LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Studienabschlussarbeiten
Fakultät für Mathematik, Informatik und Statistik

Klein, Benjamin:

Empirischer Vergleich von Amelia und Random Forest bei Imputation von fehlenden Daten anhand einer Simulationsstudie

Bachelorarbeit, Sommersemester 2017

Fakultät für Mathematik, Informatik und Statistik

Ludwig-Maximilians-Universität München

https://doi.org/10.5282/ubm/epub.40350

Empirischer Vergleich von Amelia und Random Forest bei Imputation von fehlenden Daten anhand einer Simulationsstudie

Prof. Dr. C. Heumann
Department of Statistics
Ludwig-Maximilans-Univeristät München

WS 16

Bachelorthesis von Benjamin Klein
B.Klein@campus.lmu.de
Statistik
Datum: 3. April 2017

Inhaltsverzeichnis

1 Einführung

1.1 Einleitung

Amelia Earhart war die erste Frau, die alleine um den atlantischen Ozean geflogen ist. Allerdings verschwand sie während eines späteren Fluges über den Pazifischen Ozean 1937 und ist seit dem, zumindest offiziell, nicht mehr aufgetaucht. Ihr verdankt folgende Anwendung in R ihren Namen, welche sich ebenfalls mit dem Problem des Fehlens beschäftigt. Und obwohl Amelia Earhart trotz intensiver Bemühungen nicht mehr gefunden wurde, bietet das R-Paket Amelia eine mögliche Lösung zu diesem Problem. Fehlende Daten kommen immer wieder in statistischen Analysen vor und können statistische Prozeduren erheblich beeinflussen. Sobald ein Datensatz vorliegt, kann es zu einzelnen oder mehrfachen Datenlücken kommen. Die Frage, wie mit eben solchen fehlenden Daten umzugehen ist, ergibt keine eindeutige Antwort und muss vom Datenanalyst immer wieder neu erwägt und hinterfragt werden. Diese Thesis beschäftigt sich mit der Imputation solcher fehlender Werte. Die Untersuchung erfolgt an Simulationsdatensätzen; die zu untersuchenden Daten, die Kovariablenmatrix X und die Reponsevariablen werden im Verlauf der Thesis (immer wieder neu) simuliert. Die fehlenden Werte, die innerhalb des Simulationsprozesses generiert werden, werden wiederum mittels Amelia und einem Random-Forest Ansatz imputiert. Anschließend werden mehrere Regressionen durchgeführt. Anhand der zurückgehaltenen wahren Werte können die verschiedenen Modelle miteinander verglichen werden. Der Vergleich erfolgt sowohl bezüglich der Imputationen, als auch der Regressionen, sodass mit den wahren Parametern beziehungsweise den eigentlichen Werten die Schätzfehler bestimmt werden können. Weiterhin soll der Einfluss einiger Parameter, wie dem Stichprobenumfang oder der Anzahl an Ziehungen des Responsevektors, untersucht werden. Um die Methoden dieser Prozeduren näher erläutern zu können, müssen zunächst die Regularien zum Thema der fehlenden Daten erläutert werden.

1.2 Thematik

Wie [8] ausführlich analysiert, wird in diesem Kontext von fehlenden Daten genau dann gesprochen, wenn wahre Informationen in Form von wahren Werten vorliegen, aber nicht verfügbar oder nicht beobachtet sind. Für die entsprechende Beobachtung liegt demnach keine Information der entsprechenden Variable vor. Zunächst sollte daher auf einige Muster und Mechanismen von fehlenden Werten in Daten eingegangen werden. Folgende Informationen, Definitionen und Notationen sind daher aus [8] entnommen. Sei D eine $n \times q$ Datenmatrix und $d(i,j)$ der Eintrag der Variable j von Beobachtung i aus D, wobei $i \in \{1,\dots,n\}$ und $j \in \{1,\dots,q\}$ in welcher fehlende Werte auftreten. Sei M die dementsprechende fehlende Daten-Matrix, sodass gilt:

$$M_{(i,j)} = \begin{cases} 1, & falls\ d_{(i,j)}\ fehlend \quad \leftrightarrow d_{(i,j)} \in D^{(miss)} \\ 0, & sonst \qquad\qquad\qquad\quad \leftrightarrow d_{(i,j)} \in D^{(obs)} \end{cases}$$

M gibt demnach für jede Variable j zu jeder Beobachtung i an, ob der jeweilige Wert vorliegt oder nicht. Das Fehlen der Daten kann nach [10] in folgende Mechanismen eingeteilt werden, welche den Zusammenhang zwischen dem Fehlen der Daten und den Daten selbst definieren. Falls das Fehlen der Daten nicht von den fehlenden Werten abhängt, also $f(M|D,\phi) = f(M|\phi)$ gilt, kann von **Missing Completely At Random**, oder kurz **MCAR** gesprochen werden. In diesem Fall hängt das Fehlen der Werte nicht von den Daten D ab, sondern nur von einem unbekannten Parameter ϕ. Weder die beobachteten noch die fehlenden Teile der Daten beeinflussen demnach die Wahrscheinlichkeit des Fehlens. Das einfache Ignorieren aller Beobachtungen mit fehlenden Werten beziehungsweise nur der fehlenden Werte selbst wäre eine Option, da die Daten nur rein zufällig fehlen, und somit keine Information in den fehlenden Daten selbst verloren ginge. Falls das Fehlen der Werte mit den fehlenden Daten selbst zusammenhängt, spricht man von **Not Missing At Random**, oder **NMAR**. Die Information der fehlenden Beobachtungen kann nicht akkurat wiedergewonnen werden, da diese nicht in den beobachteten Daten vorhanden sein

muss. Eine Imputation der fehlenden Werte ist demnach schwierig, da keine Anhaltspunkte in den beobachteten Daten vorhanden sein müssen. Ein weiterer Mechanismus von fehlenden Daten ist **Missing At Random**(**MAR**). Falls das Fehlen der Werte zwar an den Daten selbst, jedoch nur mit den beobachteten Daten zusammenhängt, spricht man von **MAR**. **MCAR** impliziert somit **MAR**, welches auch wie folgt ausgedrückt werden kann: $p(M|D) = p(M|D^{obs})$. $p(M)$ bedingt auf die Daten D entspricht also der gleichen Wahrscheinlichkeit, wie wenn diese nur auf die beobachteten Daten D^{obs} bedingt ist. Weiterhin kann **MAR** auch folgendermaßen definiert werden:

$$f(M|D,\phi) = f(M|D^{(obs)},\phi), \quad \forall D^{(miss)},\phi$$

Man gehe zum Beispiel von einer Umfrage per Fragebogen aus, bei dem nicht ausgefüllte Fragen auftreten. **MCAR** setzt voraus, dass das Fehlen der Daten nicht im Zusammenhang mit den Daten selbst steht. Falls zum Beispiel ein paar der Umfragebögen rein zufällig verloren gehen, fehlen die dementsprechenden Antworten, was jedoch nichts mit den verlorenen gegangenen Daten selbst zu tun hat. **MAR** bedeutet, dass das Fehlen mit den beobachteten Daten zusammenhängt. Falls zum Beispiel das Geschlecht der Probanden immer feststeht, aber Männer bestimmte Fragen mit höherer Wahrscheinlichkeit nicht beantworten wollen, hängt das Fehlen zwar von den Daten ab, aber nur von der beobachteten Variable *(Geschlecht)*. Von **NMAR** spricht man, falls das Fehlen der Daten auch mit den fehlenden Daten selbst zusammenhängt. Arbeiter mit höherem Gehalt könnten sich deswegen schämen und die dementsprechende Frage über Gehalt auslassen. Somit hängt das Fehlen dieses Wertes mit diesem Wert selbst zusammen. Eine mögliche Schätzung des Gehalts wäre somit unterschätzt und damit verzerrt. Da viele statistische Prozeduren vollständige Datensätze benötigen, gibt es einige Ansätze zum Umgang mit fehlenden Daten. Die einfachste Methode wäre das Ignorieren von Beobachtungen, in denen mindestens eine Variable nicht vorliegt.(*Complete-Case-Analysis*). Diese Methode kann aber zu erheblichen Bias führen, falls die fehlenden Werte von den übrigen Werten abweichen. Zudem führt die *Complete-Case-Analysis* zu einer Reduzierung des Stichprobenumfangs, was zu einem deutlichen Informationsverlust führen kann, vor allem wenn fehlende Werte in mehreren Variablen auftreten. Zudem wird, sofern man von einem **MAR**-Mechanismus in den Daten ausgeht, die eigentliche Information des Fehlens, welche sich noch in den Daten befindet, nicht berücksichtigt. Eine weitere einfache Alternative bietet die *Available-Case-Analysis*. Hier werden nur die einzelnen fehlenden Werte nicht berücksichtigt, alle anderen verfügbaren Daten gehen in die Untersuchung ein. Je nach Fragestellung wird demnach auf alle interessierenden verfügbaren Daten zurückgegriffen, sodass für unterschiedliche Probleme verschiedene Subdatensätze verwendet werden. Somit sind unterschiedliche Modelle aufgrund unterschiedlicher Datengrundlagen weniger vergleichbar. Diese beiden Methoden versuchen, mit den übrigen beobachteten Werten zu arbeiten. Daher sind diese Ansätze besonders sinnvoll, falls nur wenige der Daten fehlen. Ein weiterer Ansatz, bei dem die Dimension der Datenmatrix unverändert bleibt, ist die Imputation der Daten. Die fehlenden Werte werden hierbei geschätzt. *Single Imputation* ersetzen jeden fehlenden Wert durch einen Schätzwert. Als Schätzung bietet sich zum Beispiel der Mittelwert (*Mean-Imputation*) oder Median der jeweiligen Variable an. Diese können aber, wie [6] berichtet, zu Verzerrung, vor allem bezüglich Varianz- und Kovarianzschätzungen, führen. *Multiple Imputation*-Methoden verwenden komplexere Strukturen, um einen fehlenden Wert zu schätzen. Zudem werden in multiplen Imputationsschritten mehrere Schätzungen generiert, sodass entweder mehrere Datensätze entstehen oder die Schätzungen zu einem Schätzwert generiert werden. Zur tieferen mathematischen Analyse dieser Methode kann, neben [8] auch [4] hinzugezogen werden, welche auch über weitere Methoden zum Umgang mit fehlenden Daten berichten. In dieser Thesis werden zwei multiple Imputationsmethoden verwendet, um die fehlenden Daten zu schätzen. Zunächst soll auf **Amelia** eingegangen werden.

2 Methoden

2.1 Amelia

Amelia ist ein R-Paket, dass zur multiplen Imputation von fehlenden Daten benutzt wird und von James Honaker, Gary King und Metthew Blackwell entwickelt wurde. Gegenüber einfachen Methoden wie der bereits zuvor beschriebenen Complete-Case-Analysis soll Amelia via multipler Imputation den Bias verringern und die Effizienz steigern. Daher soll zunächst der Aufbau von Amelia näher erläutert werden. Dieser Abschnit basiert auf [6]. Zunächst nimmt Amelia multivariat normalverteilte Daten an. Für einen Datensatz D mit den Dimensionen $(n \times q)$ wird also angenommen:

$$D \sim N(\mu, \Sigma)$$

D folgt somit einer multivariaten Normalverteilung mit Erwartungswertvektor μ und Kovarianzmatrix Σ. Die weitere Annahme Amelias betrifft den Mechanismus des Fehlens von Daten. Dabei wird von **MAR** ausgegangen. Verbindet man beide Annahmen und geht zudem davon aus, dass M nicht von den vollständigen Daten Parametern abhängt, für $\theta = (\mu, \Sigma)$, gilt:

$$p(D^{obs}, M|\theta) = p(M|D^{obs})p(D^{obs}|\theta)$$

Dann folgt für die Likelihood Funktion

$$L(\theta|D^{obs}) \propto p(D^{obs}|\theta) = \int p(D|\theta)\, dD^{miss}$$

Gerade unter **MAR** ist eine Imputation der Daten besonders sinnvoll, da die eigentlichen Informationen in den beobachteten Daten noch vorliegen. Um die fehlenden Werte auszufüllen, bedient sich Amelia eines EMB-Algorithmus. Dieser ergibt sich aus einem EM(*expectation-maximization*)-Algorithmus, der dann wiederum auf mehrere Bootstrap-Stichproben angewandt wird. Der EMB-Algorithmus kombiniert das EM-Modell, welches abwechselnd neue Imputationen anhand der zuvor berechneten Parameter und anschließend neue Parameter anhand der zuvor bestimmten Imputationen iteriert, mit dem Bootstrap-Ansatz. Dieser wiederum simuliert die Unsicherheit des Modells aufgrund fehlender Daten mit der Unsicherheit des Bootstraps. Diese wird durch das Ziehen mit Zurücklegen simuliert, da somit nur eine Stichprobe dem EM-Algorithmus übergeben wird. [2] beschreibt den mathematischen Hintergrund des EM-Algorithmus ausfürlich, während [5] näher auf den EMB-Algorhitmus eingeht.

Im Amelia-Algorithmus werden Zunächst m vervollständigte Datensätze erstellt, in dem für jede Datenlücke m verschiedene Werte eingefüllt werden. Zur Imputation werden zunächst die Parameter μ und Σ anhand gebootstrapter Samples der Daten geschätzt. Da die gebootstrapten Samples nur Subdatensätze des wahren Datensatz darstellen, entstehen aus den verschiedenen Samples unterschiedliche Schätzungen für μ und Σ. Im $E - Step$ des EM-Algorithmus werden dann die Erwartungswerte der fehlenden Werte bedingt der geschätzten Parameter $\hat{\mu}$ und $\hat{\Sigma}$ bestimmt. Der EM-Algorithmus selbst setzt sich aus folgenden beiden Schritten zusammen:

1. *Expectation-Step:* In diesem Schritt wird jeder fehlende Wert aus der vorherigem Iteration berechnetem θ imputiert. Zur Schätzung werden zudem die beobachteten Daten aus X hinzugezogen. Amelia verwendet für diese Schätzung eine Regression mit dem fehlenden Wert als Responsevariable.

2. *Maximization-Step:* Die *Maximum-Likelihood-Methode* wird angewandt, um neue Parameter θ zu bestimmen. In diesem Fall besteht θ aus neuen Schätzungen der Parameter μ und Σ. Die Berechnung erfolgt anhand der im *Expectation-Step* bestimmten Daten.

Diese Schritte werden so lange wiederholt, bis der Prozess konvergiert und sich die berechneten Parameter also nicht mehr wesentlich ändern. Der mathematische Algorithmus dieser Prozedur stellt sich folgendermaßen dar:

E-Step: Schätze $Q(\theta; \theta^{(t)})$, wobei

$$Q(\theta, \theta^{(t)}) = E_{\theta^{(t)}}[l(\theta; y|y_{obs})]$$

M-Step: Berechne $\theta^{(t+1)}$ aus θ, sodass:

$$Q(\theta^{(t+1)}; \theta^{(t)}) \geq Q(\theta; \theta^{(t)})$$

Im E-Step wird also der Erwartungswert der Log-Likelihood von y gegeben den beobachteten Daten y_{obs} bestimmt. Im Allgemeinen lässt sich für normalverteilte Daten die Likelihood-Funktion nach [3] aus der Dichte von $\{y_1 \ldots, y_q\}$ bestimmen:

$$L\left(\mu, \sigma^2|y\right) = f\left(y_1, \ldots, y_n \mid \mu, \sigma^2\right) = \prod_{i=1}^{n} f\left(y_i \mid \mu, \sigma^2\right) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left(-\frac{\sum_{i=1}^{n}(y_i - \mu)^2}{2\sigma^2}\right)$$

Für die Log-Likelihood ergibt sich damit

$$\ell(\mu, \sigma) = \log L(\mu, \sigma^2) = -\frac{n}{2} \cdot \log\left(2\pi\sigma^2\right) - \frac{\sum_{i=1}^{n}(y_i - \mu)^2}{2\sigma^2}$$

Anhand dieser Log-Likelihood können somit die ML-Schätzer für μ und σ^2 bestimmt werden, welche dann wiederum im E-Step zur Schätzung der fehlenden Werte verwendet werden können. Für die fehlenden Werte y_{obs} können somit anhand der aus dem Bootstrap generierten Parameterschätzungen für θ Schätzungen getätigt werden. Diese werden wiederum in den weiteren Iterationen des EM-Algorithmus verbessert. Nach [10] kann im Normalfall $m = 5$ als Richtwert benutzt werden. Nachdem der EM-Algorithmus auf die gebootstrapten Stichproben angewandt wurde, können die Schätzer nach *Rubins rule* kombiniert werden. In [10] wurden diese Kombinationsregeln zur Bestimmung der Schätzer und deren Varianzen definiert, die folgende Notation ist nach [1] gewählt. Seien $\hat{Q}^{(k)}$ der Punktschätzer und $U^{(k)}$ dessen Varianz jeweils im k-ten Datensatz, $k \in \{1, \ldots, m\}$. Dann gilt für die Vereinigung aller m Schätzer:

$$\bar{Q} = \frac{1}{m} \sum_{t=1}^{m} \hat{Q}^{(t)}$$

In diesem Fall wurden die m aus den imputierten Ameliadatensätzen erstellten Regressionsmodelle und die daraus resultierenden Koeffizientenschätzern $\hat{\beta}_{j,Amelia}$, $j \in \{1, \ldots, q\}$ zusammengefügt.

$$\hat{\beta}_{j,Amelia} = \bar{\beta}_j = \frac{1}{m} \sum_{t=1}^{m} \hat{\beta}_{j,t}$$

Der aus der Kombination der m Amelia-Regressionen (Regressionen basierend auf den imputierten Amelia-Datensätzen) gewonnene Koeffizientenschätzer einer Variable j entspricht also nur dem arithmetischen Mittel der m Koeffizientenschätzer der Amelia-Regressionen. Neben den kombinierten Punktschätzer der Koeffizienten wurde diese Regel auch auf die Kombination der Punktschätzer der fehlenden Werte der Imputationsdatensätzen von Amelia angewandt. Zur Bestimmung der Varianz des Punktschätzers müssen sowohl die Streuungen innerhalb der Schätzungen als auch die Streuungen zwischen den Schätzungen berücksichtigt werden. Der Varianzteil innerhalb der Schätzungen ergibt sich wiederum aus dem Mittelwert der m geschätzten Varianzen der einzelnen Modelle.

$$W = \frac{1}{m} \sum_{t=1}^{m} U^{(t)}$$

Die Varianz des Koeffizientenschätzers der kombinierten Amelia-Regressionen ergibt sich also aus deren gemittelten einzelnen Koeffizientenschätzern. Die weitere Varianzteil zwischen den

Schätzungen lässt sich aus der unverzerrten Stichprobenvarianz der einzelnen Punktschätzer bestimmen:

$$B = \frac{1}{m-1} \sum_{t=1}^{m} \left(\hat{Q}^{(t)} - \bar{Q} \right)^2$$

Die gesamte Varianz T ergibt sich nach [10] aus Kombination der beiden Varianzteile:

$$T = W + (1 + \frac{1}{m})B$$

Wie die Multivariate-Noramlvereteilungsannahme zeigt, ist Amelia eine paramterische Imputationsmethode. Bei Verletzung dieser Annahme ist die Wirksamkeit dieser Methode eingeschränkt und kann zu Verzerrungen der Ergebnisse führen. Allerdings kann nach [6] auch bei nicht normalverteilten Daten bei ausreichendem Stichprobenumfang mit akkuraten Ergebnissen gerechnet werden. Daher soll in dieser Thesis das Abschneiden von Amelia mit einer nichtparametrischen Imputationsmethode verglichen werden. Dies soll im folgenden Abschnitt näher erläutert werden.

2.2 Random Forest

Nichtparametrische Methoden nehmen keine parametrischen Verteilungen an. Es wird also nicht ein passende Verteilungsfunktion F und deren entsprechende Parameter gesucht, sondern vielmehr wird eine Funktion f an die Datenpunkte angepasst. Der Random Forest-Algorithmus, welcher in [7] entwickelt und erläutert wird, bietet also eine nichtparametrische Alternative zur Datenanalyse. Dabei bedient sich das Modell einfachen Entscheidungsbäumen, die durch Kombination für Schätzungen verwendet werden. Der Random Forest Algorhitmus geht im Allgemeinen folgendermaßen vor:

1. Zunächst werden n_{tree} Bootstrap-Stichproben aus den Daten gezogen.
2. Nun wird für jede einzelne Stichprobe ein Regressions- oder Klassifikationsbaum erstellt, je nach Typ der Daten.
3. An jedem Zweig/Knoten des Baumes werden m_{try} der Variablen ausgewählt und der beste Split der Daten anhand dieser Variablen bestimmt.
4. Um Daten vorherzusagen, werden die Vorhersagen der n_{tree} Entscheidungsbäume kombiniert.

In diesem Fall wird der Random Forest Algorithmus verwendet, um Schätzungen für die fehlenden Werte anhand der übrigen beobachteten Daten zu generieren. Zur Imputation wurde hier das R-Package *missForest* verwendet([11]). Sei daher die Kovariablenmatrix $\mathbf{X}(n \times q)$ vorhanden, sodass $X = (X_1, X_2, \ldots, X_q)$. Sei zudem $X_s, s \in \{1, \ldots, q\}$ eine Variable mit fehlenden Werten der Stellen $i_{miss}^{(s)} \subseteq \{1, \ldots, n\}$ (und demnach mit vorhandenen beobachteten Werten $i_{obs}^{(s)} \subseteq \{1, \ldots, n\}$). Somit kann der Datensatz in 4 Teile eingeteilt werden:

1. y_{obs}^{s} Die beobachteten Werte der Variable X_s
2. y_{miss}^{s} Die fehlenden Werte der Variable X_s
3. x_{obs}^{s}: Die Werte aller Variablen außer X_s an den Stellen $i_{obs}^{(s)}$
4. x_{miss}^{s} Die Werte aller Variablen außer X_s an den Stellen $i_{miss}^{(s)}$

y_{obs} und y_{miss} sind demnach die beobachteten bzw. fehlenden Werte der Variable X_s, während x_{obs} und x_{miss} die dementsprechenden Werte aller anderen Variable an der jeweiligen Stelle beschreiben. *MissForest* schätzt zunächst die fehlenden Werte mit einfachen Prozeduren wie mean

imputation, bei welcher die Mittelwerte der Variablen eingesetzt werden. Dann werden die Variablen X_s einzeln nach ihrem Anteil von fehlenden Werten sortiert. Beginnend mit der Variable mit den wenigsten fehlenden Werten wird jeweils ein RandomForest nach [7] mit Responsevariable $y_{obs}^{(s)}$ und mit den Prädiktoren $x_{obs}^{(s)}$ berechnet. Anschließend wird dieses zuvor trainierte Modell zur Schätzung der fehlenden Werte $y_{miss}^{(s)}$ anhand der übrigen Prädiktorvariablen $x_{miss}^{(s)}$ angewandt. Das stop-criterion γ wird erreicht, sobald sich die neuen Imputationen nicht mehr ausreichend von der vorherigen Imputation unterschieden. Sobald also die Differenz einer stetigen Variable N γ unterschreitet, wird der Prozess abgebrochen, wobei die Differenz der Imputationen Δ_N folgendermaßen definiert ist:

$$\Delta_N = \frac{\sum_{j \in N}(X_{new}^{imp} - X_{old}^{imp})^2}{\sum_{j \in N}(X_{new}^{imp})^2}$$

Im nächsten Abschnitt soll die Aufgabenstellung dieser Thesis erörtert werden.

2.3 Methodik

Sowohl Amelia als auch Random Forest, beziehungsweise missForest, eignen sich also zur Imputation von fehlenden Werten. Daher sollen diese beiden Methoden im weiteren Verlauf dieser Thesis anhand eines simulierten Datensatzes sowohl hinsichtlich der Imputationen als auch einer aus den Daten resultierenden Regression verglichen werden. Dafür wird zunächst eine Kovariablenmatrix X simuliert. Mittels festgelegter, variierender Parameter wird anschließend anhand einer Linearkombination der X-Variablen ein Responsevektor generiert. Abhängig dieser Responsevektoren werden daraufhin fehlende Werte innerhalb von X erzeugt, wobei die entfernten Daten für spätere Untersuchungen zurückgehalten werden. Diese fehlenden Werte werden dann mit dem vollständigen Responsevektor und den übrigen Variablen anhand Amelia und RandomForest imputiert. Die imputierten Datensätze werden zunächst mit dem wahren Datensatz verglichen. Anschließend werden Regressionen der ursprünglichen und imputierten Datensätze auf die Responsevariablen berechnet. Unter Verwendung der Koeffizientenschätzer der Regressionsmodelle kann wiederum das Abschneiden der Imputation verglichen und bewertet werden. Zur Bewertung des Abschneidens wird der **Mean squarred Error** (**MSE**) herzugezogen, der wie folgt definiert ist:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(Y_{true} - \hat{Y}\right)^2$$

Der MSE beschreibt also die quadrierte Differenz der wahren Y_{true} von den geschätzten $\hat{Y}$ Werten einer Variable Y. Zur Bewertung der Imputationen wird zudem der nach [9] definierte **NRMSE** (Normalized root mean squarred error) verwendet:

$$NRMSE = \sqrt{\frac{mean[(y_{guess} - y_{answer})^2]}{Var(y_{answer})}}$$

Somit misst der NRMSE auch die durchschnittliche (quadrierte) Differenz der imputierten y_{guess} und der wahren Daten y_{answer}, normiert diese aber zudem über die Varianz der wahren Daten. Für einen NRMSE von 0 gilt perfekte Übereinstimmung der Datensätze, bei einem NRMSE von 1 liegen die Daten um die Varianz vom Datenpunkt entfernt. Variablen unterschiedlicher Streuungen können aufgrund dieser Standardisierung miteinander verglichen werden. Weiterhin wird der MSE in 2 Komponenten aufgeteilt:

$$
\begin{aligned}
MSE &= \frac{1}{S}\sum_{s=1}^{S}(\hat{\beta}_{sj}-\beta_j)^2 = \frac{1}{S}\sum_{s=1}^{S}(\hat{\beta}_{sj}-\bar{\beta}_j+\bar{\beta}_j-\beta_j)^2 \qquad \text{,wobei } \bar{\beta}_j = \frac{1}{S}\sum_{s=1}^{S}\hat{\beta}_{sj} \\
&= \frac{1}{S}\sum_{s=1}^{S}\Big((\hat{\beta}_{sj}-\bar{\beta}_j)^2+(\bar{\beta}_j-\beta_j)^2+2*(\hat{\beta}_{sj}-\bar{\beta}_j)*(\bar{\beta}_j-\beta_j)\Big) \\
&= \frac{1}{S}\sum_{s=1}^{S}(\hat{\beta}_{sj}-\bar{\beta}_j)^2+\frac{1}{S}\sum_{s=1}^{S}(\bar{\beta}_j-\beta_j)^2+2*\frac{1}{S}\sum_{s=1}^{S}(\hat{\beta}_{sj}-\bar{\beta}_j)*(\bar{\beta}_j-\beta_j)
\end{aligned}
$$

Weiterhin gilt: $\frac{1}{S}\sum_{s=1}^{S}(\hat{\beta}_{sj}-\bar{\beta}_j)*(\bar{\beta}_j-\beta_j)=0$, da:

$$
\begin{aligned}
\frac{1}{S}\sum_{s=1}^{S}(\hat{\beta}_{sj}-\bar{\beta}_j)*(\bar{\beta}_j-\beta_j) &= \frac{1}{S}\sum_{s=1}^{S}\Big[\hat{\beta}_{sj}\bar{\beta}_j-\bar{\beta}_j\bar{\beta}_j-\hat{\beta}_{sj}\beta_j+\bar{\beta}_j\beta_j\Big] \\
&= \frac{1}{S}\sum_{s=1}^{S}\Big[\hat{\beta}_{sj}\bar{\beta}_j-\bar{\beta}_j\bar{\beta}_j+\bar{\beta}_j\beta_j\Big]-\frac{1}{S}\sum_{s=1}^{S}(\hat{\beta}_{sj}\beta_j) \\
&= \bar{\beta}_j\frac{1}{S}\sum_{s=1}^{S}\Bigg[\underbrace{\hat{\beta}_{sj}}_{\frac{1}{S}\sum_{s=1}^{S}\hat{\beta}_{sj}=\bar{\beta}_j}-\bar{\beta}_j+\beta_j\Bigg]-\beta_j*\frac{1}{S}\sum_{s=1}^{S}(\underbrace{\hat{\beta}_{sj}}_{\frac{1}{S}\sum_{s=1}^{S}\hat{\beta}_{sj}=\bar{\beta}_j}) \\
&= \bar{\beta}_j\left(\left[\frac{1}{S}\sum_{s=1}^{S}\hat{\beta}_{sj}\right]-\bar{\beta}_j+\beta_j\right)-\beta_j*\bar{\beta} \\
&= \bar{\beta}_j\left(\bar{\beta}_j-\bar{\beta}_j+\beta_j\right)-\bar{\beta}_j*\beta_j \\
&= \bar{\beta}_j*\beta_j-\bar{\beta}_j*\beta_j \\
&= 0
\end{aligned}
$$

Somit gilt:

$$
MSE = \frac{1}{S}\sum_{s=1}^{S}(\hat{\beta}_{sj}-\beta_j)^2 = \frac{1}{S}\sum_{s=1}^{S}(\hat{\beta}_{sj}-\bar{\beta}_j)^2+\frac{1}{S}\sum_{s=1}^{S}(\bar{\beta}_j-\beta_j)^2
$$

Anstatt einer Komponente des MSEs wird die (quadrierte) Abweichung vom wahren Wert nun in einen Varianz- und einen Biasteil aufgeteilt. Der Varianzteil beschreibt dann die Streuung um den Mittelwert der Schätzungen, während der Biasteil die Präzision misst, also die (quadrierte) Abweichung der mittleren Schätzung zum wahren Wert an sich. Der kombinierte MSE-Schätzer ist somit eine Schätzung der empirischen Varianz der einzelnen Koeffizientenschätzer β_j. Dieser kann sich dann wiederum mit den geschätzten Varianzen der Koeffizientenschätzer ergeben. Diese lassen sich im linearen Modell aus der Kovarianzmatrix des ML-Schätzers von β bestimmen, die nach [3] wie folgt definiert ist:

$$
\hat{\mathrm{Cov}}(\beta) = \hat{\sigma}^2(X'\ X)^{-1} = \frac{1}{n-q}\hat{\epsilon}'\ \hat{\epsilon}(X'\ X)^{-1}
$$

Da $\hat{\sigma}^2$ somit von ϵ abhängt, welches sich in einem linearen Regressionsmodell als Schätzfehler $\hat{\epsilon} = y - X\hat{\beta}$ ergibt, verändert sich die geschätzte Kovarianzmatrix in jedem Regressionsmodell. Im Logit-Modell ist der ML-Schätzer wiederum nach [3] asymptotisch normalverteilt. Die geschätzte Kovarianzmatrix ergibt sich aus der inversen Fisher-Matrix, wobei die Diagonalelemente a_{rr} der inversen Fishermatrix somit Schätzer der Varianz der r-ten Komponente von β (β_r) sind.

$$
\hat{\mathrm{Cov}}(\hat{\beta}) = \mathbf{F}^{-1}(\hat{\beta})
$$

Somit hängen beide Kovarianzen mit den Regressionsmodellen zusammen. Daher werden diese nur bezüglich der Untersuchungen analysiert, bei denen X nur einmal simuliert wird. Im nächsten Abschnitt soll beschrieben werden, wie die Daten simuliert wurden.

2.4 Daten

Die Ergebnisse der Untersuchung hängen natürlich sehr stark von den Daten X selbst ab. Demnach wurde X vielseitig simuliert, um möglichst aussagekräftige Ergebnisse zu erzielen. Generell wurden $q = 10$ Variablen verwendet, so dass X eine $(n \times 10)$ Matrix darstellt. Im Verlauf der Untersuchung wird der Stichprobenumfang der Daten mehrfach variiert. Weiterhin werden einige Parameter der Verteilungen, aus denen die X Werte gezogen sind, ebenso wie einige Faktoren, anhand derer die Responsevariabeln bestimmt werden, differiert. Die Stichprobenanzahl wird im Verlauf der Untersuchung, wie auch sowohl die Parameter der Verteilungen, aus denen die X-Werte gezogen wurden, als auch die Faktoren, anhand derer die Responsevariablen aus X bestimmt werden, mehrfach variiert. Für die Simulation von X werden folgende Ansätze verwendet:

1. $X \sim \mathcal{N}(\mu, \Sigma)$, d.h. X folgt einer multivariaten Normalverteilung. Damit wären die Anforderungen für Amelia erfüllt

2. X folgt keiner multivariaten Normalverteilung. Die X-Variablen werden aus anderen Verteilungen generiert.

Falls X einer multivariaten Normalverteilung folgt, gilt:

$$
\begin{aligned}
X &\sim \mathcal{N}(\mu, \Sigma), \qquad \text{,wobei} \\
\mu &= (\theta_1/2, 0, 1, , 4, 0.5, 2, 10, 3, 5) \\
\Sigma(i,j) &= \begin{cases} 1, & \text{falls } i = j \\ 0.4, & \text{falls } i \neq j \end{cases} \qquad \forall i \in \{1, \ldots, 10\},\, j \in \{1, \ldots, 10\}
\end{aligned}
$$

Die Kovarianz zwei verschiedener Variablen x_i und x_j beträgt damit 0.4.

Da im Allgemeinen aber meistens nicht alle Variablen einer Normalverteilung folgen, wird der Fokus dieser Thesis auf den zweiten Fall der Daten gelegt. Tabelle 1 beschreibt daher die detaillierte Konstruktion der einzelnen Variablen.

Die zugehörigen Parameter der Verteilungen, anhand welcher die X-Variablen konstruiert werden, werden aus folgenden Werten ausgewählt, welche in Tabelle 2 dargestellt sind:

Demnach sind auch die Abhängigkeiten der einzelnen Variablen wesentlich komplexer als bei ausschließlich multivariat normalverteilten Daten. Abbildung 1 soll daher eine Vorstellung der Abhängigkeiten der Variablen vermitteln. In Kapitel 3.2 wird ein Grundmodell erstellt, anhand dem die Variationen einzelner Parameter verglichen werden können. Bei der Simulation des Grundmodells von X ergeben sich folgende Korrelationen, die in Abbildung 1 dargestellt sind.

Für jede Simulation von X wird demnach genau einer der vorliegenden Werte pro Parameter zufällig bestimmt. Zur Konstruktion der Responsevariablen aus X wurden zwei generelle Herangehensweisen benutzt:

1. Lineares Modell

2. Logit-Modell

Im Folgenden wird zur besseren Unterscheidung Z für die Zielvariable des linearen Modell und Y für die Zielvariable des Logit-Modells gewählt.

Tabelle 1: Konstruktion der X-Kovariablenmatrix

Var	**konstruiert aus**	**detailliert**
x_1	Gleichverteilung	$x_1 \sim \mathcal{U}(0, \theta_1)$
x_2	Normalverteilung	$x_2 \sim \mathcal{N}(0, 1)$
x_3	Exponentialverteilung	$\forall i \in \{1, \ldots, n\} x_3[i] \sim Exp(\lambda = \theta_2 + \sqrt{\lvert x_2[i] \rvert})$
x_4	*diskret*	x_4 nimmt mit Wahrscheinlichkeit $\begin{cases} 0.7 & \text{l an, falls } x_1[i] \in \theta_1 \left[\frac{l-1}{4}, \frac{l}{4}\right], l \in \{1, \ldots, 4\} \\ 0.1 & k_1 \text{ an, wobei } k_1 \in \{1, \ldots, 4\} \quad ; k_1 \neq k_2 \neq k_3 \neq l \\ 0.1 & k_2 \text{ an, wobei } k_2 \in \{1, \ldots, 4\} \quad ; k_1 \neq k_2 \neq k_3 \neq l \\ 0.1 & k_3 \text{ an, wobei } k_3 \in \{1, \ldots, 4\} \quad ; k_1 \neq k_2 \neq k_3 \neq l \end{cases}$
x_5	Poissonverteilung	$\forall i \in \{1, \ldots, n\} x_5[i] := \begin{cases} x_5[i] \sim Pois(1, \theta_4) \begin{cases} \text{if} & x_4[i] = 1 \lvert 4 \quad \& \\ & x_1 \in [0.25 * \theta_3, 0.75 * \theta_3] \\ \text{if} & x_4[i] = 2 \lvert 3 \quad \& \\ & x_1 \notin [0.25 * \theta_3, 0.75 * \theta_3] \end{cases} \\ x_5[i] \sim Pois(1, 10 - \theta_4) \begin{cases} \text{if} & x_4[i] = 2 \lvert 3 \quad \& \\ & x_1 \in [0.25 * \theta_3, 0.75 * \theta_3] \\ \text{if} & x_4[i] = 1 \lvert 4 \quad \& \\ & x_1 \notin [0.25 * \theta_2, 0.75 * \theta_2] \end{cases} \end{cases}$
x_6	Normalverteilung	$\forall i \in \{1, \ldots, n\} x_6[i] \sim \mathcal{N}(\bar{x}_3, \sqrt{x_5[i]})$
x_7	multivariate Normalverteilung	$x_7 \sim \mathcal{N}(\mu_7, \Sigma_7)$ $\mu_7 = (-5, 20, -10)$ $\Sigma_7 = \begin{pmatrix} 1 & 0.3 & 0.1 \\ 0.3 & 1 & 0.2 \\ 0.1 & 0.2 & 1 \end{pmatrix}$
x_8	Gammaverteilung	$\forall i \in \{1, \ldots, n\} x_8[i] \sim \mathcal{G}(1, \sqrt{\lvert x_6[i] \rvert})$

Tabelle 2: Ausprägungen von θ und anderen Parametern

Parameter	mögliche Ausprägungen
θ_1	(50,75,100)
θ_2	(0.3,0.4,0.6)
θ_3	(0.1,0.25,0.4)
θ_4	(3,5,7)
σ	*(2.25,9,36)*
π	*(0.05,0.1,0.25)*
S	*(10,50,100)*

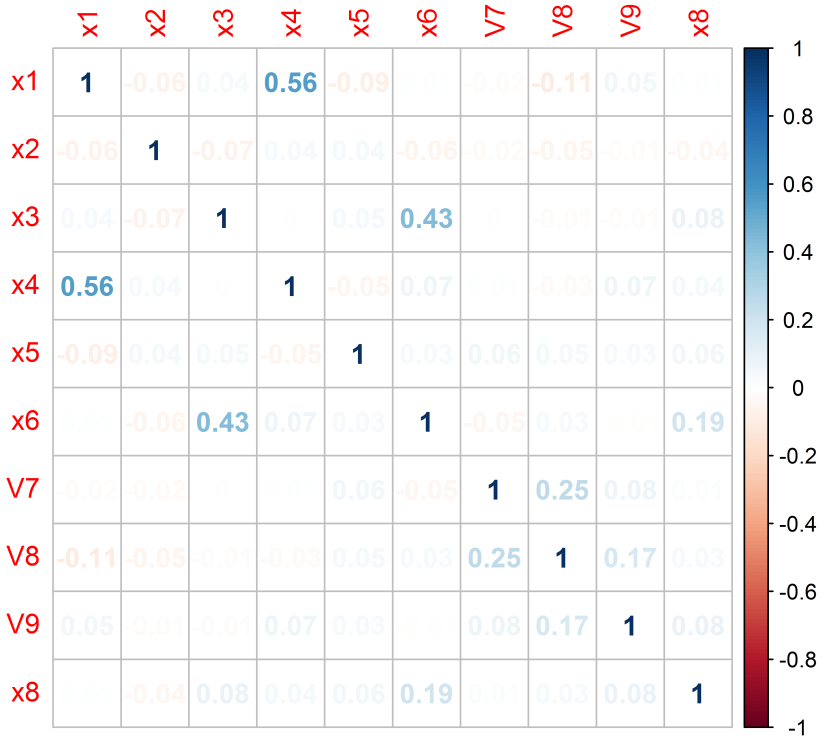


Abbildung 1: Korrelationen eines simulierten Datensates

Im Fall des Linearen Modells wird Z als Linearkombination aus X erstellt, wobei die Faktoren aus $\beta_1, \ldots, \beta_q$ bestehen:

$$\begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{pmatrix} = \begin{pmatrix} x_{(1,1)} & x_{(1,2)} & \cdots & x_{(1,q)} \\ x_{(2,1)} & x_{(2,2)} & \cdots & x_{(2,q)} \\ \vdots & \vdots & \ddots & \vdots \\ x_{(n,1)} & x_{(n,2)} & \cdots & x_{(n,q)} \end{pmatrix} \times \begin{pmatrix} \beta_1 & \beta_2 & \ldots & \beta_q \end{pmatrix}^T + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

Die einzelnen Komponenten lassen sich dann folgendermaßen berechnen.

$$z_i = \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_q x_{i,q} + \epsilon_i$$

$(\beta_1, \ldots, \beta_q)$ sind wiederum festgelegte Parameter. Zudem wird ein Fehlerterm ϵ simuliert, für welchen gilt: $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Die wahren β der Faktoren der Linearkombination, mit denen Y und Z konstruiert werden, nehmen folgende Werte an, die in Tabelle 3 zu sehen sind:

Tabelle 3: Ausprägungen β

Parameter	mögliche Ausprägungen
β_1	(0.02,0.05,0.1)
β_2	(-1,1,2)
β_3	(-0.1,0.1,0.3)
β_4	(1,2,4)
β_5	(-.1,-0.7,-0.3)
β_6	(-0.5,-0.4,-0.2)
β_{7_2}	(0.4,0.6,0.9)
β_{7_2}	(-1,-0.8,-0.6)
β_{7_3}	(0.5,0.7,1)
β_8	(-0.7,-0.5,-0.3)

Im Logit-Modell wird Y nicht direkt aus X bestimmt. Y ist hier eine binomiale Variable, wobei die Wahrscheinlichkeit Π , dass Y_i 1 annimmt, von einer weiteren Variable η abhängt:

$$P(Y = 1|X) = \Pi = \frac{\exp(\eta)}{1 + \exp(\eta)}$$

η entspricht dann wiederum eine Linearkombination aus X, jedoch im Gegensatz zum linearen Modell ohne Fehlerterm ϵ:

$$\begin{pmatrix} \eta_1 \\ \eta_2 \\ \vdots \\ \eta_n \end{pmatrix} = \begin{pmatrix} x_{(1,1)} & x_{(1,2)} & \dots & x_{(1,q)} \\ x_{(2,1)} & x_{(2,2)} & \dots & x_{(2,q)} \\ \vdots & \vdots & \ddots & \vdots \\ x_{(n,1)} & x_{(n,2)} & \dots & x_{(n,q)} \end{pmatrix} \times \begin{pmatrix} \beta_1 & \beta_2 & \dots & \beta_q \end{pmatrix}^T$$

Sowohl das lineare als auch das Logit-Modell beinhalten keinen konstanten Intercept β_0, da Y beziehungsweise Z auch ohne eine solche Konstante konstruiert wurden. Für die einzelne Komponente η_i gilt dann demnach:

$$\eta_i = \beta_1 x_{i,1} + \beta_2 x_{i,2} + \dots + \beta_q x_{i,q}$$

Y_i ist somit binomial verteilt, wobei die Wahrscheinlichkeit, dass $Y_i = 1$ immer von η_i und letztlich somit auch von $x_{i,1}, \dots, x_{i,q}$ abhängt. Im weiteren Verlauf der Untersuchung werden Y und Z jeweils mehrfach gezogen, und die Regression der Kovariablen somit auf mehrere Responsevektoren angewandt. Dabei bleiben die sonstigen Parameter für beide Ziehungen gleich, es gilt:

$$z_i = \eta_i + \epsilon_i$$

Nach der bereits beschriebenen Simulation von X werden anschließend die fehlenden Werte generiert. Die Wahrscheinlichkeit, dass die Beobachtung i in der Variable j nicht vorhanden ist, hängt sowohl vom vorher festgelegten Parameter π als auch von Π ab, wobei Π, wie bereits erwähnt, die Wahrscheinlichkeit $P(Y = 1|X)$ darstellt. Sei dafür m wiederum die fehlende-Daten Matrix, d.h. $m_{i,j} = 1, i \in \{1, \dots, n\}; j \in \{1, \dots, q\}$, falls $x_{i,j}$ nicht vorhanden ist und $m_{i,j} = 0$, sonst. Dann wird für jede Variable j genau i mal, also für jede Beobachtung i ein Bernoulliexpermient mit $p = P(m_{j_i} = 1|X, Y)$ durchgeführt, um zu bestimmen, ob $x_{i,j}$ vorhanden ist oder zurückgehalten wird. Für p gilt:

$$p = \pi * \frac{\Pi_i}{\frac{1}{n}\sum_{j=1}^{q} \Pi_j}$$

Die Wahrscheinlichkeit, dass x_{i_j} fehlt entspricht also dem Verhältnis von Π_j zu dem Mittelwert aller Π ($\bar{\Pi}$) multipliziert mit einem festgelegten Parameter π. π entspricht somit der Wahrscheinlichkeit p, dass x_{i_j} fehlend ist, falls Π_i genau dem Mittelwert $\bar{\Pi}$ entspricht. Je nach Verhältnis $\frac{\Pi_j}{\bar{\Pi}}$

ergibt sich demnach p, wobei für den Term gilt, dass das Produkt dieser Verhältnisse 1 ergibt: $\prod i = 1^n \frac{\Pi_i}{\Pi} = 1$. Im Folgenden wird daher unter π der erwartete Anteil der fehlenden Werte verstanden. Die fehlenden Werte werden dann anhand Amelia und missForest imputiert. Amelia wird zudem übergeben, dass x_4 als ordinale Variable bestimmt werden soll (sofern nicht alle x_j normalverteilt sind). Die Anzahl an Endscheidungsbäumen für die Random-Forest-Kalkulation innerhalb des missForest-Algorithmus wird zudem auf 50 gesetzt, um die Berechnungsdauer etwas zu senken. Nun können die imputierten Datensätze hinsichtlich der MSEs der Schätzungen der fehlenden Werte berechnet werden. Wie zuvor beschrieben, wird der MSE in eine Varianz- und Biaskomponente unterteilt. Anschließend werden jeweils vier Regressionsmodelle erstellt:

1. Regression auf Datensatz ohne fehlende Werte (volles Modell)
2. Regression auf Datensatz mit fehlenden Werten. Hierbei berücksichtigt R automatisch ausschließlich vollstänige Beobachtungen ohne fehlende Werte (*Complete-Case-Analysis*)
3. Regression auf mit Amelia vervollständigtem Datensatz
4. Regression auf mit missForest vervollständigtem Datensatz

Demnach werden vier Regressionsmodelle der vier Datensätze sowohl auf Y als auch auf Z erstellt. Diese werden dann hinsichtlich der MSEs der Koeffizienten im Vergleich zu den wahren Faktoren, mit denen Y beziehungsweise Z simuliert wurden, verglichen. Um die Aussagekraft der Ergebnisse zu verstärken, werden sowohl die Kovariablen als auch die Responsevariablen mehrfach simuliert. Die Responsevektoren werden S mal gezogen, wobei S (, außer bei Variationen dieses Parameters) 50 beträgt. Demnach werden auch je $4 * S$ Regressionsmodelle der vier Kovariablendatenmatrizen auf die S gezogenen Responsevektoren berechnet. Der MSE wird somit über die S Ziehungen gemittelt. Der gesamte Prozess wird im späteren Verlauf der Thesis dann wiederum W mal durchgeführt, wobei bei jeder w-ten Wiederholung ein neuer X Datensatz mit neuen Parametern (welche, wie bereits erläutert, aus dem Pool der möglichen Ausprägungen (2) zufällig ausgewählt werden) simuliert wird. Somit können am Ende gemittelte Ergebnisse über die W Simulationen von X, auf welche dann wiederum S Simulationen von Y bzw. Z folgen, bestimmt werden. Zusammengefasst wurden im Experiment also folgende Schritte durchgeführt:

1. Simuliere X
   - multivariat normalverteilt
   - generiert aus mehrere Verteilungen
2. Simuliere Y und Z abhängig von X
3. Generiere fehlende Werte in X abhängig von Y
4. Berechne Regressionsmodelle von X auf Y und Z
   - volles Modell
   - Complete Cases
   - anhand mit Amelia imputierter Daten
   - anhand mit missForest imputierter Daten
5. Auswertung anhand MSE(-Komponenten)
   - bezüglich Koeffizientenschätzer
   - bezüglich Imputationen

3 Variation einzelner Parameter bei festem Datensatz X

3.1 Probleme und Fehler bei der Simulation

Bevor die eigentlichen Ergebnisse präsentiert werden sollen, werden anschließend einige Probleme erläutert, die bei der Simulation von Daten auftreten können beziehungsweise aufgetreten sind.

In den Ergebnissen treten häufig extrem hohen Fehler in den Modellen der Complete-Case-Analysis auf. Gerade bei Reduzierung des Stichprobenumfangs oder Erhöhung von π kann es zu einer sehr niedrigen Anzahl an Beobachtungen ohne jegliche fehlenden Werten kommen. Da allerdings nur diese für die Complete-Case-Analysis berücksichtigt werden, ist somit schon die Schätzung stark beeinträchtigt. Weiterhin kann beim Logit-Modell ein sehr geringer Stichprobenumfang n_{CC} im Vergleich zu der Anzahl an Variablen q oder ungünstige Datenkonstellationen nach [3] zu einer Nichtexistenz, oder zumindest einer Nicht-Konvergenz des ML-Schätzers führen. In diesem Fall divergieren die sukzessiven Differenzen $||\hat{\beta}^{(k+1)} - \hat{\beta}^{(k)}||$, sodass der Unterschied der Schätzungen der Iterationsschritte nicht gegen β konvergieren, sondern immer weiter anwachsen. Somit geht mindestens eine Variable β_j gegen $\pm\infty$. Bei einer Divergenz und daraus resultierenden Nicht-Konvergenz des Schätzers, kann der Output der Funktion nicht sinnvoll interpretiert werden, da keine sinnvollen Schätzungen für β_j getätigt werden.

Dieser Fall tritt vor allem im Logit-Modell und bei der Complete-Case-Analysis auf, zum Beispiel beim reduzierten Stichprobenumfang auf $n = 250$ oder beim erhöhten erwarteten Anteil von fehlenden Werten. In beiden Fällen trifft zu, dass n im Vergleich zu π reduziert wird. Somit können einige Ergebnisse, bei denen die Complete-Case-Analysis Werte extreme Werte annehmen, nicht berücksichtigt werden.

3.2 Grundmodell

Im ersten Teil der Untersuchung sollen nun einige Parameter hinsichtlich ihres Einflusses auf die Regression untersucht werden. Um die neuen und alten Modelle voneinander unterscheiden zu können und die Notation zu erleichtern, wird im Folgenden unter dem Grundmodell das Modell mit den im folgenden festgelegten Parametern verstanden. Anschließend wird jeweils ein Parameter des Grundmodells variiert, während die anderen Parameter auf den festgelegten Werten festgehalten werden. Es wird zunächst nur ein Datensatz X simuliert. Zwar sind die Ergebnisse somit weniger valide als bei W Ziehungen von X, da diese von der Simulation von X abhängen. Jedoch sind für die Variationen der Parameter immer die gleichen Grundvoraussetzungen bezüglich der Daten geschaffen, sodass eine Variation der Kovariablen die Ergebnisse nicht beeinflussen kann. Eine Betrachtung der Fehler hinsichtlich der Imputationen ist demnach hier noch nicht sinnvoll, da die fehlenden Werte nur einmal generiert werden. Für die Parameter wurden folgende Werte ausgewählt. Die fett hinterlegten Parameter werden zudem in den folgenden Modellen variiert.

$\mathbf{n = 500}$

$\mathbf{S= 50}$

$\theta_1 = 75$

$\theta_2 = 0.4$

$\theta_3 = 0.25$

$\theta_4 = 5$

$\boldsymbol{\sigma = 9}$

$\mu_7 = (-5, 20, -10)$

$$\Sigma_7 = \begin{pmatrix} 1 & 0.3 & 0.1 \\ 0.3 & 1 & 0.2 \\ 0.1 & 0.2 & 1 \end{pmatrix}$$

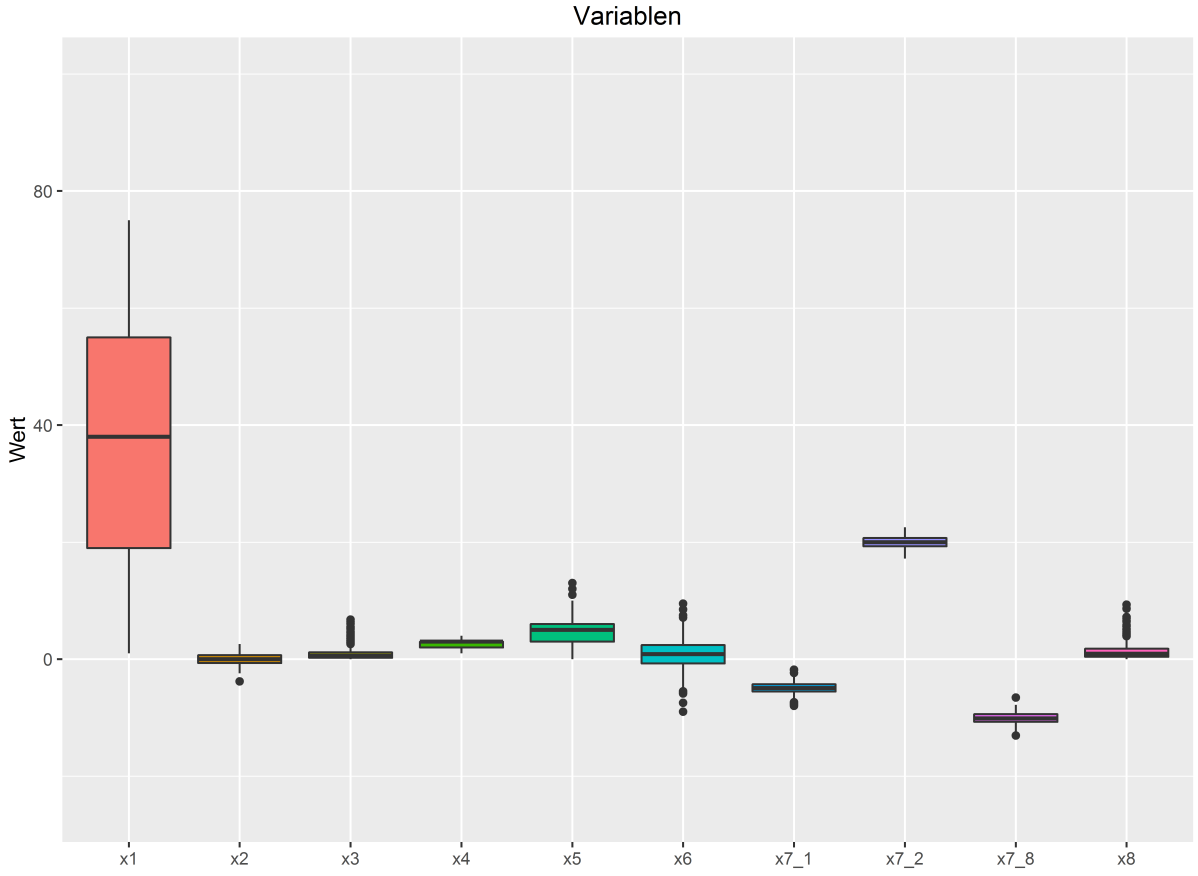
$\boldsymbol{\pi} = \mathbf{0.1}$

Zunächst werden also ausschließlich die Fehler hinsichtlich der Regressionsschätzung betrachtet. Die Faktoren der Linearkombination $\beta_1, \ldots, \beta_q$ werden also über die S Responsevektoren mit den geschätzten Regressionskoeffizienten der vier Regressionsmodelle (je 4 für das lineare und das Logit-Modell) verrechnet.

$$MSE_j = \frac{1}{S} \sum_{s=1}^{S} \left(\beta_j - \hat{\beta}_{j,s} \right)^2$$

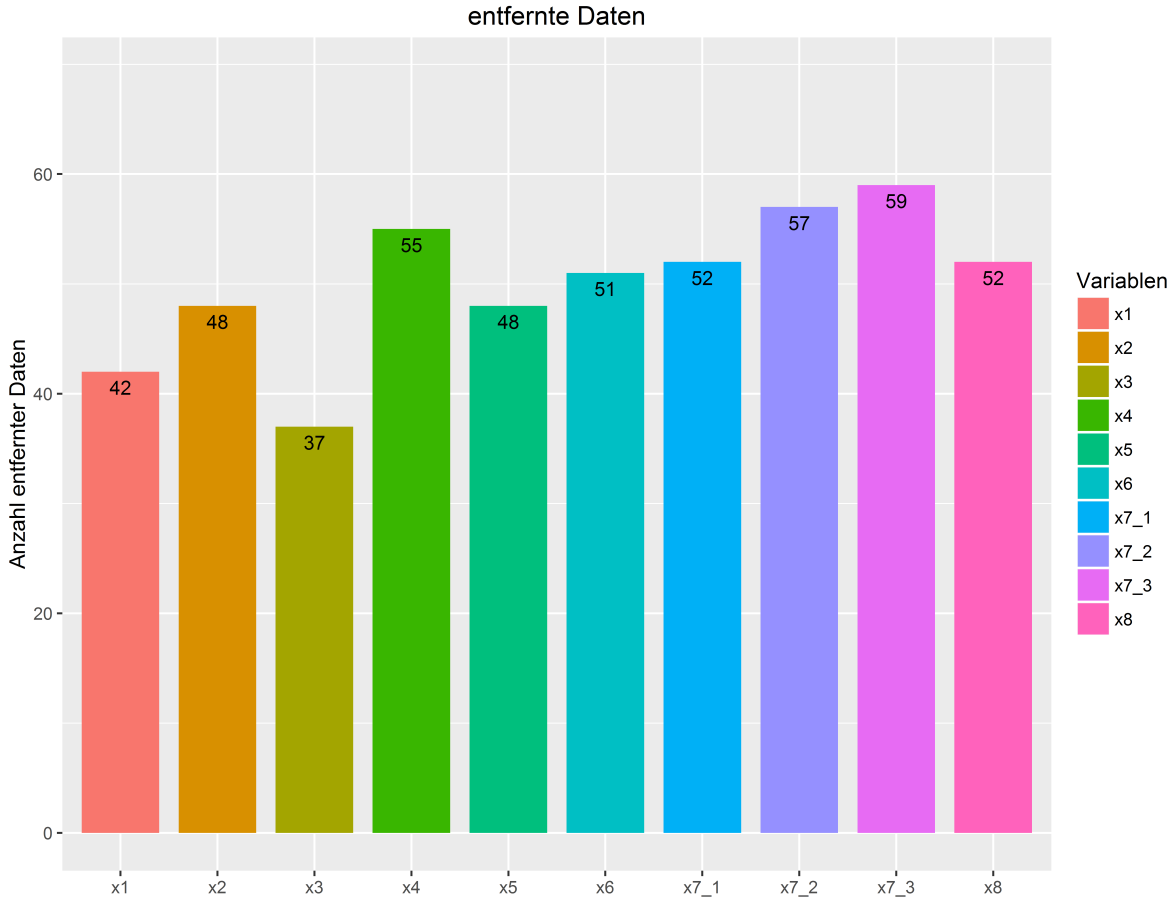
Im Folgenden wurde ein Modell mit den oben genannten Parametern gerechnet, und mehrfach variiert. Die folgenden Grafiken 2 und 3 sollen den dabei entstandenen Datensatz näher beschreiben.

Abbildung 2: Boxplot-Variablen



In Abbildung 2 ist zunächst zu sehen, dass die meisten Variablen Werte im Intervall $[-10, 10]$ um den Nullpunkt annehmen. Für x_1 hingegen liegt der Mittelwert bei ungefähr 35 und nimmt Werte bis über 70 an. Abbildung 3 bestätigt zudem, dass die Anzahl an fehlenden Werten in den Variablen ungefähr um 50 schwankt, wie es auch durch den Parameter π, dem erwarteten Anteil an fehlenden Werten, vorhergesehen ist ($n * \pi = 0.1 * 500 = 50$). Die Anzahl an fehlenden Werten pro Variable schwankt dabei allerdings recht stark, während in x_3 nur 37 Daten nicht beobachtet sind, sind es bei Variable x_{7_3} 59.

Abbildung 3: Anzahl fehlender Werte in Variablen



Wenn im Folgenden von der Varianzkomponente berichtet wird, ist damit der empirische Teil der MSE-Zerlegung, der sich aus dem Mittelwert der quadrierten Differenzen von $\hat{\beta}_{s,j}$ und den Mittelwert der Schätzungen $\bar{\beta}_j$ ergibt, gemeint. Die Biaskomponente beschreibt anschließend die quadrierte Abweichung des Mittelwerts $\bar{\beta}_j$ und dem wahren Parameter β_j. Um die Notation in der anschließenden Untersuchung zu vereinfachen, wird im folgenden Abschnitt unter der Varianzkomponente, der Varianz oder der Streuung beziehungsweise der Biaskomponente, dem Bias oder dem Präzisionsfehler die beiden Summen der MSE-Zerlegung verstanden. Die Tabellen 3 und 4 beschreiben die Varianz- und die Biaskomponente, die aus den Modellen mit den angegebenen Parametern resultieren.

Zunächst fällt auf, dass bezüglich der Varianzkomponente im Allgemeinen das volle Modell erwartungsgemäß am Besten abschneidet. Die Varianzen im Logit-Modell sind im Mittel pro Variable um etwa 20% bei Amelia und 28% bei Random Forest höher als beim vollen Modell, wobei die durchschnittliche Varianz bei der Complete-Case-Analysis den 6.8 fachen Wert beträgt. Für Schätzer β_6 beträgt die Varianzkomponente bei Amelia sogar weniger als im vollen Modell. Die imputierten Daten müssen hier also über die $S = 50$ Ziehungen der Y-Vektoren so simuliert worden sein, dass die Streuung der Koeffizientenschätzer im Vergleich zu den wahren Werten reduziert wurde. Auch im linearen Modell ist die Varianz im Durchschnitt im vollen Modell am kleinsten, allerdings sind hier die Unterschiede deutlich geringer. Die Complete-Case-Analysis verfügt nur über eine im Mittel 2.32 mal höhere Varianzkomponente, während die Varianz bei Random Forest im Durschnitt nur 14% und bei Amelia sogar nur 5.4% höher ist. Obwohl die Differenz der Varianzen durchschnittlich bei Amelia um den Wert 0.004 höhere Werte annimmt, ist die Varianzkomponente in den Koeffizientenschätzern β_1, β_4, β_5 und β_{7_1} niedriger als im vollen Modell. Was die Biaskomponente betrifft, sind im Logit-Modell bei β_2 sowie bei β_4 höhe-

Tabelle 4: Varianzkomponente ,oben: Logit, unten: Linear

	Var_voll	Var_CC	Var_Amelia	Var_RF
β_1	<0.0001	0.0003	<0.0001	<0.0001
β_2	0.0259	0.1862	0.0303	0.0310
β_3	0.0255	0.3623	0.0273	0.0304
β_4	0.0291	0.1553	0.0371	0.0413
β_5	0.0055	0.0226	0.0061	0.0065
β_6	0.0054	0.0225	0.0052	0.0064
β_{7_1}	0.0179	0.1733	0.0263	0.0279
β_{7_2}	0.0060	0.0362	0.0072	0.0076
β_{7_3}	0.0178	0.1472	0.0227	0.0246
β_8	0.0216	0.1239	0.0262	0.0278
β_1	0.0008	0.0013	0.0007	0.0009
β_2	0.1721	0.2824	0.1951	0.2168
β_3	0.1815	0.4042	0.2250	0.2266
β_4	0.2387	0.6351	0.2095	0.2706
β_5	0.0305	0.0543	0.0274	0.0308
β_6	0.0266	0.0341	0.0272	0.0284
β_{7_1}	0.1144	0.2329	0.1380	0.1423
β_{7_2}	0.0309	0.1051	0.0317	0.0306
β_{7_3}	0.0853	0.3704	0.0918	0.0965
β_8	0.0741	0.1674	0.0867	0.0956

Tabelle 5: Biaskomponente,oben: Logit, unten: Linear

	Bias_voll	Bias_CC	Bias_Amelia	Bias_RF
β_1	0.0015	0.0005	0.0013	0.0014
β_2	2.7546	2.1560	2.6573	2.6222
β_3	0.0101	0.0046	0.0075	0.0087
β_4	2.7505	2.6006	2.8534	2.7574
β_5	0.3155	0.2461	0.3078	0.2999
β_6	0.0975	0.0777	0.0895	0.0861
β_{7_1}	0.2764	0.3227	0.2670	0.2627
β_{7_2}	0.5211	0.4625	0.5440	0.5511
β_{7_3}	0.3523	0.3213	0.3249	0.3132
β_8	0.1311	0.0640	0.1334	0.1358
β_1	0.0000	0.0000	0.0000	0.0000
β_2	0.0001	0.0128	0.0000	0.0074
β_3	0.0050	0.0398	0.0080	0.0019
β_4	0.0023	0.0060	0.0077	0.0101
β_5	0.0005	0.0010	0.0014	0.0048
β_6	0.0029	0.0028	0.0089	0.0114
β_{7_1}	0.2626	0.3305	0.2074	0.1898
β_{7_2}	0.5334	0.4621	0.6095	0.6128
β_{7_3}	0.4836	0.6313	0.4325	0.4099
β_8	0.0005	0.0024	0.0015	0.0003

re Verzerrungen in allem Regressionsmodellen zu sehen. Auffällig ist weiterhin, dass bei dieser Untersuchung im Logit-Modell bezüglich des Bias das volle Modell, wenn man die durchschnittliche Biaskomponente pro Variable betrachtet, schlechter abschneidet als alle anderen Methoden,

sowohl die Complete-Case-Analysis als auch die imputierten Datensätze. Auch bei komponentenweiser relativer Betrachtung sind ähnliche Effekte zu sehen. Die Biaskomponente ist in der Complete-Case-Analysis um etwa 25%, in Amelia und Random Forest noch um etwa 5% geringer als im vollen Modell. Somit verbessert sich die Schätzung hinsichtlich ihrer Präzision in diesem Fall, wenn man die fehlen Werte ignoriert, wobei eine Imputation hier sogar noch besser abschneidet als die Verwendung der ursprünglich wahren Daten. Die Daten, die bei dieser Simulation von X fehlen, tragen demnach negativ zur Schätzung von Y bei, sodass eine nicht Berücksichtigung und somit eine Reduzierung des Stichprobenumfangs, zur Verbesserung der Präzision führt. Eine weitere mögliche Erklärung wäre, dass die übrig gebliebenen Beobachtungen alle günstig bezüglich der wahren Schätzer β liegen. Im linearen Modell sind die Biaskomponenten sehr klein, weshalb es zu großen relativen Verhältnissen verglichen mit dem vollen Modell kommt. Während der Bias also zum Beispiel in β_2 um den Faktor 167.2 mal höher ausfällt als im vollen Modell, beträgt dieselbe Biaskomponente bei Amelia nur 9% der Verzerrung des vollen Modells. Im Durchschnitt über alle Variablen ist der Bias allerdings in der Complete-Case-Analysis um den Faktor 19.3 (vor allem aufgrund der Schätzung von β_2), bei Amelia um den Faktor 2.2 und bei Random Forest um den Faktor 12 mal größer als beim vollen Modell. Betrachtet man die absoluten Verzerrungen gegenüber dem vollen Modell, beträgt die Differenz bei der Complete-Case-Analysis 0.02 im Mittel zugunsten dem vollen Modell, während Amelia um 0.001 und Random Forest um 0.004 besser abschneidet. Allerdings sind, wie bereits gesagt, bei solch geringen Biaskomponente die Ergebnisse stark abhängig von einzelnen (ungünstigen) ungenaueren Schätzungen, die bei sonst geringer Verzerrung einen größeren Einfluss verfügen. Im Folgenden sollen die Ergebnisse dieses Durchgangs mit den Resultaten verglichen werden, die durch einzelne Variationen der Parameter entstehen. Dabei werden die einzelnen neuen Modelle jeweils vor allem hinsichtlich absoluter Differenz und relativem Unterschied verglichen. Unter Differenz wird im Folgenden die absolute Differenz zwischen dem neuen Modell mit variierendem Parameter und Grundmodell verstanden. Eine positive Differenz zeigt demnach auf, dass sich die Fehlerkomponente im Vergleich zum gerade beschriebenen Modell vergrößert haben, und das Modell somit schlechter abschneidet. Da eine reine Betrachtung der Differenzen nicht mitberücksichtigt, in welchem Verhältnis sich die Variablen bezüglich der Fehlerkomponente im Grundmodell verändert haben, wird neben der Differenz auch noch der relative Unterschied gemessen. Somit wird das Verhältnis der entsprechenden Fehlerkomponente des Modells mit den variierenden Parametern zu dem Modell mit den ursprünglich festgelegten Parametern bestimmt. Ein Verhältnis von a impliziert somit eine um $((a-1)*100)\%$ höhere Fehlerkomponente durch die Variation. Falls in der Analyse der Varianz- und Biaskomponenten von gemittelten oder durchschnittlichen Ergebnissen berichtet wird, ist damit das arithmetische Mittel über alle Koeffizientenschätzer $\beta_1, \ldots, \beta_8$ beziehungsweise den entsprechenden Variablen $x_1, \ldots, x_8$ gemeint.

3.3 Variationen

3.3.1 Stichprobenumfang

Zuerst soll der Einfluss des Stichprobenumfangs auf die MSE-Komponenten untersucht werden. Die Ergebnisse bezüglich der MSE-Komponenten sind in folgender Tabellen 6 und 7 noch einmal zusammengefasst. Wie in Kapitel 2.1 beschrieben, zeigt folgende Grafik die für die einzelnen Koeffizientenschätzer $\{\beta_1, \ldots, \beta_8\}$ durchschnittlichen absoluten Differenzen des variierenden und des Grundmodells auf. Weiterhin werden auch die relativen Unterschiede angegeben, die sich aus dem Verhältnis des variierenden und des Grundmodells ergeben.

Im Logit-Modell fallen die extrem hohen Varianzkomponenten beim Stichprobenumfang $n = 250$ der Complete-Case-Analysis auf. Aufgrund der bereits erläuterten Probleme kommt es hier zu einer enormen Vezerrung der Schätzer, weshalb diese hier nicht sinnvoll interpretiert werden können. Ignoriert man diese Modelle, steigt die Varianzkomponente in den anderen Regressionmodelle ca. um 0.035 im Mittel an, was einer durchschnittlich um den Faktor 3 erhöhten Va-

Tabelle 6: Stichprobenumfang-Varianzkomponente

	$n = 250$ absolut	$n = 250$ relativ	$n = 1000$ absolut	$n = 1000$ relativ
voll-log	0.0279	2.6904	-0.0074	0.5209
CC-log	*76435.7508*	*597301.9155*	-0.0777	0.4534
Amelia-log	0.0470	3.2870	-0.0090	0.5300
RF-log	0.0436	3.0240	-0.0099	0.5201
voll-lin	0.1471	2.5078	-0.0433	0.5843
CC-lin	0.3087	2.6177	-0.0926	0.6414
Amelia-lin	0.1886	2.9071	-0.0426	0.6238
RF-lin	0.1898	2.8130	-0.0445	0.6448

Tabelle 7: Stichprobenumfang-Bias

	$n = 250$ absolut	$n = 250$ relativ	$n = 1000$ absolut	$n = 1000$ relativ
voll-log	-0.1442	0.8368	0.0236	1.0502
CC-log	*2021.6288*	*23988.9578*	-0.0002	1.3909
Amelia-log	-0.1816	0.7988	0.0348	1.0374
RF-log	-0.1556	0.7800	0.0341	1.0572
voll-lin	-0.0088	5.9320	0.0098	0.9160
CC-lin	-0.0152	1.4314	-0.0108	0.5911
Amelia-lin	-0.0029	47.8816	0.0139	19.4256
RF-lin	0.0024	1.3747	0.0148	1.6302

rianzkomponente entspricht. Aufgrund der Reduzierung des Stichprobenumfangs von $n = 500$ auf die Hälfte, hat sich die jeweiligen Varianzen also durchschnittlich ungefähr verdreifacht. Für die linearen Regressionsmodelle treten ähnliche Ergebnisse auf, wobei hier die Complete-Case-Analysis interpretiert werden kann. Die Varianzkomponente steigt ebenfalls durchschnittlich bei allen Regressionsmodellen an, allerdings ca. um 0.15 bei dem vollen Modell, und um ungefähr 0.18 bei Random Forest und Amelia, und um 0.31 bei der Complete-Case-Analysis. Dies entspricht wiederum einer Steigerung um durchschnittlich den Faktor 2.7. Bei Erhöhung des Stichprobenumfangs erzielt das Logit-Modell recht ähnliche Ergebnisse. Die Varianz sinkt in allen Modellen in etwa durchschnittlich um 50%; die Differenzen betragen bei der Complete-Case-Analysis ungefähr 0.08, und in den übrigen Modellen im Mittel etwa 0.01. Während sich bei der Verdoppelung von $n = 250$ auf $n = 500$ im Logit-Modell die Varianzkomponente im Mittel noch verdreifacht hat, ist sie bei einer weiteren Erhöhung auf $n = 1000$ nochmal auf das doppelte angestiegen. Im linearen Fall sind weiterhin ähnliche Effekte aufzufinden, wenn auch hier die Varianzkomponente nur um ungefähr 40% sinkt. Für die Complete-Case-Analysis reduziert sich somit die Streuung durchschnittlich um 0.095 und bei den anderen Modellen im Mittel um ca. 0.045. Wiederum fällt auf, dass auch bezüglich der Biaskomponente extreme Werte in der Complete-Case-Analysis ergeben, sodass auch diese Werte nicht sinnvoll analysiert werden können. Weiterhin sind die Resultate bezüglich des Bias deutlich weniger konstant als die entsprechenden Varianzkomponenten. Beim reduzierten Stichprobenumfang vermindert sich der Bias der übrigen Modelle im Logit-Modell gegenüber dem Grundmodell um ca. durchschnittlich 20% und um die durchschnittliche Differenz von ungefähr 0.16. Vor allem in β_4 und in β_2 treten Verbesserungen von ca. 0.9 beziehungsweise 0.7 auf, was wiederum einer Verbesserung von 25% beziehungsweise 30% gegenüber dem Grundmodell entspricht. Auch die linearen Regressionsmodelle zeigen im Allgemeinen leichte Verbesserungen bezüglich des Bias auf, wenn man die durchschnittlichen absoluten Differenzen in Betracht zieht. Allerdings fällt vor allem β_2 auf, da der Bias hier gegenüber dem Grundmodell im vollen Modell um den Faktor 41 und bei Amelia um den Faktor 470 steigt, was auch gemittelt über alle Koeffizientenschätzer im vollen Modell zu einer verhältnismäßigen

Erhöhung des Bias um Faktor 5.93 und bei Amelia um den Faktor 47.88 führt. Diese sehr hohen relativen Veränderungen liegen aber auch an den sehr niedrigen Biaskomponenten, sodass kleine Veränderungen großen verhältnismäßige Steigerungen bedeuten. Die absolute Veränderung liegen bei β_2 nur jeweils bei 0.003.

Bei Erhöhung des Stichprobenumfangs auf $n = 1000$ sind ebenfalls nur bedingt die erwarteten Ergebnisse eingetroffen. Obwohl der Stichprobenumfang, und damit erwartungsgemäß eigentlich auch die Genauigkeit der Schätzungen, erhöht wurde, nimmt die Verzerrung der Koeffizientenschätzer im Logit-Modell durchschnittlich in allen Modellen zu; bei der Complete-Case-Analysis um 40% und bei den anderen Regressionsmodellen ungefähr durchschnittlich um ungefähr 5% . Wiederum sind die absoluten Differenzen gegenüber dem Grundmodell sehr gering, und betragen im Mittel maximal 0.035 bei Amelia, wohingegen in der Complete-Case-Analysis deer Bias bei erhöhtem Stichprobenumfang um 0.0002 im Mittel geringer ausfällt. Im linearen Modell hingegen sind die Veränderungen gegenüber dem Grundmodell nicht konstant. Während sich das volle Modell um durchschnittlich ungefähr 9% verbessert (aber absolut im Mittel um 0,01 an Verzerrung zunimmt), beträgt die Verzerrung bei der Complete-Case-Analysis mehr als 40% weniger (und nimmt im Mittel um 0.011 ab). Für Amelia beträgt die Biaskomponente im Mittel den 20 fachen Wert. Wiederum ist dies auf Variable β_2 zurückzuführen, bei der die Verzerrung um den Faktor 186.42 steigt, was aber nur einer absoluten Differenz von unter 0.0013 entspricht. Ein weiterer Anstieg um 64% im Mittel tritt hier bei Random Forest auf, was im Durchschnitt einer erhöhten Verzerrung von 0.015 entspricht.

Beim verglichenen Abschneiden von Amelia und Random Forest spielt der Stichprobenumfang eine Rolle. Die anschließende Tabelle 8 vergleicht die absoluten und relativen Unterschiede von Random Forest und Amelia. Positive Werte bei den Differenzen und Werte über 1 bezüglich der relativen Veränderungen bedeuten somit, das Random Forest über eine höhere Fehlerkomponente verfügt. Die mittleren beiden Spalten beziehen sich in diesen wie auch allen weiteren Tabellen zum Vergleich von Random Forest und Amelia auf das Grundmodell und bleiben somit in allen Variationen konstant. Zur besseren Übersicht von fortlaufenden Veränderungen bleiben diese Werte trotzdem in den Tabellen.

Tabelle 8: Differenz und Verhältnis von Random Forest und Amelia

	n_{250}	n_{500}	$n = 1000$
Varianzkomponente			
$logit_{absolut}$	-0.0020	0.0015	0.0006
$logit_{relativ}$	-0.0020	0.0015	0.0006
$lin_{absolut}$	0.0118	0.0106	0.0087
$lin_{relativ}$	0.0118	0.0106	0.0087
Biaskomponente			
$logit_{absolut}$	0.0113	-0.0148	-0.0154
$logit_{relativ}$	0.8935	1.0033	1.0336
$lin_{absolut}$	0.0024	-0.0029	-0.0020
$lin_{relativ}$	8.3560	106.9827	2.6196

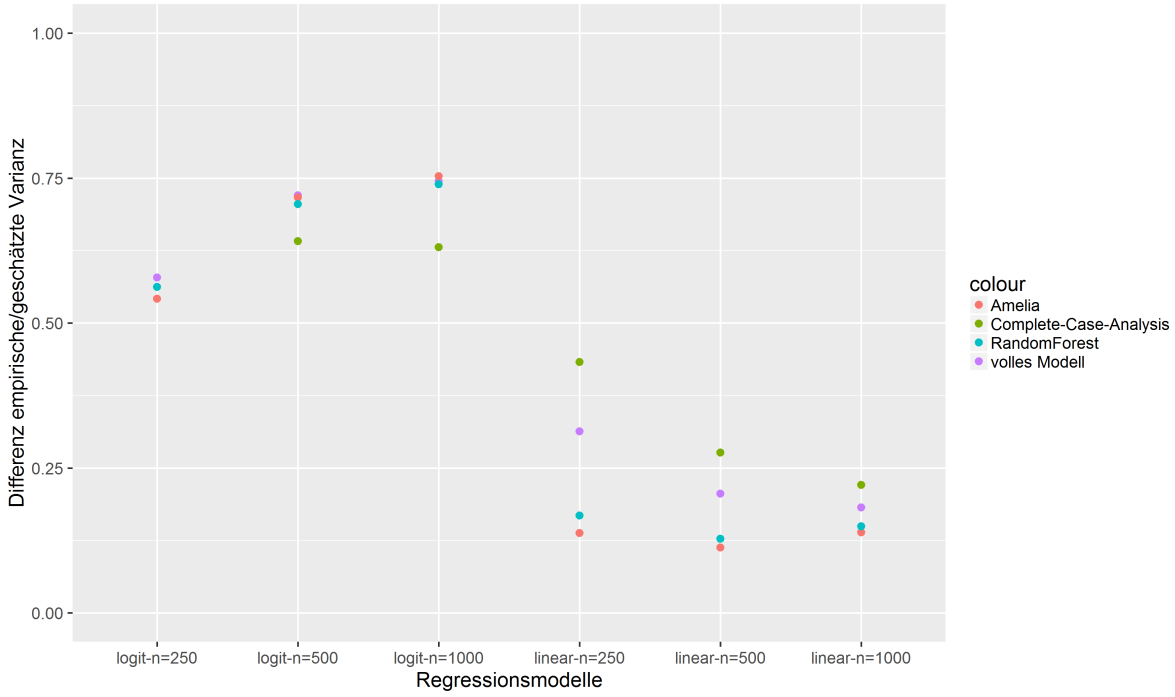
Für das Modell mit $n = 250$ schätzt Random Forest die Varianzkomponente um 0.002 genauer im Logit Modell, während im linearen Modell Amelia um 0.012 geringere Varianzkomponente aufweist. In den übrigen Modellen weist Amelia sowohl im linearen als auch im Logit-Modell geringere Varianzkomponenten auf. Auch mit Bezug zum mittleren Verhältnis der einzelnen Variablen sind nur geringe Unterschiede feststellbar, beim kleinsten Stichprobenumfang liegt die Varianzkomponente im Logit-Modell bei Random Forest fast 1% unter Amelia, wobei Amelia

für $n = 500$ eine im Mittel um 5% und für $n = 1000$ eine im Mittel um 9% geringere Streuung aufweist.

Während bei $n = 250$ Amelia im Durchschnitt bezüglich der Biaskomponente noch um 0.011 im Logit- und 0.002 im linearen Modell besser als Random Forest abschneidet, erzielt Random Forest für $n = 500$ ein um 0.014 beziehungsweise 0.003 geringeres Ergebnis, wobei sich für das Ergebnis für $n = 1000$ ähnliche Werte (-0.015 und -0.002) ergeben. Die Ergebnisse sind demnach genau andersherum, als bei der Varianzkomponente. Betrachtet man die relativen Veränderung pro Variable, fällt auf, dass sich vor allem im linearen Modell bei Random Forest eine deutliche höhere durchschnittliche Verzerrung zeigt. Bei $n = 1000$ ergibt sich im Vergleich zu den Amelia-Varianzkomponenten somit ein um den Faktor 2.62 , für $n = 250$ ein um den Faktor 8.35 und bei $n = 500$ sogar ein um Faktor den 106.98 höheren Bias. Die starke verhältnismäßige Steigerung ergibt sich aufgrund dem Schätzer für β_2, bei dem der Präzisionsfehler bei RandomForest 1060.19 fach größer ist als bei Amelia. Bei x_2 handelt es sich um die Standard-Normalverteilte Variabel, welche auch für die anderen linearen Modelle ($n = 250 : +821\%, n = 1000 : +543\%$) um deutlich höhere Verzerrungen annimmt. Allerdings kann die verbesserte Performance Amelias bei normalverteilten Daten im linearen Modell nicht an Variable x_7 bestätigt werden. Obwohl x_7 ebenfalls einer (in diesem Fall multivariaten) Normalverteilung folgt, schneidet Random Forest bezüglich der Biaskomponente besser ab als Amelia.

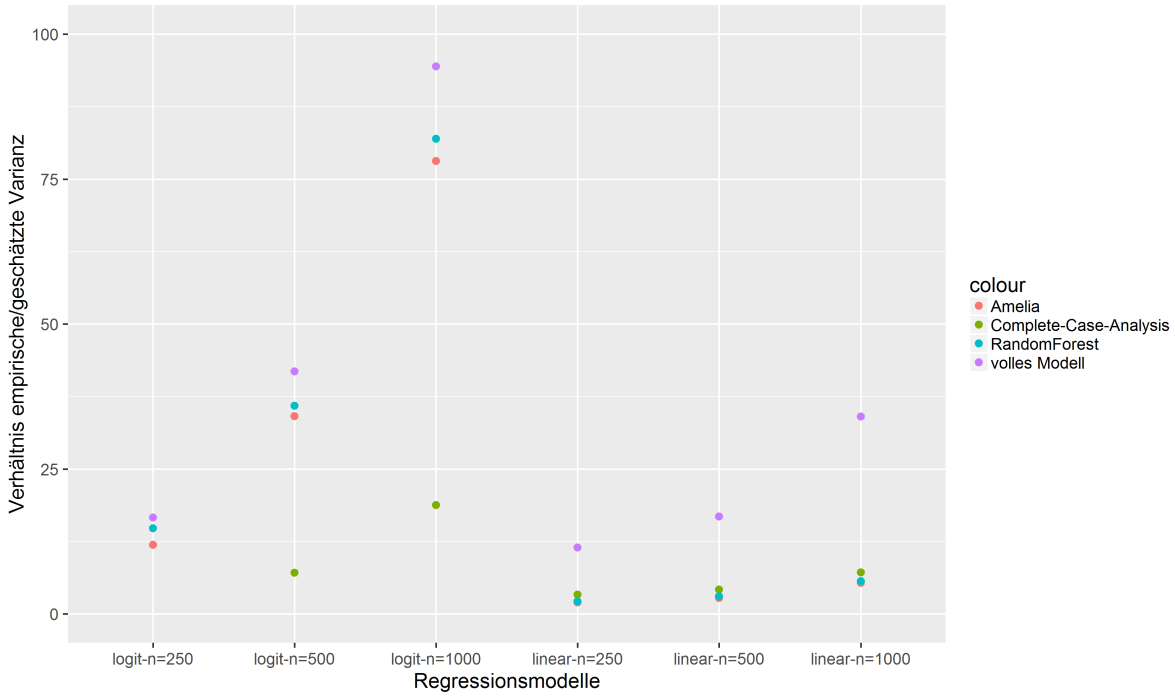
Neben den MSE-Komponenten der Koeffizientenschätzer sollen auch die empirischen und geschätzten Varianzen und deren Veränderung bei Variation der Parameter untersucht werden. Die folgenden beiden Grafiken Abbildung 4 und Abbildung 5 zeigen sowohl die Differenz als auch das Verhältnis des empirischen MSEs und der geschätzten Varianz der Regressionsmodelle.

Abbildung 4: Varianzvergleich-Differenzen



Da im Logit-Modell beim Stichprobenumfang von 250 das Complete-Case-Regressionsmodell nicht konvergiert ist, wird diese Differenz nicht berücksichtigt. Ansonsten ist aber zu sehen, dass alle Differenzen positiv sind. Da das volle Modell fast die gleichen Differenzen wie Amelia und Random Forest aufweist, ist in Abbildung 5 der Punkt nur schwer zu erkennen. Bei Anstieg des Stichprobenumfangs ist zu sehen, dass die Differenz der Varianzen im Logit-Modell von ca. 0.5 auf ungefähr 0.7 ansteigt. Während die Complete-Case-Analysis im Logit-Modell, soweit berücksichtigt, mit der geringsten Differenz eingeschätzt wird, ist die Differenz im linearen Modell am

Abbildung 5: Varianzvergleich-Verhältnisse



größten. Die Differenzen im linearen Modell sind im allgemeinen geringer als beim Logit-Modell. Während zudem bei Logit-Modell mit steigendem Stichprobenumfang die Differenzen zunahmen, reduzieren sich die Diffferenzen der Varianzen beim linearen Modell bei ansteigendem n. Auch bei Betrachtung der Verhältnisse der empirischen und der geschätzten Varianzen kann man eine generelle Unterschätzung des MSEs feststellen. Im Logit-Modell steigt die Überschätzung des MSEs mit ansteigendem Stichprobenumfang von ungefähr 20% für $n = 250$ auf im ungefähr 70%, wobei wieder die Complete-Case-Analysis die geringsten Differenzen aufweist. Bei Anstieg des Stichprobenumfangs von $n = 250$ auf $n = 1000$ steigen auch die Verhältnisse der Varianzen im Schnitt auf die ungefähr 2.5 fache Verhältnis an, wobei das volle Modell eine etwa 6 mal höhere Unterschied der Verhältnisse aufweist.

Im Allgemeinen kann beim Stichprobenumfang bezüglich der Varianzkomponente sowohl im linearen als auch im Logit-Modell ein Einfluss auf die Schätzung der Koeffizietenschätzer festgestellt werden. Zwischen dem Bias der Schätzungen und dem Stichprobenumfang zeigen die Daten zumindest keinen eindeutigen Zusammenhang auf, die Biaskomponente nimmt bei den hier vorliegenden Daten bei steigendem Stichprobenumfang sogar eher zu. Im Vergleich von Amelia und Random Forest ist Amelia in der Varianzkomponente meistens genauer, während Random Forest eine geringere Biaskomponente aufweist. Random Forest schneidet im Allgemeinen zudem beim Logit-Modell besser ab, wohingegen Amelia vor allem bei steigendem Stichprobenumfang geringere Fehlerkomponenten aufweist. An dem Vergleich der Varianzen lässt sich vor allem feststellen, dass mit zunehmendem Stichprobenumfang die relativen Unterschiede der Schätzungen steigen, wobei vor allem im Logit-Modell die empirische Varianz generell höher ist als die geschätzte Varianz. Beim vollen Modell ist der Unterschied der Verhältnisse am höchsten, während die Complete-Case-Analysis im Logit-Modell die geringsten Differenzen zwischen geschätzter und empirischer Varianz aufweist.

3.3.2 Erwarteter Anteil fehlender Werte

Im nächsten Abschnitt soll der Anteil an erwarteten fehlenden Werten betrachtet werden, welcher von den $\pi = 0.1$ auf $\pi = 0.05$ vermindert und auf $\pi = 0.25$ erhöht wird. Im Folgenden sind daher

wieder die Tabellen der mittleren absoluten und relativen Veränderungen (Tabelle 9 und Tabelle 10) bezüglich der Vazianz- und Biaskomponente gelistet.

Tabelle 9: erwarteter Anteil fehlender Werte-Varianzkomponente

	$\pi = 0.05$ absolut	$\pi = 0.05$ relativ	$\pi = 0.25$ absolut	$\pi = 0.25$ relativ
voll-log	0.0165	1.9374	-0.0056	0.5989
CC-log	-0.0157	1.0084	*18391.2912*	*180201.4334*
Amelia-log	0.0172	1.8537	0.0007	0.9698
RF-log	0.0155	1.7165	0.0041	1.1688
voll-lin	0.0009	1.0759	0.0139	1.1144
CC-lin	-0.0444	0.9411	0.3524	2.4331
Amelia-lin	0.0017	1.0882	0.0771	1.5762
RF-lin	-0.0065	1.0365	0.1199	1.9574

Tabelle 10: erwarteter Anteil fehlender Werte-Biaskomponente

	$\pi = 0.05$ absolut	$\pi = 0.05$ relativ	$\pi = 0.25$ absolut	$\pi = 0.25$ relativ
voll-log	-0.0673	0.9065	-0.1701	0.9362
CC-log	-0.0665	1.0973	*1680.3258*	*20773.4474*
Amelia-log	-0.0567	0.9308	-0.2240	0.9282
RandomForest-log	-0.0497	0.9384	-0.2448	1.0941
voll-lin	0.0030	7.7946	-0.0090	5.4924
CC-lin	-0.0159	0.9382	-0.0239	2.2026
Amelia-lin	-0.0018	95.5301	-0.0010	243.5382
RandomForest-lin	0.0012	0.9168	0.0319	8.5574

Da im vollen Modell keine fehlenden Werte vorliegen, sollte sich eigentlich an den Schätzungen zum Grundmodell keine wesentlichen Änderungen aufzeigen. Zur Anschauung und zum Vergleich mit den Veränderungen in den anderen Modellen, sind die Unterschiede im vollen Modell dennoch aufgeführt. Zunächst ist feststellbar, dass auch hier bei einer Simulation die Complete-Case-Ergebnisse extrem über den anderen Ergebnissen liegen. Allerdings ist, bei $\pi = 0.25$ die Wahrscheinlichkeit einer Beobachtung, für keine der 10 Variablen fehlende Werte zuhaben, auch sehr gering (bei 10 Variablen und einer durchschnittlich erwarteten Wahrscheinlichkeit $1 - p$, dass die fehlende Variable j der Beobachtung i vorhanden ist: $0.75^{10} \sim 6\%$), wodurch der Stichprobenumfang sich deutlich reduziert, und n im Vergleich zu q zu klein wird. Aufgrund der Nicht-Konvergenz wird die Complete-Case-Analysis im Logit-Modell nicht berücksichtigt Während im Logit-Modell für $\pi = 0.05$ relativ gesehen alle Varianzkomponenten zunehmen, veringert sich nur die mittlere absolute Differenz im Complete-Cases Modell um 0.016 (, wohingegen die anderen Regressionsmodelle ungefähr durchschnittlich um den gleichen Betrag zunehmen). Die relative Erhöhung der Streuung im Complete-Case-Analysis-Modell beträgt allerdings durchschnittlich 0.1%, wobei sich die Streuung bei Random Forest um 72%, bei Amelia um 85% und beim vollen Modell um 94% steigert. Was die Präzision der Koeffizientenschätzer betrifft, verbessert sich die Biaskomponente in allen Regressionsmodellen im Durchschnitt jeweils um ungefähr 0.06. Während bei der Complete-Case-Analysis, was die relativen Veränderungen betrifft, eine Erhöhung des Bias um ca 10% im Mittel auftritt, verringert sich der Bias ansonsten um durchschnittlich zwischen 7 und 10% . Hinsichtlich dem linearen Model sind nur geringere Veränderungen feststellbar. Während bei der Complete-Case-Analysis eine Verminderung der Varianzkomponente von durchschnittlich 6% erzielt wird(absolut: um 0.044), erhöhen sich die übrigen Modelle bezüglich relativer Abweichung gegenüber dem Grundmodell. Bei Random Forest sinkt die durchschnittliche Varianzkomponente zwar um 0.007, im durchschnittlichen Verhältnis der Varianzkomponenten schneidet das Modell mit $\pi = 0.05$ allerdings um ca. 4% schlechter ab.

Bezüglich der Biaskomponente fällt vor allem die im Mittel um 95.5 fach höheren Biaskomponenten bei Amelia auf, wobei hinsichtlich der absoluten Differenzen die Präzision im Durchschnitt um 0.002 genauer ist. Der hohe Faktor kann wiederum durch β_2 erklärt werden, die für $\pi = 0.05$ zwar nur um 0.007 höher ist, was aber einer Erhöhung um den Faktor 952 gegenüber dem Grundmodell entspricht. Auch im vollen Modell führt eine um 0.005 höhere Verzerrung in β_2 zu einer 6657% Erhöhung der Biaskomponente, weshalb auch beim vollen Modell die relative Erhöhung um 679% recht hoch ausfällt. Bei der Complete-Case-Analysis wie auch beim Random Forest Modell nimmt die Biaskomponente im Vergleich zum ursprünglichen Modell jeweils um ca. 8% im Durchschnitt ab.

Bei einer Erhöhung des erwarteten Anteils der fehlenden Werte auf 25% kann, wie bereits erläutert, die Complete-Case-Analysis nicht berücksichtigt werden. Ansonsten fällt beim Logit-Modell auf, dass im Vergleich der Varianzkomponenten zum Modell mit $\pi = 0.1$ nur sehr geringe absolute Differenzen auftreten. Im vollen Modell ergibt sich eine im Mittel um 0.006 geringere Varianzkomponente, allerdings sind die einzelnen Varianzkomponenten im Schnitt um ca. 40% geringer. Während bei Amelia die Unterschiede sowohl relativ als auch absolut nur gering sind, ergibt das Random Forest Modell eine um pro Variabel im Durchschnitt um 0.004 beziehungsweise um 17% höhere Varianzkomponente. Allerdings verbessern sich alle Regressionsmodelle bezüglich der Biaskomponente im Logit-Modell um ungefähr 0.2, was einer durchschnittlichen Verbesserung von 6% im vollen Modell und von 7% bei Amelia entspricht. Bei Random Forest ist zwar im Mittel eine um 10% höhere Biaskomponente beobachtbar, allerdings erhöht sich der Bias nur in 3 Schätzern ($\beta_4, \beta_{7_1}, \beta_{7_3}$). Im linearen Modell entsprechen die Ergebnisse für $\pi = 0.25$ eher den instinktiven Erwartungen. In allen Modellen steigt die Varianz gegenüber dem Modell mit $\pi = 0.1$ im Durchschnitt an, bei der Complete-Case-Analysis sogar um 0.35. Betrachtete man die relativen Verbesserungen, ist auch dort eine Erhöhung der Varianzkomponente erkennbar (volles Modell: +11%, CC: +143%, Amelia: +58%, Random Forest:+96%). Bezüglich der Biaskomponente wird die Präzision jedoch im vollen Modell um 0.009, bei der Complete-Case-Analysis um 0.024 und bei Amelia um 0.001 verbessert. Bezüglich der relativen Veränderungen der Biaskomponente steigt die Verzerrung jedoch beim vollen Modell um den Faktor 5.49, bei der Complete-Case-Analysis um den Faktor 2.20 und bei Amelia sogar um den Faktor 243. Wiederum fällt vor allem β_2 aufgrund der hohen relativen Steigerungen auf; die Biaskomponenten im Vergleich mit dem Modell der ursprünglichen Simulation beim vollen Modell, Amelia und Random Forest ungefähr um die Faktoren 28, 2425 und 25, was zu den hohen relativen Verhältnissen führt.

Im Folgenden soll wiederum Random Forest mit Amelia verglichen werden. Die anschließende Tabelle 11 bildet die durchschnittlichen Differenzen und Verhältnissen von Random Forest und Amelia in den Modellen ab.

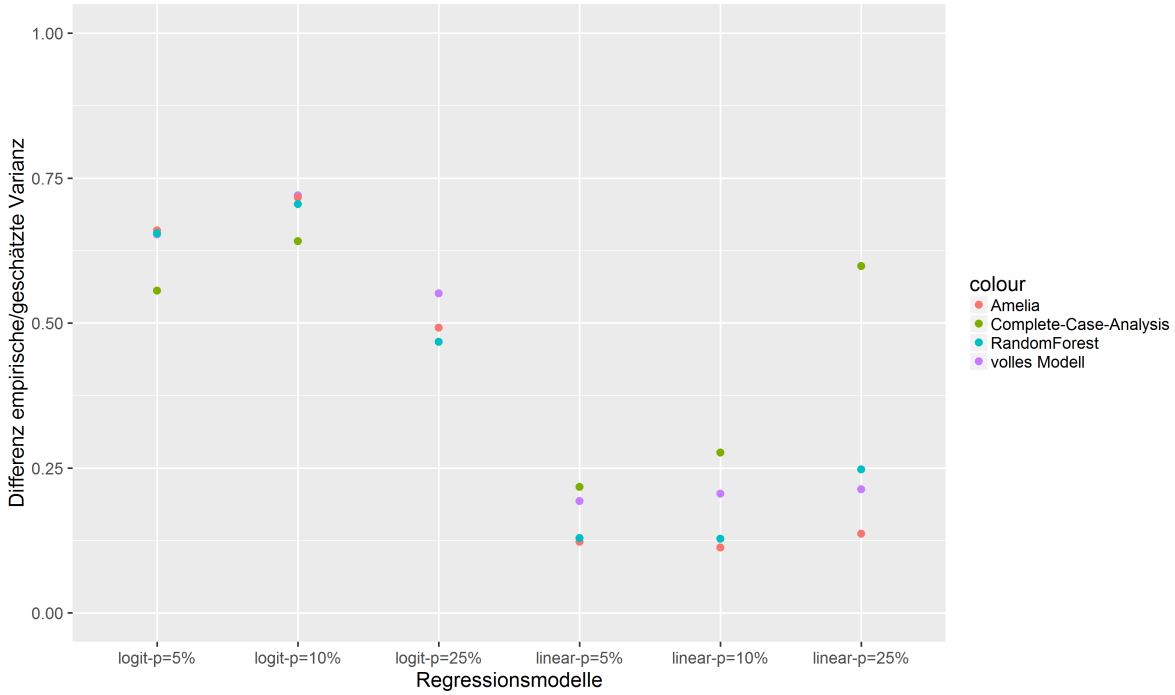
Tabelle 11: Differenz und Verhältnis von Random Forest und Amelia

	$\pi_{0.05}$	$\pi_{0.1}$	$\pi 0.25$
Varianzkomponente			
$logit_{absolut}$	-0.0003	0.0015	0.0049
$logit_{relativ}$	0.9855	1.0722	1.3404
$lin_{absolut}$	0.0024	0.0106	0.0087
$lin_{relativ}$	1.0372	1.0936	1.3663
Biaskomponente			
$logit_{absolut}$	-0.0078	-0.0148	-0.0356
$logit_{relativ}$	0.9957	1.0033	1.0935
$lin_{absolut}$	0.0002	-0.0029	0.0301
$lin_{relativ}$	379.3109	106.9827	31.5157

Bezüglich der Varianzkomponente liegt nur im Logit-Modell für $\pi = 0.05$ eine höhere Streuung in Amelia vor, während Amelia ansonsten, was sowohl die durchschnittlichen Differenzen als auch die durchschnittlichen Verhältnisse betrifft, besser abschneidet. Zudem fällt sowohl im linearen als auch im Logit-Modell auf, dass sowohl der relative als auch der absolute Unterschied zwischen Amelia und Random Forest mit zunehmenden π zunimmt, sodass Amelia mit zunehmendem π im Vergleich mit Random Forest immer besser abschneidet. Was den Bias der beiden Modelle angeht, fällt jedoch auf, dass sich im Logit-Modell die Differenz der Varianzkomponenten von Amelia im Vergleich zu Random Forest zwar von -0.003 auf 0.0049 verbessert, gleichzetig die Biaskomponente sich von -0.0078 auf -0.0356 verschlechtert. Somit beträgt die Differenz, um die der empirische MSE bei Amelia gegenüber Random Forest höher ist, für $\pi = 0.05$ im Mittel -0.008, sodass hier Random Forest besser abschneidet. Für das Modell mit dem größten Anteil an fehlenden Werten, beträgt diese durchschnittlichere Differenz ungefähr 0.03, sodass hier Amelia den geringeren Fehler aufweist. Für β_6 ist zudem die Biaskomponente im linearen Modell bei $\pi = 0.05$ um 3678 mal höher als bei Amelia, was zu der durchschnittlich um Faktor 379.31 höheren Biaskomponente im selben Modell führt. Weiterhin wird β_6 auch im linearen Modell mit $\pi = 0.25$ deutlich unpräziser (um Faktor 220) bei Random Forest geschätzt, was die auch hier hohe Biaskomponente erklärt. x_6 folgt wie auch x_2, welches im Grundmodell für den um den Faktor 106.98 höhere Biaskomponente des Random Forest Modells sorgt, einer Normalverteilung.

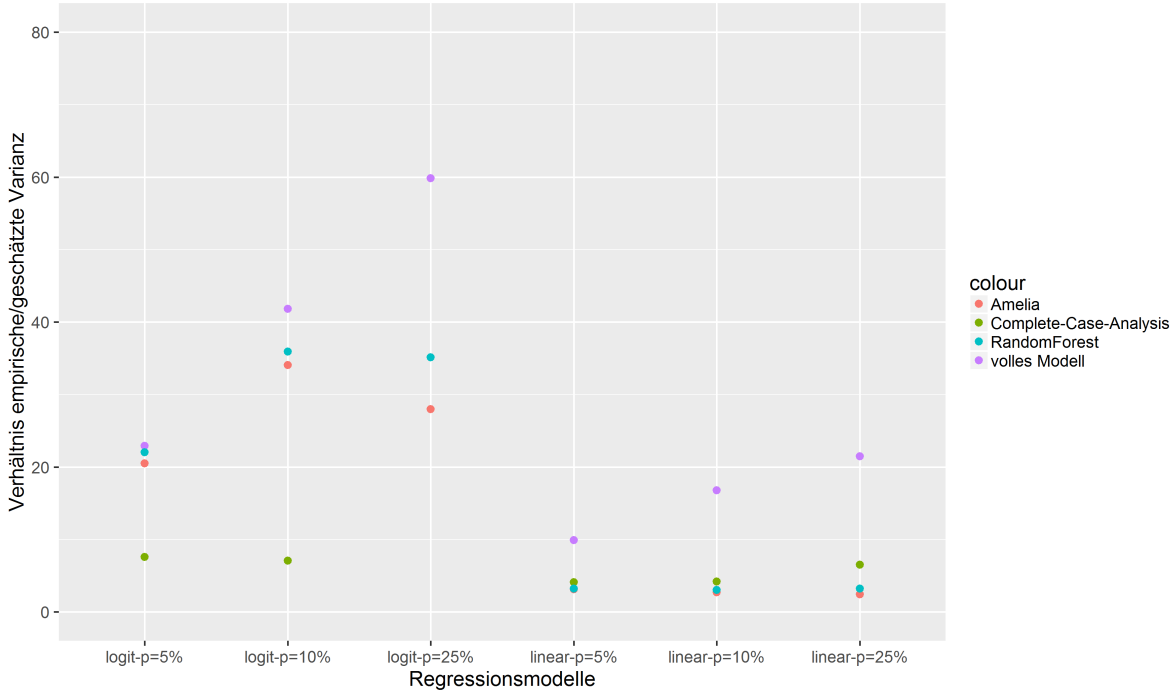
Im Folgenden sind wieder die beiden Grafiken (Abbildung 6 und Abbildung 7)zum Vergleich der Varianzen bei Variation von π abgebildet:

Abbildung 6: Varianzvergleich-Differenzen



Auch in diesen Grafiken wurde die Complete-Case-Analysis für $\pi = 0.25$ aus den Grafiken aufgrund der Nicht-Konvergenz des Modells entfernt. Wiederum übersteigen alle empirisch gemessenen Varianzen den geschätzten Varianzen des Modells. Im Logit-Modell fällt zunächst hinsichtlich der Differenz kein konstanter Einfluss des Anteils der fehlenden Werte auf die Schätzung der Varianz auf. Die geringste Differenz ergibt sich für $\pi = 0.25$, wobei die Differenzen für $\pi = 0.05$ bei allen Regressionsmodellen immer noch geringer sind wie für $\pi = 0.1$ ist. Die erhöhte Unsicherheit aufgrund der größeren Anzahl von fehlenden Werten wird also von der geschätzten Varianz besser aufgefasst, als wenn nur ein kleinerer Anteil an fehlenden Werten vorliegt. Dies bedeutet nicht, dass die Varianzen selbst bei Erhöhung der erwarteten fehlenden Werten sinkt,

Abbildung 7: Varianzvergleich-Verhältnisse



sondern vielmehr, dass das erhöhte Risiko durch einen größeren Anteil an fehlenden Werten besser geschätzt werden kann, zumindest im Logit-Modell. Beim linearen Modell hingegen steigen die Unterschiede bezüglich der Differenzen der Schätzung im Durchschnitt zunehmend an. Wie auch beim Stichprobenumfang ist die Differenz bei der Complete-Case-Analysis im Logit-Modell am geringsten und im linearen Modell am höchsten. Betrachtet man die relativen Veränderungen, steigt die Überschätzung der Varianz wiederum in beiden Modellen mit zunehmendem Anteil fehlender Werte an. Vor allem das volle Modell wird bei $\pi = 0.05$ noch um den Faktor 22 unterschätzt, während der empirische MSE bei $\pi = 0.25$ sogar den fast 60 fachen Wert beträgt. Die sehr großen Verhältnisse bezüglich der Varianz lassen sich wiederum auf die dazu recht kleinen Varianzen zurückführen, sodass eine kleine absolute Veränderung eine hohe relative Steigerung bedeutet. Im linearen Modell wird die empirische Varianz bei Amelia und Random Forest bezüglich der mittleren Differenz durchweg um etwa ein Drittel unterschätzt, während die Differenz zwischen geschätzter Varianz und empirischem MSE bei der Complete-Case-Analysis mit zunehmenden π zunehmend ansteigt.

Vor allem beim linearen Modell wirkt sich der variierte erwartete Anteil an fehlenden Werte auf die Schätzung aus. Bei einer Reduzierung von π in der Varianzkomponente sogar meistens erhöhte Fehlerkomponenten auftreten, sodass sich die Koeffizientenschätzungen trotz weniger fehlende Daten verschlechtert haben. An diesem Beispiel zeigt sich auf, wie abhängig die Ergebnisse von den Simulationen der Daten ist. Obwohl das volle Modell von diesen Schwankungen nicht beeinflusst werden sollte, da sich die Datengrundlage hier nicht ändert, sind auch im vollen Modell veränderte Fehlerkomponenten sichtbar. In der Complete-Case-Analysis sind jedoch (sofern berücksichtigt) die erwarteten Ergebnisse aufgetreten, sodass mit zunehmenden π die Fehlerkomponenten ansteigen. Allerdings ist für $\pi = 0.25$ eine deutliche Steigerung der Fehler im linearen Modell erkennbar. Während Random Forest für im Logit Modell für $\pi = 0.05$ geringere Fehlerkomponenten aufweist, schneidet Amelia bei steigendem π besser ab. Auch im linearen Modell weist Amelia, besonders bei hohen π, die geringeren Fehlerkomponenten auf. Im Vergleich der Varianzen wird wiederum das volle Modell am meisten unterschätzt, während die Complete-Case-Analysis bezüglich der Differenz und dem Verhältnis der empirischen und geschätzten Varianz die geringsten Unterschiede aufweist.

3.3.3 Ziehungen des Responsevektors

Als nächstes soll der Einfluss der Anzahl von wiederholten Ziehungen der Responsevariablen untersucht werden. Tabelle 12 und Tabelle 13 vergleichen die Varianz- und Biaskomponenten, wenn die Anzahl der Ziehungen von $S = 50$ auf $S = 10$ beziehungsweise $S = 100$ gesetzt wird:

Tabelle 12: Ziehungen von Y-Varianzkomponente

	$S = 10$ absolut	$S = 10$ relativ	$S = 100$ absolut	$S = 100$ relativ
voll-log	-0.0061	0.6243	0.0004	1.0614
CC-log	-0.0466	0.8216	-0.0367	0.9163
Amelia-log	-0.0084	0.6100	-0.0001	1.0670
RF-log	-0.0086	0.6231	-0.0014	1.0103
voll-lin	0.0263	1.3439	0.0102	1.1216
CC-lin	0.0210	1.2087	0.0831	1.4632
Amelia-lin	0.0202	1.2427	0.0234	1.2510
RF-lin	0.0212	1.2498	0.0242	1.2550

Tabelle 13: Ziehungen von Y-Biaskomponente

	$S = 10$ absolut	$S = 10$ relativ	$S = 100$ absolut	$S = 100$ relativ
voll-log	0.0355	1.0268	0.1132	1.1469
CC-log	0.0133	1.4933	0.0963	1.5008
Amelia-log	0.0257	1.1076	0.1066	1.2100
RF-log	0.0338	1.1080	0.1130	1.2234
voll-lin	0.0467	20.4592	0.0272	1.1124
CC-lin	0.0287	4.3279	0.0101	1.9487
Amelia-lin	0.0824	1078.5615	0.0335	23.6247
RF-lin	0.0576	15.0840	0.0340	1.9560

Zunächst soll wiederum das Logit-Modell betrachtet werden. Sowohl im absoluten als auch im relativen Vergleich des Modells mit $S = 10$ Ziehungen gegenüber den 50 Ziehungen sinkt die Varianzkomponente in allen Regressionsmodellen. Die größte mittlere absolute Verbesserung wird dabei im Complete-Case-Modell erzielt, die Varianzkomponente sinkt um 0.047 im Mittel und die Varianzkomponente verringert sich durchschnittlich um ungefähr 18% in jedem Modell. Der größte relative Unterschied wird bei Amelia festgestellt; für $S = 10$ wird eine um fast 39% geringere Varianzkomponente festgestellt, wobei diese durchschnittlich um 0.008 abnimmt. Allerdings nimmt die Biaskomponente in allen Modellen gleichzeitig zu. Während die Complete-Case-Analysis bezüglich der Varianzkomponente die größte absolute und kleinste relative Verbesserung vorweist, nimmt bezüglich des Bias die Complete-Case-Analysis verglichen mit den anderen Modellen absolut mit 0.013 am wenigsten zu, obwohl der relative Unterschied mit durchschnittlich fast 50% am höchsten ist. Im linearen Modell hingegen schneidet das Modell im Vergleich sowohl der absoluten als auch der relativen Veränderungen hinsichtlich der Varianzkomponenten schlechter ab, wobei die Varianzkomponente im Mittel bei allen Modellen um etwa 0.02 bis 0.026 höher abschneidet, was einer durchschnittlichen Erhöhung um etwa 20-36% entspricht. Auch anhand der Biaskomponenten zeigt sich im Vergleich der absoluten Differenzen und der relativen Veränderungen im Durchschnitt, dass die reduzierte Anzahl an Ziehungen zu höheren Verzerrungen führt. Während die durchschnittlichen Differenzen zum Grundmodell sich alle im Intervall zwischen 0.03 und 0.08 befinden, variieren die mittleren relativen Unterschiede zwischen deutlich stärker.Bei der Complete-Case-Analysis vervierfacht sich die Biaskomponente durch Reduzierung der Anzahl von Ziehungen des Responsevektors, wobei allein bei x_4 eine um den Faktor 19 erhöhte Biaskomponente auftritt. Das Random Forest Modell kann die Erhöhung um Fak-

tor 15 vor allem auf die Schätzer von β_8 (Faktor 72) und β_3 zurückgeführt werden, während die um den Faktor 1078 erhöhte Biaskomponente bei Amelia anhand der um den Faktor 10727 gestiegenen Biaskomponente des Schätzers für β_2 erklärt werden kann, wobei die absolute durchschnittliche Differenz bei β_2 0.07 beträgt. Bei Erhöhung des Stichprobenumfangs auf $S = 100$ ist weiterhin ein ähnlicher Trend zu sehen. Bei der Complete-Case-Analysis sinkt die Varianzkomponente im Vergleich zum Grundmodell um 0.037 und die Variablen sinken im Mittel um 9% bezüglich der Varianz. Die absoluten Differenzen betragen in den anderen Modellen im Durchschnitt weniger als 0.001, während die relativen Differenzen im vollen Modell und bei Amelia um 6 beziehungsweise 7% steigen. Obwohl die Anzahl an Ziehungen höher ist, nimmt die Präzision der Schätzungen im Durchschnitt in allen Modellen ab, sodass sowohl alle relativen Verhältnisse als auch absoluten Differenzen bezüglich der Biaskomponente im Vergleich höher ausfallen. Die Unterschiede gegenüber dem ursprünglichen Modell betragen im Mittel bezüglich der absoluten Differenzen in allen Modellen je ungefähr 0.1, während die Variablen durchschnittlich bei der Complete-Case-Analysis um 50%, im vollen Modell um 15% und in Amelia und Random Forest jeweils um etwa 20% steigen. Im linearen Modell kann sowohl bei der Varianzkomponente als auch bei der Biaskomponente im Vergleich zum ursprünglichen Modell bei der Erhöhung der Anzahl von Ziehungen des Responsevektor ein Verschlechterung des MSEs festgestellt werden. Das Complete-Case-Analysis-Modell nimmt dabei am meißten (0.083 absolut, +46%) durchschnittlich an Varianz zu. Was den Bias betrifft, kann bei der Complete-Case-Analysis eine mittlere absolute Differenz von 0.01 beobachtet werden, während die anderen Regressionsmodelle um durchschnittlich etwa 0.03 zunehmen. Die relativen Unterschiede ergeben im Durchschnitt 11% beim vollen Modell, während bei der Complete-Case-Analysis und dem Random Forest Modell eine im Durchschnitt um je etwa 95% größere Verzerrung auftritt. Die um den Faktor 23.62 höhere durchschnittliche Biaskomponente kann wiederum auf Variable x_2 zurückgeführt werden, in welcher der Bias um 0.0015 zunimmt, was allerdings einem Anstieg um Faktor 223.6 bedeutet.

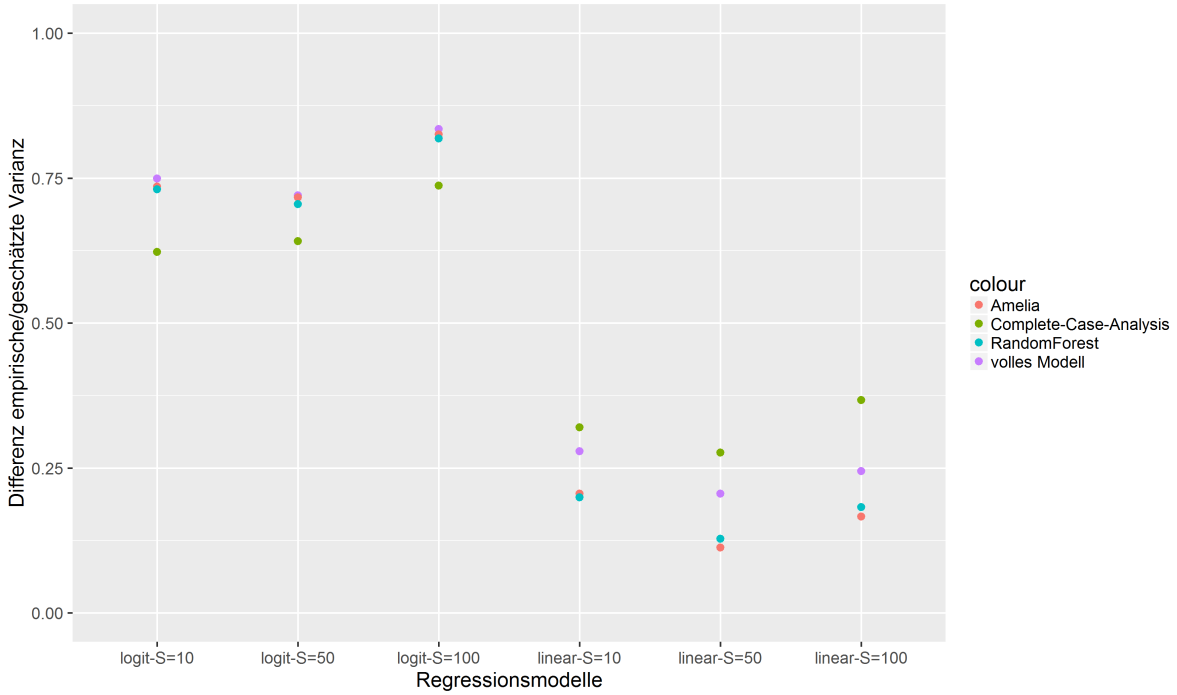
Tabelle 14: Differenz und Verhältnis von Random Forest und Amelia

	S_{10}	S_{50}	S_{100}
Varianzkomponente			
$logit_{absolut}$	0.0013	0.0015	0.0002
$logit_{relativ}$	1.1003	1.0722	1.0118
$lin_{absolut}$	0.0117	0.0106	0.0114
$lin_{relativ}$	1.1229	1.0936	1.0946
Biaskomponente			
$logit_{absolut}$	-0.0067	-0.0148	-0.0083
$logit_{relativ}$	0.9793	1.0033	1.0109
$lin_{absolut}$	-0.0277	-0.0029	-0.0024
$lin_{relativ}$	1.3130	106.9827	4.2729

Die Tabelle 14 vergleicht wiederum die Ergebnisse von Random Forest und Amelia. Bei multiplen Ziehungen des Responsevektors hat nur einen geringen Einfluss auf den Vergleich des Abschneidens von Amelia und Random Forest. Sowohl im linearen als auch im Logit-Modell schneidet Amelia bezüglich der Varianzkomponente im Logit-Modell um ca. 0.0014 und im linearen Modell um ca. 0.0115 besser ab. Für $S = 100$ sinkt im Logit-Modell die Differenz auf 0.002, die Varianzkomponente weist also bei Random Forest ähnliche Fehler wie bei Amelia auf. Während bei Random Forest im linearen Modell die Varianzkomponente immer etwa 10% höher ist, fällt wie auch bei den Differenzen auf, dass für $S = 100$ Random Forest bezüglich der Streuung der Koeffizientenschätzer nur noch um unter 2% schlechter abschneidet. Hinsichtlich des Bias fällt wiederum auf, dass sich für Random Forest in allen Modellen im Durchschnitt

präzisere Schätzungen ergeben. Während im linearen Modell die Differenz der Biaskomponenten im Durchschnitt mit zunehmender Anzahl an Ziehungen konstant abnimmt, ist kein Trend beim Logit Modell zu erkennen, wobei beide MSE-Komponenten je im Modell mit $S = 50$ über die größten Fehler verfügen. Im linearen Modell ist die MSE-Komponente, die sich als Summe der Varianz- und Biaskomponente ergibt, also bei Amelia höher, während für $S = 50$ und $S = 100$ Amelia einen geringeren MSE aufweist. Auch bei Betrachtung der mittleren Verhältnisse fällt auf, das Amelia nur im Logit-Modell schlechter abschneidet als Random Forest. Im Folgenden sollen wiederum die Differenzen und Verhältnise der empirischen und geschätzten Varianzen anhand der Scatterplots Abbildung 8 und Abbildung 9 betrachtet werden.

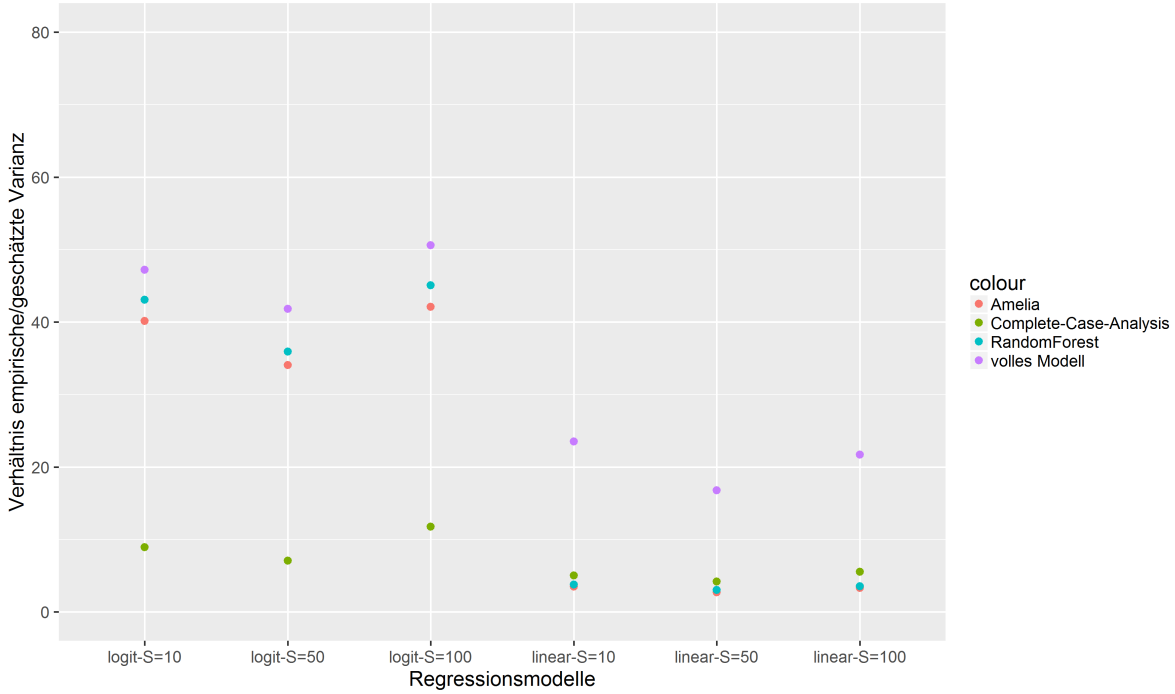
Abbildung 8: Varianzvergleich-Verhältnisse



Zwischen den unterschiedlichen Regressionsmodellen sind durchgängige Unterschiede bezüglich der Differenzen feststellbar, was allerdings die Ziehungen des Responsevektors angeht, kann keine Konstanz festgestellt werden. Während wiederum die Complete-Case-Analysis im Logit-Modell die kleinsten Differenzen aufweist, sind diese im linearen Modell wieder verglichen mit den anderen Regressionsmodellen am größten. Die Differenzen der beiden Imputationsmethoden sind recht ähnlich, die Differenzen beim vollen Modell sind beim Logit-Modell am höchsten und im linearen Modell am zweithöchsten hinter der Complete-Case-Analysis. Die Differenzen betragen dabei beim Logit-Modell ungefähr 0.6 und 0.8, während die Unterschiede sich im linearen Modell auf ungefähr zwischen 0.1 und 0.3 belaufen. Für $S = 50$ erfolgt bezüglich des Verhältnisses der Varianzen der kleinste Unterschied. Sowohl im linearen als auch im Logit-Modell wird wiederum im vollen Modell die empirische Varianz am meisten unterschätzt, wobei der Unterschied der Verhältnisse der Varianzen im Logit-Modell nur etwa ein Fünftel und im linearen Modell ungefähr ein Viertel der Unterschiede bei der Compelte-Case-Analysis betragen. Für Random Forest und Amelia ergibt sich derweil ein ähnlich hohes Verhältnis wie beim vollen Modell beim Logit-Modell, während im linearen Modell sich ähnliche Verhältnisse wie bei der Complete-Case-Analysis ergeben.

Zusammenfassend zeigt vor allem das lineare Modell einen Einfluss von Variationen des Stichprobenumfangs auf. In der Biaskomponente wurden bei Reduzierung auf 10 Ziehungen des Responsevektors durchgehend schlechtere Ergebnisse erzielt, allerdings schnitten sowohl Logit- als auch lineares Modell ebenfalls bei einer Erhöhung von S schlechter ab als beim Grundmodell.

Abbildung 9: Varianzvergleich-Differenzen



Auch in der Varianzkomponente kann nur in der Complete-Case-Analysis ein besseres Ergebnis für $S = 100$ erzielt werden, während wiederum im Logit-Modell die Reduzierung der Ziehungen auch zu einer Reduzierung der Varianzkomponente führt. Im Vergleich von Amelia und Random Forest ist kaum ein auffälliger Einfluss von S sichtbar, während Random Forest bezüglich der Biaskomponenten immer kleinere Verzerrungen aufweist, wurden kleinere Varianzkomponenten bei Amelia gemessen. Was die Varianzvergleiche betrifft, können im linearen Modell nur geringe Unterschiede zwischen den geschätzten und den empirischen Varianzen bei Random Forest und Amelia festgestellt werden.

3.3.4 Variation von ϵ

Anschließend soll der Einfluss des Fehlerterms ϵ auf das Modell untersucht werden. Je größer dieser Fehler ist, desto mehr wird die Schätzung der Responsevektoren beeinflusst, was wiederum den Einfluss der Kovariablen auf die Responsevektoren (und die Regressionsmodelle als Schätzung jener) schwächt. Da der Fehlerterm nur im linearen Modell auftritt, werden auch nur die linearen Modelle untersucht.Während im ursprünglichen Modell $\sigma = 9$ verwendet wurde, wird σ in folgenden Modellen durch $\sigma_1 = 2.25$ und $\sigma_2 = 36$ ersetzt. Die folgenden Tabelle 15 zeigen also die Varianzkomponente und die Biaskomponente dieser Variablen.

Sowohl bezüglich der Bias- als auch der Varianzkomponenten verlaufen die absoluten Differenzen erwartungsgemäß; je größer die Varianz des Störterms σ, desto ungenauer die Schätzung bezüglich des MSEs, beziehungsweise bezüglich der Varianz- oder der Biaskomponente. Allerdings kann man bei der Varianzkomponente sehen, dass die Erhöhung von $\sigma = 2.25$ auf $\sigma = 9$ mit einer durchschnittlichen Steigung der Varianzen über alle Modelle und über alle Variablen von ungefähr 0.08 eine geringere Verbesserung mit sich bringt als die erneute Erhöhung σ von $9 \rightarrow 36$ von durchschnittlich ungefähr 1.812. Während durch die Reduzierung von σ somit eine relative Verminderung der Varianz um ca. 50% bei Amelia und ca. 60% bei den anderen Modellen feststellbar ist, erhöht sich die Varianz jeweils um ungefähr den Faktor 15 (beziehungsweise 16 bei Random Forest) für $\sigma = 36$. Auch in der Biaskomponente sinkt die Varianz in allen Regressionsmodellen, wenn σ auf 2.25 gesetzt wird. Allerdings steigt das durchschnittliche Verhältnis beim

Tabelle 15: Einfluss ϵ- Oben: Varianzkomponente, unten: Biaskomponente

	$\sigma = 2.25$ absolut	$\sigma = 2.25$ relativ	$\sigma = 36$ absolut	$\sigma = 36$ relativ
voll-var	-0.0602	0.3918	1.3901	15.9168
CC-var	-0.1536	0.3715	2.8611	14.8371
Amelia-var	-0.0518	0.5182	1.4347	15.9501
RF-var	-0.0685	0.4193	1.5857	16.3878
voll-bias	-0.0115	2.4555	0.0245	42.4325
CC-bias	-0.0305	2.1671	0.1057	13.2163
Amelia-bias	-0.0054	270.2753	0.0521	26.2486
RF-bias	-0.0097	21.0308	0.0488	41.0923

vollen Modell und bei der Complete-Case-Analysis um den Faktor 2, bei Amelia und Random Forest sogar um den Faktor 270 beziehungsweise den Faktor 21. Der sehr hohe Wert bezüglich der Verzerrung von Amelia ist wiederum auf x_2 zurückzurühren, welche im Vergleich zum Modell mit $\sigma = 9$ eine um 0.001 höhere beziehungsweise eine um 2665 fach größere Biaskomponente verfügt. In x_8 ist im Verhältnis die Biaskomponente bei Random Forest um 17286% gestiegen. Bei erneuter Erhöhung von σ auf 36 sind die Resulatate konstanter. In allen Regressionsmodellen ist eine Steigung der Verzerrung um zwischen 0.025 beim vollen Modell und 0.1 bei der Complete Case-Analysis zu sehen, was zu durchschnittlichen relativen Erhöhung der Biaskomponente um den Faktor 40 beim vollen Modell und Random Forest und einer 13 beziehungsweise 26 fachen Erhöhung bei Complete-Case-Analysis und Amelia sorgt.

Wiederum werden im Anschluss in Tabelle 16 die Differenzen und Verhältnisse der Amelia und Random Forest Modelle in folgenden Tabellen miteinander verglichen: Auch hier wird nur

Tabelle 16: Vergleich und Verhältnis von Random Forest und Amelia- σ

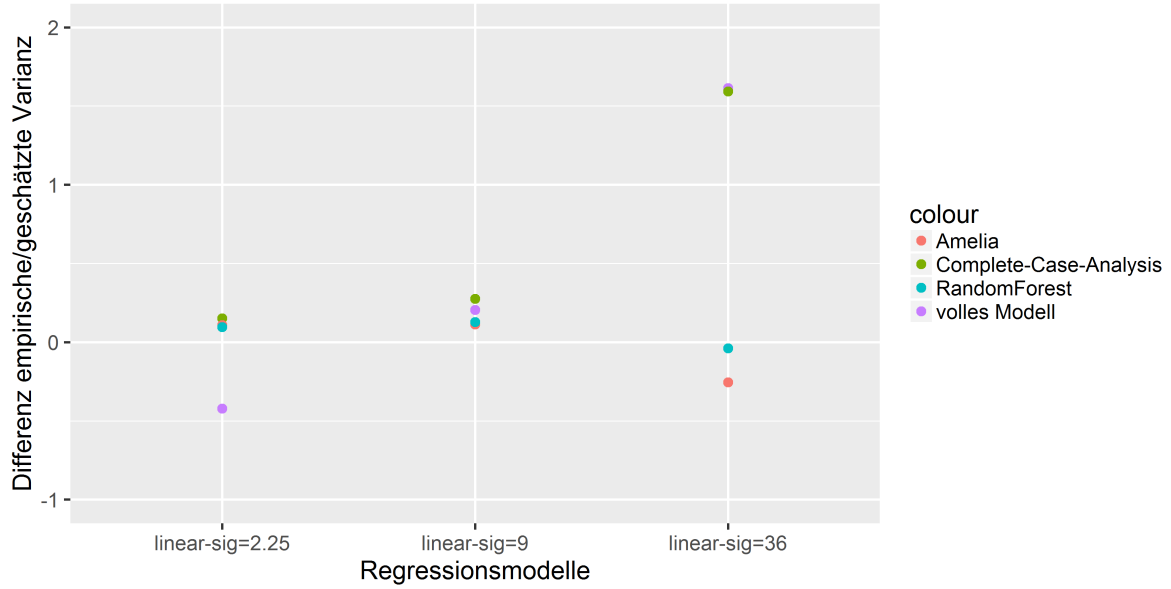
	$\sigma = 2.25$	$\sigma = 9$	$\sigma = 36$
Var_absolut	-0.0061	0.0106	0.1615
Var_relativ	0.8976	1.0936	1.1133
Bias_absolut	-0.0072	-0.0029	-0.0062
Bias_relativ	0.9557	106.9827	2.4686

Rücksicht auf die linearen Fehlerterme genommen, da nur in diesen ϵ vorhanden ist. Zunächst lässt sich erkennen, dass Amelia bezüglich der Varianzkomponente mit zunehmenden σ im Vergleich zu Random Forest einee zunehmend geringere Streuung aufweist. Während für $\sigma = 2.25$ Random Forest noch eine um 0.061 im Mittel und durchschnittlich 10% geringere Varianzkomponente aufweist, schneidet diese für $\sigma = 9$ beziehungsweise $\sigma = 36$ um 0.01 und 0.162 schlechter ab, was auch eine mittlere Erhöhung der Varianzkomponente von 9 und 11% bedeutet. Bezüglich des Bias schneidet Random Forest hinsichtlich der durchschnittlichen absoluten Unterschiede jeweils besser ab, wobei Amelia für $\sigma = 0.03$ am besten abschneidet. Die um 106 mal höhere Random Forest-Biaskomponente wurde bereits in der Sektion zu den Stichprobenumfängen erklärt.

Nachfolgend sind in Abbildung 10 und Abbildung 11 die Varianzvergleiche graphisch dargestellt.

Auch bezüglich der Varianzen werden nur die linearen Modelle berücksichtigt, da im Logit-Modell kein Fehlerterm ϵ erscheint, und somit eine Veränderung von σ keinen Einfluss zeigen kann. Für $\sigma = 2.25$ im vollen Modell und für $\sigma = 36$ bei Amelia und Random Forest liegt die geschätzte Varianz über dem empirischen MSE. Im vollen Modell ist mit zunehmender Varianz des Fehlerterms ϵ also eine deutliche Steigung von -0.45 auf 0.2 und schließlich 1.6 feststellbar, sodass hier eine konstanter Einfluss zu sehen ist. Ein ähnlicher Effekt lässt sich auch bei der Complete-Case-Analysis erkennen, während bei Amelia und Random Forest die Differenz der empirischen und der geschätzten Varianz bei $\sigma = 2.25$ und $\sigma = 9$ um 0.1 schwankt, während für

Abbildung 10: Varianzvergleich-Verhältnisse



$\sigma = 36$ die Differenzen jeweils unter 0 fallen. Bei zunehmendem Fehlerterm im linearen Modell trifft die geschätzte Varianz also eher zu. Im Umkehrschluss könnte die geschätzte Varianz eben diesen Fehlerterm unterschätzten, sodass die Varianz der Kovariablen zu klein eingeschätzt wird. Betrachtet man die Verhältnisse der Varianzen, ist auch dort bezüglich dem vollen Modell zu sehen, dass für ansteigendes σ die empirische Varianz deutlich höher ausfällt, als die geschätzte Varianz, sodass für $\sigma = 36$ eine im Mittel um den Faktor 71 fache empirische Varianz gemessen wird. Sowohl Amelia als auch Random Forest nehmen bezüglich dem relativen Unterschied der Varianzen mit zunehmendem σ ab. Während für $\sigma = 2.25$ die empirische Varianz den fünfachen Wert der geschätzten Varianz annimmt, liegt bei $\sigma = 9$ nur noch ungefähr der dreifache Wert vor. Für $\sigma = 36$ beträgt die empirische Varianz bei Random Forest im Mittel nur etwa 4% über der geschätzten Varianz, während bei Amelia die empirische Varianz um 7% geringer ist.

Der Einfluss der Standardabweichung σ der Fehlerkomponente ϵ ist sowohl bezüglich der Varianz-, als auch der Biaskomponente recht deutlich zu sehen. Eine Erhöhung von σ führt somit zu einem eindeutigen Anstieg der Fehler. Während Amelia wiederum bei steigendem σ im Verhältnis zu Random Forest weniger an Varianz zunimmt, schneidet Random Forest b(bei ebenfalls steigendem σ) bezüglich der Biaskomponente besser ab. Allerdings treten keine konstante Veränderung der Differenzen oder Verhältnisse auf. Für $\sigma = 36$ kann zudem die kleinste Differenz zwischen empirischer und geschätzter Varianz festgestellt werden.

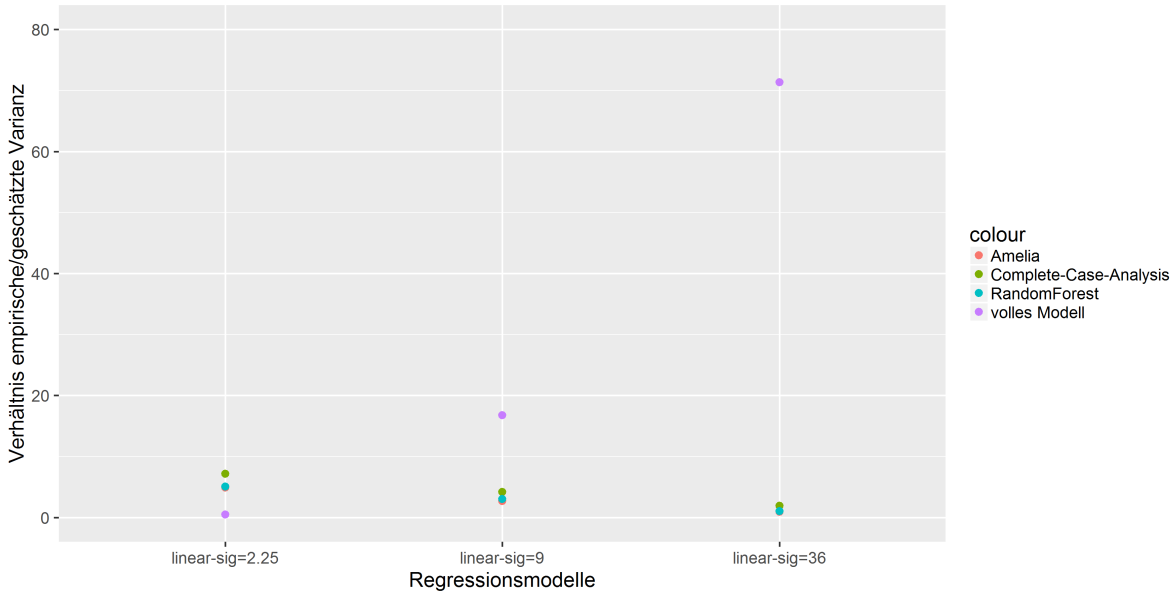
Die Untersuchung der Variationen einzelner Parameter zeigt auf, wie sehr die Ergebnisse doch von der Simulation des X-Datensatzes abhängen. Daher soll im anschließenden Abschnitt nun auch X mehrmals gezogen werden.

4 Ergebnisse bei multiplen Ziehungen von X

4.1 Modell mit Parametern aus 3.2

Zunächst werden im folgenden Modell die Parameter des Grundmodells wieder verwendet. Allerdings wird in diesem Fall X $W = 20$ mal simuliert. Die Ergebnisse sind hier also nicht nur über die S Ziehungen von Y oder Z, sondern zusätzlich über die W Ziehungen von X gemittelt. Da sich somit die wahre Kovarianzmatrix bei jeder Ziehung ändert, wird hier kein Vergleich der empirischen und geschätzten Varianzen durchgeführt. Allerdings wurden die MSE-Komponenten bei multiplen Ziehungen von X nicht nur auf die Koeffizientenschätzer der Regressionen, sondern

Abbildung 11: Varianzvergleich-Verhältnisse



auch auf die anhand Amelia und Random Forest imputierten Daten angewandt. Die Schätzungen der fehlenden Werte bei Amelia werden wiederum nach [10] als die Mittelwerte aller m vervollständigten Datensätze bestimmt.

Die im Kapitel 3.2 festgelegten Parameter sollen anschließend in einem Modell angewandt werden, bei dem X mehrfach simuliert wird. Tabelle 17 und Tabelle 18 zeigen daher die Varianz- und Biaskomponenten der Koeffizientenschätzer für dieses Modell:

Tabelle 17: Varianzkomponente, oben: Logit, unten: Linear

	Var_voll	Var_CC	Var_Amelia	Var_RF
β_1	0.0001	0.0004	0.0002	0.0002
β_2	0.0531	0.1972	0.0642	0.0728
β_3	0.0430	0.0537	0.0390	0.0462
β_4	0.0480	0.1915	0.0471	0.0581
β_5	0.0068	0.0179	0.0073	0.0084
β_6	0.0087	0.0273	0.0079	0.0099
β_{7_1}	0.0310	0.1026	0.0250	0.0272
β_{7_2}	0.0070	0.0146	0.0064	0.0065
β_{7_3}	0.0258	0.0629	0.0215	0.0234
β_8	0.0335	0.0774	0.0277	0.0284
β_1	0.0004	0.0007	0.0005	0.0005
β_2	0.1259	0.4138	0.1852	0.1864
β_3	0.1560	0.2021	0.1617	0.1784
β_4	0.1631	0.4958	0.1671	0.1907
β_5	0.0404	0.0613	0.0469	0.0535
β_6	0.0310	0.0582	0.0397	0.0372
β_{7_1}	0.1114	0.2143	0.1049	0.1191
β_{7_2}	0.0353	0.0815	0.0417	0.0443
β_{7_3}	0.1234	0.2326	0.1463	0.1507
β_8	0.1065	0.1238	0.1075	0.1141

Zunächst soll dieses Modells dem aus nur einem Datensatz resultierenden Modell aus 3.2 ge-

Tabelle 18: Biaskomponente, oben: Logit, unten: Linear

	Bias_voll	Bias_CC	Bias_Amelia	Bias_RF
β_1	0.0000	0.0001	0.0000	0.0000
β_2	0.0214	0.0635	0.0007	0.0569
β_3	0.0020	0.0053	0.0092	0.0082
β_4	0.0129	0.0805	0.0003	0.0445
β_5	0.0014	0.0094	0.0005	0.0090
β_6	0.0012	0.0034	0.0001	0.0022
β_{7_1}	0.2949	0.3423	0.3771	0.3788
β_{7_2}	0.5244	0.5093	0.5615	0.5481
β_{7_3}	0.4295	0.4135	0.3885	0.3914
β_8	0.0041	0.0253	0.0027	0.0027
β_1	0.0000	0.0000	0.0001	0.0000
β_2	0.0001	0.0099	0.0004	0.0126
β_3	0.0002	0.0005	0.0032	0.0001
β_4	0.0111	0.0074	0.0320	0.0011
β_5	0.0000	0.0000	0.0011	0.0053
β_6	0.0004	0.0004	0.0015	0.0000
β_{7_1}	0.4924	0.5362	0.4972	0.5239
β_{7_2}	0.5197	0.5275	0.5669	0.5605
β_{7_3}	0.3370	0.3189	0.3368	0.3273
β_8	0.0009	0.0023	0.0001	0.0003

genübergestellt werden. In der Varianzkomponente zeigt sich in den Daten nur bei der Complete-Case-Analysis eine Reduzierung der Streuung durch multiple Simulation von X auf. Im Durchschnitt liegt die Varianzkomponente pro Variabel bei der Complete-Case-Analysis demnach um 0.048 unter der Streuung beim Modell mit nur einer Simulation von X, wobei sich für $W = 20$ beim vollen Modell, bei Amelia und bei Random Forest eine um 0.010, 0.006 und 0.008 höhere Varianzkomponente ergibt. Somit erzielt das Modell mit nur einer Variable für die Complete-Case-Analysis eine im Mittel pro Variabel um 22% geringere Varianzkomponente, während die Streuungen beim vollen Modell, Amelia und Random Forest um 57%, 31% und 40% höher sind. Die Biaskomponente fällt allerdings im Modell mit $W = 1$ über alle Variablen gemittelt zwischen 0.48 bei der Complete-Case-Analysis und 0.591 beim vollen Modell höher aus, sodass das Modell mit $W = 20$ Ziehungen von X im Durchschnitt pro Variable beim vollen Modell nur etwa 35% und bei den übrigen Modellen etwa die Hälfte der Streuung des Modells mit nur einer Simulation von X annimmt. Somit wird aufgrund der 20-fachen Ausführung der Simulation und der Berechnungen im Logit Modell die Varianzkomponente des MSE zwar nicht verbessert, allerdings sinkt der Präzisionsfehler etwa um die Hälfte, im vollen Modell wurde dieser sogar um 65% reduziert. Im linearen Modell tritt nur bei der Complete-Case-Analysis eine im Mittel um 35% geringere Biaskomponente bei wiederholten Ziehungen von W auf, wobei der Präzisionsfehler pro Variable im Mittel um 0.008 sinkt. Während sich im vollen Modell , bei Amelia und bei Random Forest die absoluten Differenzen pro Variable bei Erhöhung der Simulationen von X auf $W = 20$ um 0.007, 0.016 und 0.018 steigt, nimmt die durchschnittliche relative Veränderung der Biaskomponente pro Variable im vollen Modell um Faktor 1.60, bei Amelia um Faktor 6.78 und bei Random Forest um Faktor 1.21 zu. Im linearen Modell treten bezüglich der Biaskomponente also nur Verbesserungen in der Complete-Case-Analysis auf. Allerdings ergibt sich für die Varianzkomponente im linearen Modell bezüglich der absoluten Unterschiede in allen Modellen für $W = 20$ eine Reduzierung der Streuung um zwischen 0.040 und 0.065. Dennoch steigt die Varianzkomponente im Schnitt pro Variabel im vollen Modell um 2.4% und bei Amelia und Random Forest um 12 beziehungsweise 10.5% an, während bei der Complete-Case-Analysis die Varianzkomponente im

Modell aus nur einer Simulation im Mittel eine um 9% geringere Streuung auftritt.

Im Folgenden sollen die Imputationen bezüglich der MSE-Komponenten verglichen werden. Der MSE wird wiederum in die zwei Komponenten aufgeteilt und wird variablenspezifisch bestimmt. Sei F die Anzahl fehlende Werte in Variable j. Sei weiterhin γ einer der F fehlenden Werte in der Variable j, und $\hat{\gamma}_{g,imp}$ die g-te Imputation von γ der Schätzmethode imp an der Stelle f, wobei $g \in \{1, \ldots, G\}$.. Dann ist die Varianzkomponente demnach folgendermaßen definiert:

$$Var_{MSE} = \frac{1}{F} \sum_{f=1}^{F} \left(\frac{1}{G} \sum_{g=1}^{G} (\hat{\gamma}_{imp,g,f} - \bar{\gamma}_{imp,f})^2 \right)$$

, wobei $\bar{\gamma}_{imp,f} = \frac{1}{G} \sum_{g=1}^{G} \hat{\gamma}_{imp,g,f}$ Die Varianzkomponente wird demnach für eine Imputationsmethode imp und einen fehlenden Wert f bestimmt, und anschließend über alle fehlenden Werte F einer Variable j gemittelt. Zudem wird ein Gesamt-Imputationssfehler bestimmt, indem die Varianzkomponente nicht nur für eine Variable, sondern alle fehlenden Werte bestimmt wird. Analog dazu wird auch die Biaskomponente für die Imputationen bestimmt:

$$Bias_{MSE} = \frac{1}{F} \sum_{f=1}^{F} \left(\frac{1}{G} \sum_{g=1}^{G} (\bar{\gamma}_{imp,f} - \gamma_f)^2 \right)$$

Tabelle 19: MSE-Komponenten für Modell mit festen Parametern, W=20

	Amelia_logit	Amelia_linear	RF_logit	RF_lineaer
Varianzkomponente				
Gesamt	8.0900	7.4671	2.1126	1.8144
x_1	78.2336	72.3596	20.6018	17.7283
x_2	0.5403	0.4253	0.1498	0.1059
x_3	0.1709	0.1692	0.0195	0.0230
x_4	0.2715	0.2364	0.0943	0.0585
x_5	1.2651	1.1155	0.2959	0.2255
x_6	1.0917	1.0259	0.1637	0.1762
x_{7_1}	0.1807	0.1815	0.0241	0.0300
x_{7_2}	0.1732	0.1713	0.0241	0.0292
x_{7_3}	0.1880	0.1892	0.0245	0.0297
x_8	0.3119	0.2967	0.0363	0.0435
Biaskomponente				
Gesamt	30.4925	34.2863	32.1819	34.0226
x_1	291.8246	328.5783	308.4359	326.7837
x_2	0.8465	1.0527	1.1945	1.3719
x_3	0.6580	0.6659	0.6288	0.6294
x_4	0.5964	0.8414	0.6236	0.8273
x_5	3.6781	4.3995	4.0028	4.3843
x_6	4.0973	4.4123	4.3517	4.4323
x_{7_1}	0.9051	0.9077	0.9605	0.9578
x_{7_2}	0.9656	0.9683	1.0169	1.0053
x_{7_3}	0.9625	0.9684	0.9811	0.9751
x_8	1.1332	1.1680	1.0250	1.0311

Im Allgemeinen fällt zunächst auf, in diesem Modell bezüglich der Varianzkomponente der Imputationen Random Forest besser abschneidet als Amelia. Sowohl bei den Gesamt-Imputationen, als auch bei allen einzigen Variablen und sowohl beim Logit- als auch beim linearen Modell ist die Varianzkomponente bei den Imputationen von Amelia um zwischen 187 und 774% höher als

die entsprechende Random Forest-Komponente. Auch bei den Normalverteilten Daten schneidet Amelia hier also nicht besser ab. Weiterhin sind die imputierten Daten zumindest bezüglich der Varianzkomponente bei den Logit-Modellen in den meisten Fällen (nicht für x_{7_3}) höher als im linearen Modell. Bezüglich der Biaskomponente ist Amelia jedoch im Logit-Modell etwas präziser, wobei im linearen Modell wiederum Random Forest eine etwas geringere Biaskomponente aufweist. Berechnet man allerdings das Verhältnis der Biaskomponenten, schneidet über alle Variablen gemittelt Amelia beim Logit-Modell um 5.3 und beim linearen Modell um 1.2% besser ab als Random Forest. Berücksichtigt man nur die Variablen, die einer Normalverteilung folgen, ist die Biaskomponente bei Random Forest im Vergleich zu Amelia beim Logit-Modell um 13% und beim linearen Modell um 10% im Schnitt höher. Insgesamt schneidet Amelia also nur bezüglich des Bias deutlich besser bei Normalverteilten Daten ab, bei der Varianzkomponente ist dieser Effekt in diesem Fall nicht aufgetreten.

4.2 Modell mit zufällig gezogenen Parametern

In folgendem Abschnitt werden, wie bereits erwähnt, nicht mehr ausschließlich die für das Grundmodell ausgewählten Parameter verwendet. Stattdessen werden die Parameter aus den in den Tabellen 3 und 2 gelisteten Ausprägungen in jeder Simulation w neu gezogen. Für die Varianzkomponenten und die Biaskomponenten hinsichtlich der Koeffizientenschätzer der Regressionsmodelle ergeben sich demnach folgende Ergebnisse, die in Tabelle 20 und Tabelle 21 dargestellt werden.

Tabelle 20: Varianzkomponente,oben: Logit, unten: Linear

	Var_voll	Var_CC	Var_Amelia	Var_RF
β_1	0.0010	0.0064	0.0008	0.0012
β_2	0.1124	4.3665	0.0880	0.1245
β_3	0.2931	0.5711	0.1985	0.1970
β_4	1.0486	28.9983	0.5848	1.0544
β_5	0.1223	1.6835	0.0692	0.1159
β_6	0.0313	0.2241	0.0270	0.0381
β_{7_1}	0.0725	0.2961	0.0607	0.0870
β_{7_2}	0.0227	0.0815	0.0193	0.0240
β_{7_3}	0.0914	0.3147	0.0631	0.0776
β_8	0.1402	1.5666	0.1095	0.1459
β_1	0.0002	0.0004	0.0002	0.0002
β_2	0.0249	0.0405	0.0278	0.0267
β_3	0.0271	0.0610	0.0433	0.0334
β_4	0.0258	0.0545	0.0260	0.0270
β_5	0.0072	0.0095	0.0075	0.0069
β_6	0.0068	0.0095	0.0084	0.0066
β_{7_1}	0.0402	0.0584	0.0461	0.0453
β_{7_2}	0.0100	0.0127	0.0105	0.0102
β_{7_3}	0.0325	0.0404	0.0354	0.0348
β_8	0.0248	0.0249	0.0269	0.0274

Es lässt sich zunächst erkennen, dass Amelia und teilweise auch Random Forest über geringere Varianzkomponenten im Logit-Modell verfügen, als das volle Modelle mit den eigentlichen Daten. Für Amelia ist die Varianzkomponente in jeder Variable bei Amelia geringer, wobei das Verhältnis der Varianzkomponenten von Amelia zum vollen Modell zwischen 0.56 und 0.86. Im Durchschnitt ist der Schätzfehler bezüglich Varianzkomponente somit im Logit Modell 26 % geringer als im vollen Modell. Der große Unterschied könnte an der allgemeinen Verzerrung der Schätzer im Logit-Modell liegen, was jedoch bei derart vielen wiederholten Ziehungen und Messungen der Varianz nicht zu stark ins Gewicht fallen sollte. Die Streuung der Schätzer im Random Forest Modell liegen zwar teilweise (β_3) auch unter den anologen Varianzkomponeten im vollen Modell, durchschnittlich beträgt die Streuung in Random Forest in jeder Variable jedoch ca. 2.2% mehr als im vollen Modell, während die Varianzkomponente bei der Complete-Case-Analysis im Schnitt fast 11.8 mal so hoch ist. Im linearen Modell hingegen schneidet das volle Modell wie erwartet in gegenüber Amelia im Mittel um 14.8%, gegenüber der Complete Case Analysis um 45.8% und gegenüber Random Forest um 7.4% im Mittel in jeder Variabel besser ab. Auch im Vergleich mit dem Logit-Modell sind die Fehler deutlich geringer, die Varianzkomponenten betragen im Logit-Modell im vollen Modell nur noch ungefähr 10% , bei der Complete-Case-Analysi nur 1%, bei Amelia noch etwa 18% und bei Random Forest noch etwa 11% gegenüber den Varianzkomponenten. Was allerdings den Bias betrifft, liegt auch im Logit-Modell eine deutlich höhere Verzerrung bei Complete-Case-Analysis, Amelia und Random Forest vor, so dass im Mit-

Tabelle 21: Biaskomponente,oben: Logit, unten: Linear

	Bias_voll	Bias_CC	Bias_Amelia	Bias_RF
β_1	0.0001	0.0007	0.0001	0.0003
β_2	0.0605	0.4324	0.0196	0.0689
β_3	0.0002	0.0293	0.0238	0.0158
β_4	0.6971	4.2576	0.0768	0.8926
β_5	0.0374	0.2523	0.0013	0.0310
β_6	0.0063	0.0290	0.0057	0.0222
β_{7_1}	0.6777	0.6361	0.7088	0.6296
β_{7_2}	0.3082	0.2340	0.3173	0.2948
β_{7_1}	0.4090	0.4061	0.4643	0.4083
β_8	0.0406	0.2840	0.0100	0.0472
β_1	<0.0001	<0.0001	<0.0001	<0.0001
β_2	0.0011	0.0005	0.0002	0.0002
β_3	0.0008	0.0008	0.0150	0.0163
β_4	0.0003	0.0001	<0.0001	0.0171
β_5	<0.0001	<0.0001	0.0005	0.0016
β_6	<0.0001	<0.0001	0.0016	0.0021
β_{7_1}	0.7080	0.7440	0.7587	0.7299
β_{7_2}	0.3596	0.3523	0.3707	0.3892
β_{7_3}	0.4249	0.4355	0.3743	0.3097
β_8	<0.0001	0.0008	<0.0001	0.0001

tel die Biaskomponente um die Faktoren 16.17, 10.35 und 7.84 höher sind. Im linearen Modell liegen kaum noch Verzerrungen vor, sodass Biaskomponenten unter 0.0001 vermehrt auftreten. Für Amelia treten auch bei Betrachtung des MSE durchschnittlich um 0.1324 beziehungsweise 24% geringe Fehler als beim vollen Modell auf. Somit schneidet Amelia auch im Vergleich mit Random Forest im Logit-Modell besser ab. Die Varianzkomponente bei Random Forest nimmt in diesem Modell durchschnittlich ca. 40% höhere Werte an und liegt im Mittel um 0.065 pro Variable über der Varianzkomponente bei Amelia, wohingegen im linearen Modell die Streuung bei Amelia um 5.4% beziehungsweise um 0.0014 größer ist. Beim Vergleich der Varianzkomponente zwischen den Logit-Modellen mit wechselnden Parametern und mit festen Parametern ergibt sich eine um den Faktor 7.4 beim vollen Modell, um den Faktor 33.6 bei der Complete-Case-Analysis, um den Faktor 4.9 bei Amelia und um den Faktor 6.3 fach größere Streuung bei dem Modell mit wechselnden Parametern. Auch bezüglich der Biaskomponente schneidet das Logit-Modell im Vergleich mit den imputierten Datensätzen und der Complete-Case-Analysis zumindest teilweise schlechter ab. Während die Biaskomponente im Durchschnitt bei der Complete-Case-Analysis 92% höheren Präzisionsfehler aufweist, sind es bei Random Forest 682% und bei Amelia 934%. Allerdings fällt hier vor allem β_3 ins Gewicht, bei der die Biaskomponente in der Complete-Case-Analysis um Faktor 120, bei Amelia um Faktor 97 und bei Random Forest um Faktor 65 mal höher ist, sodass ohne Berücksichtigung dieser Variabel das Verhältnis der Biaskomponente verglichen mit dem vollen Modell im Mittel bei der Complete-Case-Analysis noch eine Steigerung um 358%, bei Random Forest um 13% beläuft. Für Amelia ergibt sich gegenüber dem vollen Modell eine im Durchschnitt um 37% geringere Verzerrung. Während im linearen Modell Amelia im Mittel eine um 0.003 und die Complete-Case-Analysis eine um 0.004 höhere Biaskomponente als das volle Modell verfügen, schneidet Random Forest im Schnitt um 0.003 besser als das volle Modell ab. Falls die relativen Veränderungen betrachtet werden, ist die Biaskomponente im Mittel dennoch bei Random Forest um den Faktor 25.8 mal höher als beim vollen Modell, bei Amelia ist die Verzerrung um den Faktor 10.2 und bei der Complete-Case-Analysis um den Faktor 2 mal höher.

Im Folgenden sollen wiederum die MSE-Komponenten bezüglich der Imputationen untersucht werden, welche in Tabelle 22 dargestellt sind.

Tabelle 22: MSE-Komponenten für Modell mit zufälligen Parametern, W=20

	Amelia_logit	Amelia_linear	RF_logit	RF_lineaer
Varianzkomponente				
Gesamt	10.4376	9.9299	2.9084	3.2983
x_1	99.7377	95.4997	27.7887	32.0251
x_2	0.6992	0.8331	0.2683	0.4011
x_3	0.1961	0.1963	0.0247	0.0290
x_4	0.2621	0.2520	0.0755	0.0886
x_5	1.9481	1.7422	0.4350	0.4017
x_6	1.0794	1.0885	0.1652	0.1783
x_{7_1}	0.1904	0.1898	0.0312	0.0366
x_{7_2}	0.1817	0.1749	0.0291	0.0345
x_{7_3}	0.1882	0.1896	0.0266	0.0320
x_8	0.3265	0.3265	0.0407	0.0483
Biaskomponente				
Gesamt	47.5642	47.9770	47.7981	46.7680
x_1	460.1865	462.4811	465.4879	453.5971
x_2	2.4631	1.1032	1.9430	1.2772
x_3	0.7549	0.7500	0.7236	0.7176
x_4	0.7509	0.7557	0.7309	0.7088
x_5	6.2828	6.3640	5.1310	5.2739
x_6	4.2091	4.3009	4.4170	4.4291
x_{7_1}	1.0287	1.0270	1.0803	1.0737
x_{7_2}	0.9662	0.9630	0.9958	0.9848
x_{7_3}	1.0206	1.0145	1.0502	1.0380
x_8	1.4409	1.4483	1.3752	1.3671

Im ersten Moment fällt im Vergleich zum Modell mit festen Parametern auf, dass die imputierten Daten bezüglich der Varianzkomponente bei Amelia um 15% im Logit-Modell und um 24% im linearen Modell und bei Random Forest um 25% im Logit- und um 60% beim linearen Modell schlechter abschneiden. Allerdings treten bezüglich der Varianzkomponente ähnliche Ergebnisse auf. Die Varianzkomponente der Imputationen von Amelia ist im Mittel mehr als 5 mal so groß wie die der Random Forest-Imputationen. Auch bei Betrachtung der Differenzen ist die Streuung der Imputationen bei Amelia, vor allem bei x_2, deutlich größer. Im Vergleich der im Logit- und im linearen Modell generierten Imputationen weist wiederum das Logit Modell einen um 7% beziehungsweise einen um 0.63 höhere Varianzkomponente im Mittel bei Amelia auf, während die Streuung bei Random Forest durchschnittlich um 0.30 beziehungsweise 6% größer ist. Auch bezüglich des Bias schneidet das Modell mit festen Parametern besser ab, sodass bei dem Logit-Modell und dem linearen Modell bei Amelia der Präzisionsfehler nach der Biaskomponente um 24 beziehungsweise 11% besser ist, während die Verzerrung bei Random Forest um 18% beim Logit- beziehungsweise 9% beim linearen Modell genauer ist. Für Amelia ergeben sich im Logit-Modell gegenüber Random Forest um im Mittel 1.75 geringere Biaskomponenten, im linearen Modell schneidet Random Forest um eine durchschnittlich 0.166 geringere Verzerrung ab. Bei Berücksichtigung der relativen Unterschiede reduziert sich der Präzisionsfehler im Mittel beim Logit Modell um 5 und beim linearen Modell um 1%.

Anschließend wird noch der $NRMSE$ des Modells in Tabelle 23 der Modelle mit festen und variierenden (zufällig gezogenen) Parametern verglichen. Zusätzlich zum MSE wird hier auch der NRMSE hinzugezogen, welcher sowohl den durchschnittlichen aboluten Fehler, als auch

Tabelle 23: NRMSE

	$logit_{Amelia}$	$linear_{Amelia}$	$logit_{RF}$	$linear_{RF}$
Parameter fest	0.4775	0.3564	0.4996	0.3639
Parameter zufällig	0.4041	0.3148	0.3977	0.3176

die relativen Verbesserungen berücksichtigt. Da durch die Standardabweichung der jeweiligen Variable normiert wird, können hier, wie bereits erwähnt, alle Variablen gleichzeitig betrachtet werden. Zunächst fällt beim Vergleich auf, das beezüglich des NRMSE das aus den zufällig generierten Parametern bei allen Imputationen besser abschneidet. Weiterhin ist der Fehlerterm auch beim NRMSE im Logit-Modell höher als beim linearen Modell. Im Vergleich von Amelia und Random Forest, liegt der NRMSE bei Amelia jeweils unter dem von Random Forest, während beim Modell mit zufälligen Parametern im logit Modell Random Forest besser abschneidet und im linearen Modell Amelia einen etwas geringeren NRMSE aufweist. Somit kann im Durchschnitt über alle Variablen festgestellt werden, dass die Imputationen im Amelia-Datensatz zumindest im Verhältnis zu deren Standardabweichung, bei den festen Parametern genauer geschätzt werden.

4.3 Modell mit multivariat-normalverteilten Daten

Bis jetzt wurden die X-Daten immer nach dem zu Beginn gewählten Muster aus vielen verschiedenen Verteilungen generiert. Damit ist, wie bereits erwähnt, die Voraussetzung Amelias, das alle Variablen einer Normalverteilung folgen, bis hierhin immer verletzt gewesen. Im Folgenden wurde noch ein Modell gerechnet, bei dem diese Annahme erfüllt ist. Der Erwartungswertvektor μ wurde wiederum zufällig gewählt, während eine feste Kovarianz von 0.4 für zwei Variablen x_i und x_j, $i \neq j$ bestimmt Die Faktoren für die Konstruktion von den Responsevariablen beziehungsweise die wahren β wurden weiterhin aus dem Pool der möglichen Ausprägungen gezogen.
:

$$\mu = (\frac{\theta_1}{2}, 0, 2, 0, 5, 1, .5, 2, 10, 3, 5)$$

$$\Sigma_{i,j} = \begin{pmatrix} 1 & 0.4 & 0.4 & \dots & 0.4 \\ 0.4 & 1 & 0.4 & \dots & 0.4 \\ 0.4 & 0.4 & 1 & \dots & 0.4 \\ \vdots & \vdots & & \ddots & \vdots \\ 0.4 & 0.4 & 0.4 & \dots & 1 \end{pmatrix}$$

Zunächst sollen wieder die Koeffizientenschätzer auf ihre Schätzfehler untersucht werden. Die entsprechenden Tabellen zur Bias-und Varianzkomponente sind im Anhang zu finden. Da vor allem das Abschneiden von Amelia mit Random Forest, auch im Vergleich der vorherigen Modelle, untersucht werden soll, bietet folgende Tabelle ein Vergleich von Random Forest und Amelia. Dafür wurden Varianz- und Biaskomponenten von Random Forest und Amelia miteinander verrecnet, und sowohl das Verhältnis als relativer Unterschied, als auch die Differenz als absoluter Vergleich, wie auch schon in den vorherigen Analysen miteinander verglichen. Somit ergibt sich folgendes Ergebnis:

Zunächst zeigen die Daten im Logit-Modell im Vergleich der ausschließlich einer Normalverteilung folgenden Variablen sogar ein besseres Abschneiden von Random Forest auf. Während der MSE beim Modell mit festen Parametern im Schnitt um 40% und beim Modell mit zufällig gezogenen Parametern um über 50% höher bei Random Forest als bei Amelia ist, beträgt der MSE beim Modell für Random Forest aus multivariat normalverteilten Daten durchschnittlich über 10% weniger. Auch bei Berücksichtigung der Differenzen schneidet Amelia im multivariat normalverteilten Modell schlechter ab als Random Forest, während die Differenz der Schätzfehler bezüglich der Koeffizienten in den anderen Modellen bei Random Forest höher ausfällt. Ähnliche Ergebnisse werden auch in den MSE-Komponenten erzielt, sodass für das Logit-Modell in allen

	log_{fest}	lin_{fest}	log_{zufall}	lin_{zufall}	log_{MVN}	lin_{MVN}
MSE-Gesamt						
Verhältnis	1.4168	1.0390	1.5642	1.0208	0.8967	1.1443
Differenz	0.0136	0.0065	0.1428	-0.0069	-0.0037	0.8045
Varianzkomponente						
Verhältnis	1.1287	1.0734	1.3973	0.9457	0.9195	1.1525
Differenz	0.0035	0.0073	0.0645	-0.0014	-0.0033	0.8167
Biaskomponente						
Verhältnis	27.8029	4.2712	5.4975	643.6043	1.2846	1.2920
Differenz	0.0101	-0.0008	0.0783	-0.0055	-0.0004	-0.0122

Messparametern der gegenteilige Effekt auftritt. Hinsichtlich der Varianzen unterschätzen die geschätzten Modell-Varianzen den empirischen MSE eindeutig. Je höher allerdings die Schätzfehler sind, desto besser wird der empirische MSE geschätzt. Im linearen Model hingegen wird die These, dass Amelia bei multivariat normalverteilten Daten besser abschneidet, bestätigt. Sowohl bezüglich der durchschnittlichen Verhältnisse und der durchschnittlichen Differenz weist Amelia gegenüber Random Forest bei multivariat normalverteilten Daten geringere Fehler auf. Auch in der Varianzkomponente erzielt Amelia bei diesem Modell die besten Ergebnisse, nur bezüglich der Biaskomponente schneidet Random Forest im linearen Modell für MVN-Daten besser als Amelia ab. Trotzdem fällt der kombinierte MSE für Amelia im Mittel um 14% oder 0.805 besser aus. Ein Vergleich der imputierten Werte mit den anderen Modellen, bei denen X mehrfach gezogen wird, ist aufgrund der unterschiedlichen Datengrundlage nicht sinnvoll. Allerdings beträgt der MSE der Variablen im Logit-Modell 0.811 bei Amelia und 0.832 bei Random Forest, während im linearen Modell der MSE für Amelia 0.736 und bei Random Forest 0.746 beträgt. Somit ergibt sich bei Amelia ein um 2.5 beziehungsweise 1.3% geringerer MSE der Imputationen. Das bessere Abschneiden von Amelia kann also nur im linearen Modellen anhand diesen Resultaten bestätigt werden, im Logit-Modell treten sogar höhere Fehler auf. Bezüglich der Imputationen kann auch eine geringfügig genauere Schätzung erzielt werden. Somit ergeben sich für den NRMSE beim Modell mit multivariat normalverteilten Daten für das Logit Modell bei Amelia 0.0748 und bei Random Forest 0.054 und im linearen Modell für Amelia 0.076 und für Random Forest 0.054. Der NRMSE ist also auch im linearen Modell geringer als beim Logit-Modell. Obwohl die Annahme von Amelia bezüglich der Verteilung der Daten erfüllt ist, schneidet Amelia trotzdem schlechter ab als Random Forest. Anhand der deutlich höheren Zusammenhänge der Kovariablen kann dieser NRMSE aber nicht wirklich mit den anderen Werten verglichen werden.

5 Fazit

5.1 Zusammenfassung der Ergebnisse

In der Thesis wurde der Einfluss der Imputation fehlender Daten anhand Amelia und Random Forest untersucht. Als Messparameter wurde der MSE verwendet, der noch in eine Varianz- und eine Biaskomponente zerlegt wurde. Bezüglich der Koeffizientenschätzer können nur geringe Unterschiede zwischen den vollständigen und den imputierten Datensätzen festgestellt werden. Sowohl Amelia und auch Random Forest schätzten die fehlenden Werte ausreichend gut ein, so dass die Schätzfehler der Koeffizientenschätzer ähnlich hoch sind wie die Fehler des auf den vollständigen Datensatz beruhenden Modells. Die Complete-Case-Analysis führt bei einem derartig großen Anteilen an fehlenden Daten zu deutlich höheren Fehlerkomponenten, im Logit-Modell sind zudem einige der Ergebnisse aufgrund Nicht-Konvergenz der Regressionsmodelle nicht mehr sinnvoll interpretierbar. Generell fällt im Verlauf der Untersuchung auf, dass Amelia über eine geringere Streuung in den Koeffizientenschätzern verfügt, während die Präzisionsfehler der Koeffizientenschätzer bei Random Forest niedriger ausfallen. Von den untersuchten Parametern beeinflusst

vor allem σ die Fehlerkomponenten. Je höher die Streuung des Fehlereterms ϵ, desto ungenauer werden die Koeffizienten geschätzt. Was die Anzahl der wiederholten Ziehungen des Responsevektors betrifft, weißen die Daten keinen Zusammenhang zu den Schätzfehlern hergestellt werden. Während eine Erhöhung des Stichprobenumfanges generell zu geringeren Varianzkomponenten führt, beeinflusst eine Veränderung des erwarteten Anteils von fehlenden Werten vor allem die Präzision der Schätzungen Bei einer mehrfachen Ziehung der Daten X führen beim Festhalten der Parameter zu einer geringere Streuung der Koeffizientenschätzer , allerdings können diese bei Variation der Parameter mit geringerem Bias geschätzt werden. Im Vergleich zwischen Amelia und Random Forest schneidet Amelia bezüglich der Koeffizientenschätzer vor allem bei Modellen mit höheren Schätzfehlern, größeren Anteilen an fehlenden Daten und größeren Stichprobenumfängen besser ab. Bezüglich der Verteilung der Variablen konnte die Annahme, dass Amelia bei normalverteilten Daten bessere Ergebnisse im Vergleich mit Random Forest erzielt, nicht bestätigt werden. Allerdings kommt es bei der Untersuchung vereinzelt zu besonders hohen relativen Verhältnissen der Fehlerkomponente zugunsten Amelia, so dass Random Forest deutlich höhere relative Fehler aufweist, was vermehrt bei normalverteilten Daten auftritt. Was die Imputationen der Daten angeht, schneidet Random Forest doch recht deutlich besser ab. Für die aus den imputierten Datensätzen resultierenden Regressionsmodelle bleibt der Unterschied dennoch nicht bestehen. Amelia biete also eine einfache und sinnvolle Lösung für das fehlende-Daten-Problem an. Trotz vieler Variationen von Parametern können fehlende Werte akkurat geschätzt werden. Der Einfluss der Verteilung der verwendeten Daten konnte in dieser Thesis nicht bestätigt werden. Während Random Forest für die direkten Imputationen dieser Daten noch besser geeignet ist, gelingt Amelia bezüglich der Koeffizientenschätzer ähnlich gute Ergebnisse. Im Allgemeinen sind die Ergebnisse sind sehr stark von den gewählten Parametern und Daten abhängig. Somit können keine grundsätzlichen Aussagen über Amelia beziehungsweise Random Forest getroffen werden, ohne diese auf die verwendeten Daten zu bedingen. Im Modell wurde keine kategoriale Variable verwendet, somit könnte das Abschneiden von Amelia und Random Forest auch vom Skalenniveau der Daten betroffen sein. Weiterhin wurde das Random Forest-Modell nicht getuned, auch so könnte noch eine bessere Performance bei Random Forest erzielt werden. Ein weiterer Parameter, der hier nicht betrachtet wurde, ist die Rechenleistung, wobei hier Amelia deutlich weniger Zeit benötigt. In der Praxis gibt es natürlich nahezu unendlich Faktoren, die über den Vergleich zwischen Amelia und Random Forest entscheiden. Dennoch bietet Amelia eine einfache und sinnvolle Lösung für das fehlende-Daten-Problem an. Trotz vieler Variationen von Parametern können fehlende Werte akkurat geschätzt werden. Der Einfluss der Verteilung der verwendeten Daten konnte in dieser Thesis nicht bestätigt werden. Während Random Forest für die direkten Imputationen dieser Daten noch besser geeignet ist, gelingt Amelia bezüglich der Koeffizientenschätzer ähnlich gute Ergebnisse.

6 Anhang

6.1 Tabellen

Im Folgenden ist die Bias- und Varianzkomponente Modells mit ausschließlich multivariat normalverteilten Daten gelistet. Für das Logit-Modell konnte die Complete-Case-Analysis wiederum aufgrund Nicht-konvergenz nicht berücksichtigt werden.

Tabelle 24: Logit-Modell

	volles Modell	Amelia	RF
Varianzkomponente			
β_1	0.0010	0.0015	0.0014
β_2	0.0343	0.0463	0.0432
β_3	0.0268	0.0351	0.0361
β_4	0.0377	0.0533	0.0485
β_5	0.0194	0.0475	0.0449
β_6	0.0235	0.0388	0.0321
β_{7_1}	0.0214	0.0338	0.0341
β_{7_2}	0.0248	0.0399	0.0326
β_{7_3}	0.0311	0.0426	0.0346
β_8	0.0312	0.0427	0.0414
Biaskomponente			
β_1	0.0036	0.0016	0.0013
β_2	0.0000	0.0072	0.0029
β_3	0.0018	0.0042	0.0008
β_4	0.0008	0.0076	0.0010
β_5	0.0010	0.0102	0.0039
β_6	0.0029	0.0078	0.0000
β_{7_1}	0.8053	0.8283	0.8450
β_{7_2}	0.6240	0.7872	0.8018
β_{7_3}	0.2627	0.2186	0.2120
β_8	0.0006	0.0000	0.0002

Tabelle 25: Lineares Modell

	volles Modell	CC	Amelia	RF
Varianzkomponente				
β_1	0.1240	1.3952	0.1950	0.2195
β_2	5.3346	38.0948	7.9411	7.9884
β_3	2.4095	60.5614	4.4518	4.9890
β_4	4.5774	45.6422	5.5112	6.7102
β_5	2.6246	66.5731	4.3017	4.9866
β_6	3.4696	38.8356	6.3965	7.1210
β_{7_1}	3.0092	42.0746	4.8547	5.8383
β_{7_2}	2.8520	52.3942	4.7173	4.9524
β_{7_3}	4.0933	32.1298	7.0420	9.0219
β_8	3.8739	76.5279	7.0225	8.7735
Biaskomponente				
β_1	0.0000	0.0612	0.0013	0.0020
β_2	0.0358	0.0735	0.0831	0.1271
β_3	0.0440	0.8851	0.0435	0.0595
β_4	0.0376	0.1436	0.2385	0.2426
β_5	0.0015	0.2063	0.0226	0.0000
β_6	0.1201	1.1598	0.4349	0.4435
β_{7_1}	0.1905	0.0439	0.0541	0.0327
β_{7_2}	1.5378	0.0438	0.8694	0.8676
β_{7_3}	1.1406	0.0787	1.1555	0.9963
β_8	0.0101	0.5003	0.0033	0.0130

6.2 R-Code

Da es sich bei der Thesis um eine Simulationsstudie handelt, wurden alle Daten in R selbst generiert. Zur Erstellung der Thesis wurden 4 Funktionen verwendet:

1. Funktion 1
2. Funktion 2
3. Auswertung 1
4. Auswertung 2

Funktion 1 erstellt die Funktionen, bei denen nur ein X-Datensatz erstellt wurde. Die Funktion gibt am Ende neben der Übersicht aller übergebenen Parameter den mittleren AIC, der Modelle, sowie MSE, geschätzte Varianz und Bias- und Varianzkomponente der MSE-Zerlegung aus, jeweils sowohl für das Logit- als auch das lineare Modell. Neben den Parametern muss der Funktion noch übergeben werden, ob X wie beschrieben generiert werden oder ob X einer multivariaten Normalverteilung folgen soll. Zudem muss die Anzahl an wiederholten Ziehungen des Responsevektors angegeben werden. Bei Funktion 2 hingegen wird X auch mehrfach simuliert. Funktion 2 gibt keine geschätzte Varianz mehr aus, dafür aber die Fehlerkomponenten der imputierten Daten und den NRMSE zusätzlich zu den gleichen Ausgabeparametern wie Funktion 1. Da die Parameter zur Bestimmung von X und den Responsevektoren in der Funktion zufällig gezogen werden, müssen nur die Anzahl an Ziehungen von X und den Responsevektoren übergeben werden. Zur Auswertung der beiden Funktionen wurden jeweils weitere Skripte in R erstellt. Somit werden je 3 der gespeicherten Ergebnisse aus Funktion 1 bzw. Funktion 2 aufgerufen und deren Ergebnisse zusammengefügt. Anschließend werden die Ergebnisse unter Angabe der gewünschten Fehlerkomponente verglichen.

6.3 Referenzen

Literatur

[1] J. B. Carlin, N. Li, P. Greenwood, C. Coffey, et al. Tools for analyzing multiple imputed datasets. *The Stata Journal*, 3(3):226–244, 2003.

[2] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B*, 39(1):1–38, 1977.

[3] L. Fahrmeir, T. Kneib, S. Lang, and B. Marx. *Regression: models, methods and applications*. Springer Science & Business Media, 2013.

[4] A. Gelman and J. Hill. *Data analysis using regression and multilevel/hierarchical models*, volume Analytical methods for social research. Cambridge University Press, New York, 2007.

[5] J. Honaker and G. King. What to do about missing values in time series cross-section data. *American Journal of Political Science*, 54(3):561–581, 2010.

[6] J. Honaker, G. King, and M. Blackwell. Amelia II: A program for missing data. *Journal of Statistical Software*, 45(7):1–47, 2011.

[7] A. Liaw and M. Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002.

[8] R. J. Little and D. B. Rubin. *Statistical analysis with missing data*. Wiley series in probability and mathematical statistics. Probability and mathematical statistics. Wiley, 2002.

[9] S. Oba, M.-a. Sato, I. Takemasa, M. Monden, K.-i. Matsubara, and S. Ishii. A bayesian missing value estimation method for gene expression profile data. *Bioinformatics*, 19(16):2088, 2003.

[10] D. B. Rubin. *Multiple Imputation for Nonresponse in Surveys*. Wiley, 1987.

[11] D. J. Stekhoven and P. Buehlmann. Missforest - non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 2012.