$	0,00	-0,86	-0,91	0,87	-0,01	-0,07	-0,18	-0,01	-0,07	-0,18
$SE(\widehat{\beta}_3)$	0,078	0,101	0,113	0,072	0,078	0,078	0,080	0,121	0,117	0,110
$CIrate$	94,7	1,5	1,8	7,6	67,5	69,4	63,3	94,4	93,2	84,6
$CIwidth$	0,307	0.396	0.445	0.283	0.306	0.307	0.314	0.533	0.513	0.466

Tabelle 4.2: Ergebnisse zu β_2 und β_3

$SE(\widehat{\beta}_j) = \sqrt{V[\widehat{\beta}_j]}$ betrachtet. Bezüglich β_2 wurde der Datenvektor der abhängigen Variable

ergänzt. Die Verzerrungen der multiplen Verfahren zu der jeweiligen singulären Entsprechung sind identisch und fallen damit für LR, BLR, PMM und CART klein aus. Die ME und das HD produzieren allerdings einen relativen Fehler von $-0,3$. Dazu sei beachtet, dass β_2 als $\frac{Cov(X_2,X_4)}{V[X_2]}$ dargestellt werden kann und die ME Null-Varianz-Werte in $\boldsymbol{Y}_4$ ergänzt, womit gemäß der Definition der empirischen Kovarianz die Schätzung von $Cov(X_2, X_4)$ ebenfalls zu klein ist. Dagegen ist das HD in der einfachen Form generell nicht in der Lage, Zusammenhänge zwischen Variablen zu rekonstruieren, wodurch die Kovarianz abnimmt. Der Standardfehler der beiden Verfahren ist – wie für multiple Imputationsverfahren – größer als BD und für die LR kleiner als BD. Die übrigen singulären Verfahren liefern hier in etwa den gleichen Standardfehler wie BD. Demzufolge realisieren sich proportional die Weiten der Konfidenzintervalle. Für die Deckungsrate von ME und HD dominiert die Verzerrung der Schätzung, was zu einer Deckungsrate von 0 führt. BLR und PMM erreichen hier das beste Ergebnis, wobei BLR um rund 1% näher am angestrebten Wert von 95% liegt. Die übrigen Verfahren weisen eine zu niedrige Deckungsrate auf. Letztlich liefert keines der Verfahren bezüglich β_2 valide Inferenz, wobei die multiplen Imputationsverfahren PMM und BLR den angestrebten Werten am nähsten kommen.

In der unteren Hälfte der Tabelle 4.2 sind die Ergebnisse zu β_3 gegeben. Gegenüber β_2 sind hier die Datenvektoren der abhängigen und der unabhängigen Variable von Datenausfällen betroffen. Dies lässt ein anspruchsvollere Ausfallproblematik entstehen. Dem folgend steigt der relative Fehler für die ME und das HD durch die eben genannte Systematik weiter an. Durch die höhere Ausfallproblematik, überwiegt die Reduktion in der Schätzung von $Cov(X_3, X_4)$ der in $V[X_3]$ für die ME. Auch die LR liefert hier eine starke Verzerrung, die entgegen der anderen Verfahren positiv ausfällt. Abbildung 4.1 stellt die ergänzten

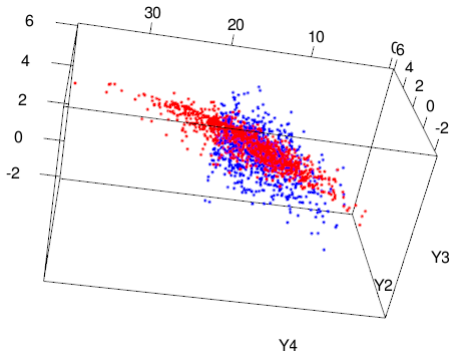

Abbildung 4.1: LR Ergänzungen (an der Ordinate gespiegelt)

Daten (rot) den tatsächlichen Daten (blau) gegenüber. Da $\boldsymbol{Y}_4$ in der Imputation von $\boldsymbol{Y}_3$ berücksichtigt wurde und vice versa, ist durch die Ergänzungen in beiden Datenvektoren ein positiv korrelierter und zunehmend linerer Zusammenhang entstanden, der in der

anschließenden LR Schätzung in der Schätzung von $Cov(X_3, X_4)$ aufgegriffen wird. Somit steigt $\widehat{\beta}_3$ übermäßig an. Die Verzerrungen zu BLR, PMM und CART sind in beiden Varianten identisch. Die Imputationsverfahren CART und PMM führen zu einem relativen Fehler der nicht zu valider Inferenz führt, wohingehen BLR mit $-0,01$ insgesamt die kleinste Verzerrung aufweist. Bezogen auf den Standardfehler des Maßstabs BD mit $0,078$ liegen für alle Imputationsverfahren die Ausprägungen der Standardfehler wie für β_2 um BD. Proportional folgt die Breite der Konfidenzintervalle. Für ME, HD und LR werden nur einstellige Deckungsraten erreicht, da die Verzerrungen zu hoch ins Gewicht fallen. Die übrigen singulären Imputationsverfahren erzielen durch niedrigere Deckungsraten schlechtere Ergebnisse als für β_2, was der erhöhten Ausfallproblematik geschuldet ist. Abgesehen von CART, ist für die multiplen Imputationsverfahren BLR und PMM eine bessere Deckungsrate, gegenüber der Ergebnisse zu β_2, erkennenbar. Allerdings liefert nur BLR valide Inferenz hinsichtlich β_3.

4.3 Verallgemeinerung

Zusammanfassend kann festgestellt werden, dass multiple Imputationsverfahren für komplexe Datenausfälle valide Ergebnisse hinsichtlich Inferenz erzielen und die singulären Entsprechungen, für hinreichend großer Ausfallproblematiken, dazu nicht in der Lage sind. Dies hält für alle betrachteten Parameter. Zudem ist ersichtlich, dass die genannten univariaten bzw. deterministischen Verfahren gegenüber den probabilistischen multivariaten Verfahren zu beträchtlichen Schätzfehlern führen können. Daher sind ME, HD und LR nur bedingt als Imputationverfahren verwendbar. Allgemein gilt, dass die Unsicherheit bezüglich des Datenausfalls für valide Inferenz mit dem Imputationsverfahren über probabilistische Komponenten abbildbar sein muss. Wird zudem mit multipel ergänzt, äußert sich die modellierte Unsicherheit durch Rubins Rules in einer Erhöhung der Varianzschätzer, die – bei einer unverzerrten Parameterschätzung – über breitere Konfidenzintervalle valide Inferenz sicherstellt. Obwohl in Kapitel 3 Annahmen linearer Modelle verletzt wurden, ist die multiple Imputation durch BLR robust genug, um diese zu handhaben und liefert so hinsichtlich valider Inferenz die besten Ergebnisse. Dabei ist das ähnlich gute abschneidende Verfahren PMM effizienter. Da mit der Simulation im Kapitel 3 primär die generelle Gegenüberstellung von singulären und multiplen Imputationsverfahren adressiert wurde, kann hierzu keine Allgemeingültigkeit folgen. Die vielfältigen Ausprägungen empirischer Daten würdigend, muss von Fall zu Fall entschieden werden, welches multiple Imputationsverfahren für valide Inferenz zu verwenden ist. Die Leistungen schwanken mit den Anwendungsfällen.

5 Fazit und Ausblick

In dieser Arbeit wurden singuläre und multiple Imputationsverfahren im Kontext der validen Inferenz gegenübergestellt. Ziel war es, Verfahren zu bestimmten, die zu valider Inferenz führen. Gemäß Kapitel 3 wurde eine Simulationsstudie durchgeführt. Mit dem fundamentalen Resultat der Vernachlässigbarkeit von ψ für Datenausfälle MAR und 2.4, wurden die simulierten unvollständigen Daten mit sechs singulären und drei multiplen Imputationsverfahren ergänzt vgl. Kapitel 2. Wegen des generellen Ausfallmusters wurde für multivariate Verfahren – Kapitel 2.1 entsprechend – `mice` verwendet. Auf den vervollständigten Daten wurden nach Unterpunkt 3.2 die drei Parameter β_2, β_3 und μ_4 geschätzt. Die Vergleichskriterien hinsichtlich valider Inferenz waren der relative Fehler, die Deckungsrate und die Weite der Konfidenzintervalle. Diese zielen auf die theoretischen Anforderungen an valide Inferenz aus 3.9 ab. Die Ergebnisse des Vergleichs aus Kapitel 4 zeigen, dass multiple Imputationsverfahren singulären im Kontext der validen Inferenz überlegen sind und insbesondere univariate bzw. deterministische Imputationsverfahren beachtliche Verzerrungen in den Parameterschätzungen führen können, die valide Inferenz ausschließen. Für valide Inferenz sind multivariate Imputationsverfahren zu verwenden, welche die vom Datenausfall verursachte Unsicherheit mit einer Erhöhung der Varianz abbilden. Dies wird durch stochastische Komponenten im Verfahren und multiple Imputation mit der Erweiterung um Rubins Rules in den Parameterschätzungen erreicht. Vergleiche einzelner Imputationsverfahren sind nur innerhalb der Arbeit gültig. Die Simulationsstudie adressiert diese Granularität nicht ausreichend um verallgemeindernde Aussagen treffen zu können.

Hierzu ist eine Modifikation der Simulation durch hinreichend große empirische Datensätze, aus denen an Stelle der Population $\mathcal{Y}$ gezogen wird, erwägenswert. Die Vergleichbarkeit der multiplen Imputationverfahren steigt dabei mit der Anzahl der empirischen Datensätze. Wechselnde Datenkonstellationen lassen vorteilhafte Eigenschaften der Verfahren ersichtlich werden. Sensitivitätsanalysen können dazu u.a. durch Änderungen in den Datenausfallsraten erfolgen. Eine Erweiterung der verwendeten multiplen Imputationsverfahren um die Bayessche Ridge Regression oder die Ensemble Methode Random Forest ist denkbar. Valide Inferenz vernachlässigend, ist ein Vergleich der Verfahren auch mit dem Perspektivenwechsel zur Prädiktionsgüte eines auf ergänzten Daten geschätzten Prognosemodells vorstellbar.

Literaturverzeichnis

Breiman, L., Friedman, J., Stone, C. J. & Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

Buuren, S. v. & Groothuis-Oudshoorn, K. (2010). mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, 1–68.

Doove, L. L., Van Buuren, S. & Dusseldorp, E. (2014). Recursive partitioning for missing data imputation in the presence of interaction effects. *Computational Statistics & Data Analysis*, *72*, 92–104.

Enders, C. K. (2010). *Applied missing data analysis*. Guilford press.

Geman, S. & Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence* (6), 721–741.

Hastie, T., Tibshirani, R. & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

Khinchine, A. (1929). Sur la loi des grands nombres. *Comptes rendus de l'Académie des Sciences*, *189*, 477–479.

Little, R. J. & Rubin, D. B. (2019). *Statistical analysis with missing data* (Bd. 793). John Wiley & Sons.

Myers, R. H. & Myers, R. H. (1990). *Classical and modern regression with applications* (Bd. 2). Duxbury press Belmont, CA.

Raghunathan, T. (2015). *Missing data analysis in practice*. CRC press.

Rubin, D. B. (2009). *Multiple imputation for nonresponse in surveys* (Bd. 307). John Wiley & Sons.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. CRC press.

Van Buuren, S. (2018). *Flexible imputation of missing data*. CRC press.